

A Project Report
on
***Feature Based Segmentation for Transport Mode Detection using
Spatial Trajectory Data in Urban Environment***

Submitted for the partial fulfilment of the requirement for the

Award of the Degree of
Bachelor of Technology

in
COMPUTER SCIENCE & ENGINEERING

by
Arvind Singh Rawat
Akash Chhetri
Ayush Bharagava
Shashwat Sinha

Under the Guidance of
Dr. Mainak Bandhyopadhyay



DIT UNIVERSITY, DEHRADUN, INDIA

April 2019



DECLARATION

This is to certify that the Project entitled “*Feature Based Segmentation for Transport Mode Detection using Spatial Trajectory Data in Urban Environment*” in partial fulfillment of the requirement for the award of the Degree of Bachelor Of Technology submitted to DIT University, Dehradun, Uttarakhand, India, is an authentic record of bona fide work carried out by me, under the guidance of Dr. Mainak Bandyopadhyay.

The matter embodied in this Project/Thesis/Dissertation has not been submitted for the award of any other degree or diploma to any University/Institution.

Student's Name and Signature

1. Arvind Singh Rawat
2. Akash Chhetri
3. Ayush Bhargava
4. Shashwat Sinha

Guide Name, Designation
Name and Signature

Dr. Mainak Bandyopadhyay
Assistant Professor
Computer Science and Engineering

Mr. Vishal Bhrati

Head of Department(CSE)

Mr. Gagan Singh

Controller of Examination

Date : 30/4/2019

Place : Dehradun



CERTIFICATE

This is to certify that the Project entitled “*Feature Based Segmentation for Transport Mode Detection using Spatial Trajectory Data in Urban Environment*” in partial fulfilment of the requirement for the award of the degree of Bachelor Of Technology in Computer Science and Engineering, submitted to DIT University, Dehradun, Uttarakhand, India, is an authentic record of bona fide research work carried out by

Mr. Arvind Singh Rawat - Roll No. 160102261,

Mr. Akash Chhetri Roll No. 160102293,

Mr. Ayush Bhargava Roll No. 160102287 and

Mr. Shashwat Sinha Roll No. 160102259,

under my supervision/ guidance.

Dr. Mainak Bandyopadhyay

Assistant Professor

Computer Science and Engineering

Date : 30/4/2019

Place : Dehradun



Acknowledgement

We would like to express our sincere gratitude to our mentor Mr. Amandeep Saini for his immense support and guidance. We would also like to thank faculty and staff at DIT University, Dehradun for providing a friendly learning working environment and of course, the internet, for always providing help.

Arvind Singh Rawat

Akash Chhetri

Ayush Bharagava

Shashwat Sinha

(B. Tech CSE – E)

(Session 2018-19)

Abstract

Human overpopulation is among the most pressing environmental issues. And one of the problems due to it is to maintain the connectivity within the peoples. The connectivity here is could be any mode of communication or any mode of transportation. Now, the world is in the IT Age thus communication is not a significant problem but transportation-related problems still persist. The need for transportation is not only limited to the availability of modes of transportation but it is also dependent on the medium of transportation. To satisfy this need humans invented subways and flyovers. But the ever going increase in demand for transportation is unstoppable. So we have to develop new ways to tackle problems related to transportation. Due to significant development in the field of GNSS, the location-based researches are now much improved. Now, most of the smartphones have one or more GNSS embedded in them. This system allows to track their daily routine, provide location-based services, location-based recommendation, etc. So we can also use this system to identify the need of an individual or general public with the help of analysis of their data and this technique is classified as Activity Recognition. Many of the users use activity recognition on a daily basis to track their fitness routine, to enjoy automated assistance in their work, etc. Thus we are going to use the same technology to determine the mode of transportation of an individual in transit. This process can become the preliminary for the other processes like transit time calculation, optimal route suggestion, and it also helps researchers in pre-processing the trajectory data because the different mode of transportation affects the nature of data and hence may require a different approach to the process.

The project starts from manually retrieving the GPS data through smartphones. Then the data is fed in the pre-processing stage. This stage resolves the problems regarding the raw data, like handling semantic errors, removing outliers etc. Then the extracted data is processed through the segmenting stage to decompose the whole data into small homogenous datasets. At last, data enters the final stage, which is to identify the mode of transportation in each segment through analysis. Our aim is to automate the process as much as possible.

The challenge is to identify the mode of transportation because the modes of transportation don't follow the same characteristics everywhere, every time. Example the speed of the vehicle is much more in the general time as compared to prime time. Here prime time is the time when

traffic is most like morning time 8:00 to 10:00 and evening time 5:00 to 7:00 when the general employee's office time starts or ends. And this time varies for different places. Also, the average speed in India is less than the western countries. So, one way is to fix the speed threshold according to location-based speed limits like in local roads the speed limit for motorcycle is 40kmph while of Car is 30kmph while on the highway it is much faster. But the problem with this solution is that speed is solely dependant on driver of the vehicle, which means some of the people doesn't follow the limits while others could not follow due to some intrinsic (like, vehicle condition, driver experience etc) or some external (like, road condition, traffic congestion, environmental conditions etc.) conditions. And, thus the decision boundary for these cases are not very crisp and thus overlaps other conditions. Thus we are using the Fuzzy Inference System to counter this problem.

Lastly, we are comparing results obtained from the various algorithms and comparing performances of the various algorithm in terms of Accuracy, Sensitivity, and other factors.

Table of Contents

Title.	Page No.
Declaration	(i)
Certificate	(ii)
Acknowledgement	(iii)
Table of Contents	(iv)
Abstract	(v)
Chapter 1 : Introduction	
Objective	1
Motivation	2
Theory	3
Chapter 2 : Overall description	
Design	11
Modules Used	13
Use Case	15
Dependencies and Limitations	16
Chapter 3 : Conclusion and Future Work	
Conclusion	17
Future Developments	18
Chapter 4 : Appendix	
Key Terms	19
References	20
Abbreviation	21
List of Figures	21

Objective

This Project aims for three objectives and hence can be decomposed in three basic parts:

Project for Mode of Transport Detection:

- This Project aims for the analysis of GPS data (commonly collected by the user with the help of their smartphones and a logger app) to determine the Mode of Transportation.
- This project uses Segmentation algorithms to segment the GPS data to make homogenous segments and thus each segment assures that the mode of transportation in that segment is common.
- The segments thus generated can be analyzed over a variety of factors to determine the Mode of Transportation. The factors involved could be Speed, Time, and Location etc.

Module for Segmentation of data:

- This project serves as the module for end user to use the algorithms implemented by this algorithm to preprocess their data.
- This module provides different classes for different algorithms and each class contains some common attributes like data, anchors etc. while some classes may contain algorithm specific attributes.
- Each method would contain a docstring that explains the use of method, its arguments and what that method returns.

Research on Segmentation Algorithms:

- The project implements multiple segmentation algorithms, thus all the algorithm will work upon same data and hence provides the opportunity to measure their performances when pitted against each other in same environment.
- Also, we are going to discuss the complexities of the different algorithm.

Motivation

The implementation of this project is motivated by the concept of Time Series segmentation. To implement this project we are using multiple algorithms which follows different approaches to meet the same goal. Thus this project also provides a comparison between different approaches with the help of F1 score and other factors like Accuracy, Sensitivity etc. This comparison will help us to choose the optimal approach to fulfil our requirement.

In this rising world of urbanization and actively throbbing real estate industry, GPS (global positioning system) is widely implemented and used in our day to day needs. Apart from this, all of the delivery apps use the same technology to deliver foods and goods.

This might seem trivial, but as per the online census, this positioning system has a whopping error of 52% in its prediction, and thereby largely making this positioning system highly unreliable.

If successfully developed, and deployed this will revolutionize the field of positioning systems.

Our project can be used in transportation science where the focus is on measuring daily travel patterns of individuals or group of individuals.

The motivation for transport mode detection is the growing need in a different kind of MaaS (Mobility as a Service) based services that require reliable information about the recognized transportation mode in order to serve their own customers with the right product offering.

Theory

SEGMENTATION

The process of decomposing a large dataset into relatively smaller homogenous chunks. Segmentation is used in Industries for Market segmentation or Customer segmentation, Image Segmentation etc. Segmentation can be either done on the basis of similarity of one or more features or on the basis of the similarity in the geometry of the data. The former one is known as Feature-based segmentation and the latter one is Geometric segmentation. We are using the Feature-based segmentation.

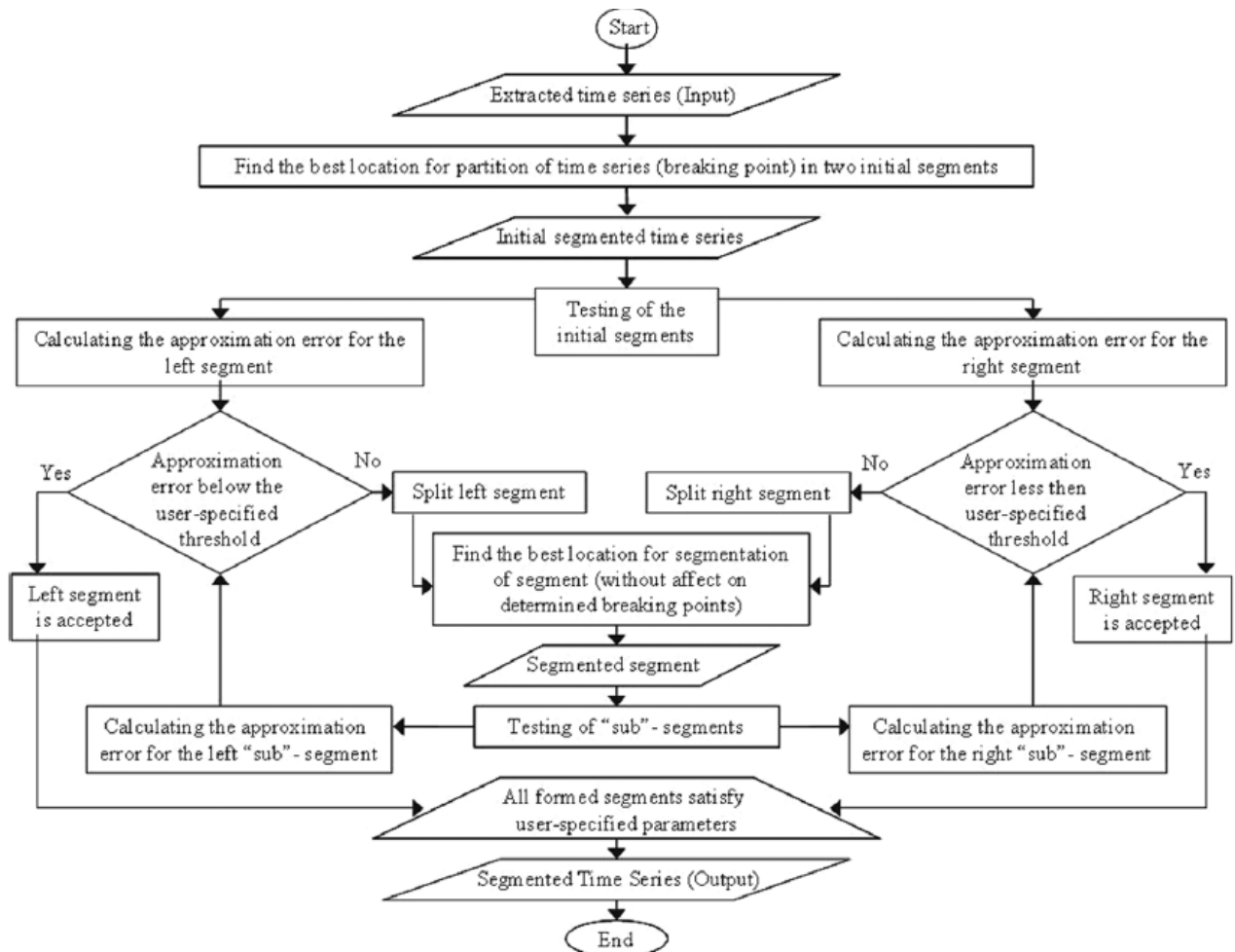


SEGMENTATION ALGORITHMS

Top Down Algorithm

The segments are computed in the Divide and Conquer manner. In this approach, Initially, the whole dataset is considered as a single segment then in subsequent iterations anchor is placed when the difference between the features of subsequent segments exceeds the user-defined threshold. That anchor acts as the start boundary of the new segment and the end boundary of the previous segment. And then for the newly created segments, the process is repeated until the user-defined uniformity is achieved.

Figure 1 : Flowchart from Top Down Algorithm

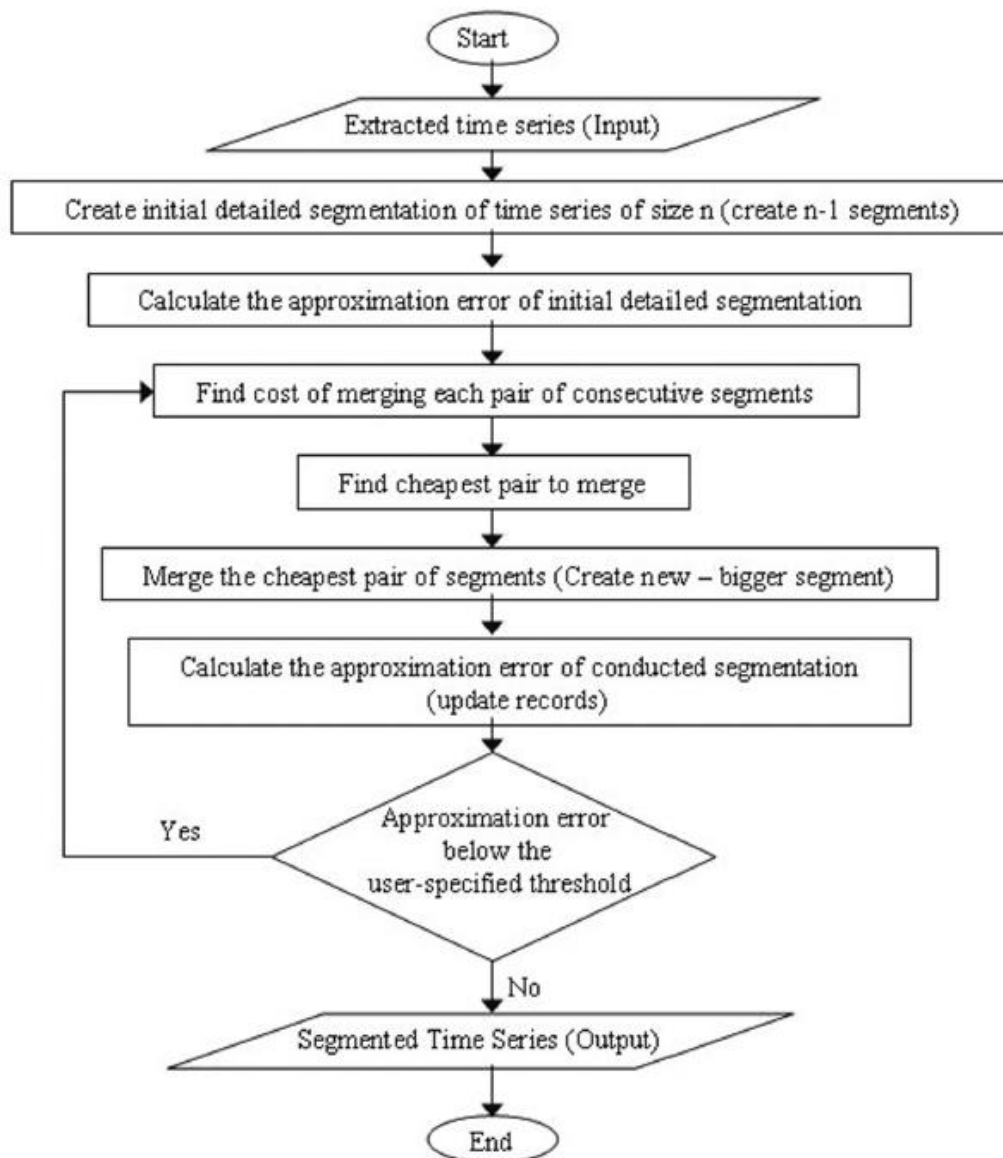


Source : [1]

Bottom Up Algorithm

The segments are computed in the Iterative merge manner. In this approach, Initially, the whole dataset is divided into atomic chunks. Then in further iterations, the pair of adjacent chunks with the lowest difference are merged. This process is repeated until the maximum possible chunks sizes are available which maintains the user-defined homogeneity.

Figure 2 : Flow Chart for Bottom Up Algorithm

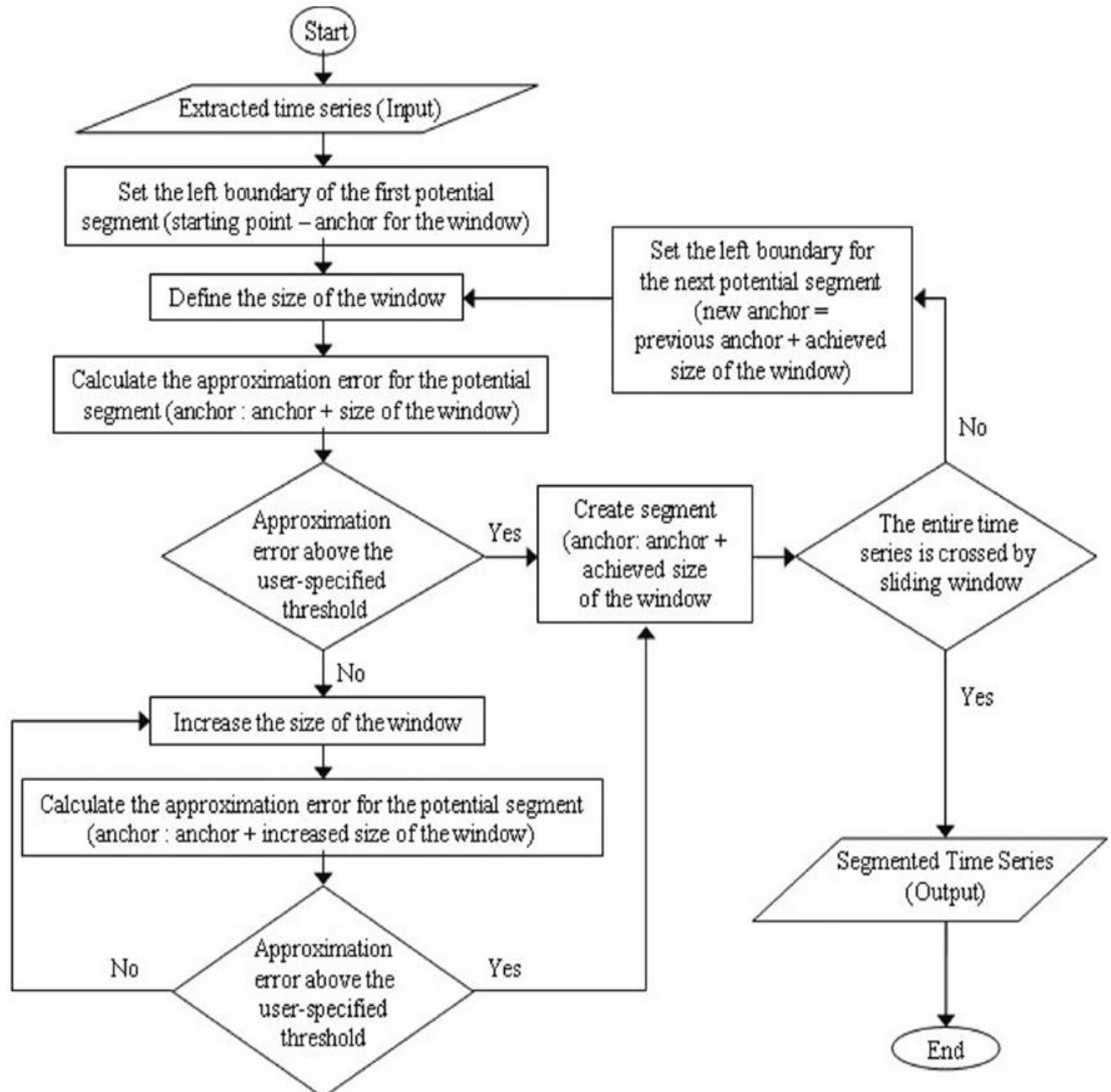


Source : [1]

Sliding Window Algorithm

This algorithm uses a Window of the user-specified length and the window is shifted over the data in iterations. The anchor is placed when the difference created by including one more data point is more than the predefined threshold.

Figure 3 : Flowchart for Sliding Window Algorithm

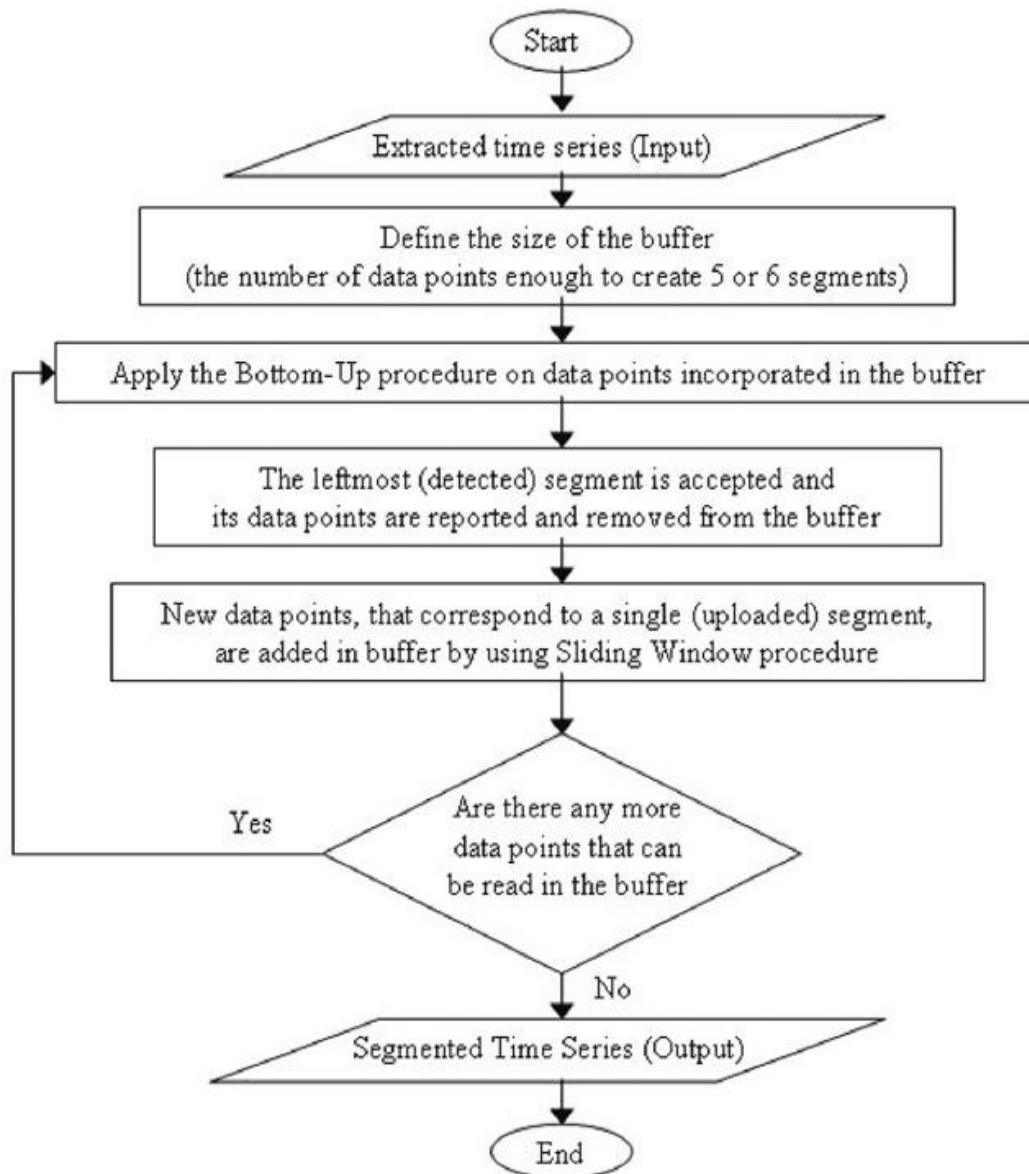


Source : [1]

Sliding Window and Bottom Up (SWAB) Algorithm :

This algorithm is the hybrid of the above two mentioned algorithms. Two fixed sized windows are created adjacent to each other and in iterations, those windows are shifted over the data. The distribution of data of both windows are then compared if the difference exceeds the limit specified by the user, then the anchor is placed between the windows.

Figure 4 : Flow Chart for SWAB Algorithm



Source : [1]

FUZZY LOGIC

Fuzzy logic is an approach to perform computation based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based.

Fuzzy logic includes 0 and 1 as extreme cases of certainty (or "the state of matters" or "fact") but also includes the various states of certainty in between, for example, the result of a comparison between two things could be not "tall" or "short" but "0.38 degree of tallness."

Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths which we aggregate further into higher truths which in turn, when certain thresholds are exceeded, cause certain further results such as motor reaction. A similar kind of process is used in neural networks, expert systems and other artificial intelligence applications. Fuzzy logic is essential in the development of human-like capabilities for AI, sometimes referred to as artificial general intelligence: the representation of generalized human cognitive abilities in software so that, faced with an unfamiliar task, the AI system could find a solution.

Fuzzy logic operators

Fuzzy logic works with membership values in a way that mimics Boolean logic. To this end, replacements for basic operators AND, OR, NOT are available. There are several ways to do this. A common replacement is called the Zadeh operators:

Boolean	Fuzzy
AND(x,y)	MIN(x,y)
OR(x,y)	MAX(x,y)
NOT(x)	$1 - x$

For TRUE/1 and FALSE/0, the fuzzy expressions produce the same result as the Boolean expressions.

FUZZY INFERENCE SYSTEM

Fuzzy Inference System is the key unit of a fuzzy logic system having decision making as its primary work. It uses the “IF...THEN” rules along with connectors “OR” or “AND” for drawing essential decision rules.

Characteristics of Fuzzy Inference System

- The output from FIS is always a fuzzy set irrespective of its input which can be fuzzy or crisp.
- It is necessary to have fuzzy output when it is used as a controller.
- A defuzzification unit would be there with FIS to convert fuzzy variables into crisp variables.

Functional Blocks of FIS

The following five functional blocks will help you understand the construction of FIS –

Rule Base – It contains fuzzy IF-THEN rules.

Database – It defines the membership functions of fuzzy sets used in fuzzy rules.

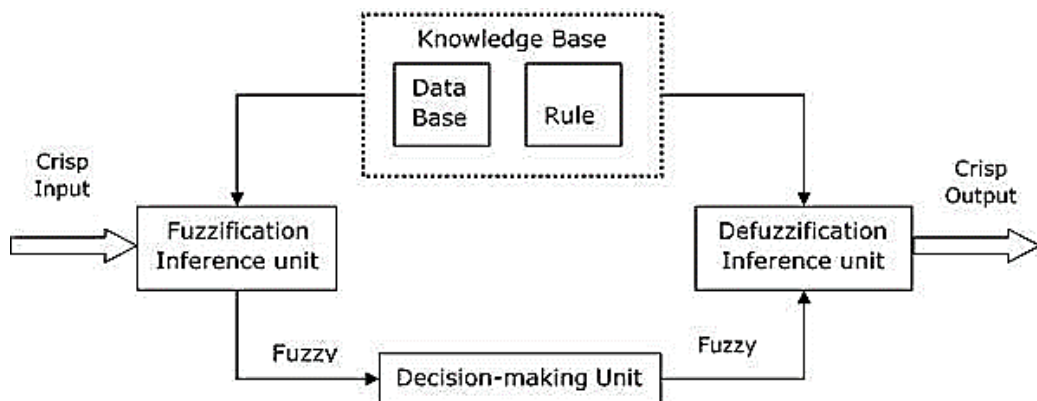
Decision-making Unit – It performs operation on rules.

Fuzzification Interface Unit – It converts the crisp quantities into fuzzy quantities.

Defuzzification Interface Unit – It converts the fuzzy quantities into crisp quantities.

Following is a block diagram of fuzzy inference system.

Figure 5 : Block Diagram of Fuzzy Inference System



MAHALANOBIS DISTANCE

The Mahalanobis distance is a measure of the distance between a point P and a distribution D, introduced by P. C. Mahalanobis in 1936. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. This distance is zero if P is at the mean of D, and grows as P moves away from the mean along each principal component axis. The Mahalanobis distance measures the number of standard deviations from P to the mean of D. If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. The Mahalanobis distance is thus unitless and scale-invariant, and takes into account the correlations of the data set.

$$d_e(t) = (e_1 - e_2)^2 \cdot \left(\frac{1}{2 \cdot \sigma_1^2} + \frac{1}{2 \cdot \sigma_2^2} \right)$$

where,

$d_e(t)$ = Mahalanobis Distance between two distribution

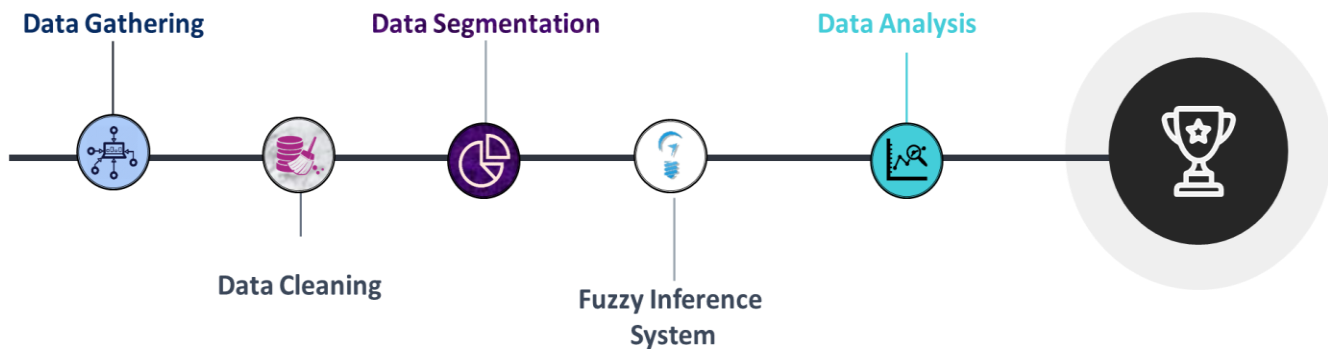
e_1 = Mean of first distribution

e_2 = Mean of second distribution

σ_1^2 = Variance of first distribution

σ_2^2 = Variance deviation of second distribution

Design



Data Gathering

- This step involves the extraction of data from the Source.
- Then, storing the data collected into a pre-defined format.
- Data is stored in a structured format

Data Cleaning

- It is also called the Preprocessing stage.
- This step involves removing redundant or error-prone data from the dataset.
- This process consists of handling NULL, NAN and any other odd values.
- Typecasting (to specify the data in user-defined type and to maintain consistency) and Data normalization (If the comparison is needed to be done between two different features) are also the part of this phase.
- This is one of the crucial phases and directly influence the outcome of the process. Without pre-processing, data can result in unexpected results while results obtained from pre-processed data are more consistent and reliable.

Data Segmentation

- This step involves Segmentation of data on the basis of similarity of features.
- This process will decompose the whole dataset into segments in which each data follows a general trend in one or more features.
- In this phase, we are implementing four segmentation algorithm (Top-Down, Bottom-Up, Sliding Window, Dual Sliding Window).
- Each Algorithm works on a different set of hyper parameters e.g. sliding window algorithm needs window size and threshold while top down algorithm requires the Threshold only.

Fuzzy Inference System

- There is no crisp decision boundary to differentiate between modes of Transportation, thus this project uses the concept of Fuzzy Inference system.
- Fuzzy Inference System uses a Rule base that specifies a set of If- Else conditions to determine the final Crisp Output and an inference engine that evaluates the values on the basis of different membership functions.

Data Analysis

This step involves multiple subtasks:

- Identification of optimal threshold for segmentation.
- Identification factor (or factors) of classification of Mode of transportation.
- Comparison of different segmentation algorithms on different factors like time complexity, space complexity, accuracy etc.

Modules Used

Python Modules:-

1. Pandas:

In PYTHON programming, Pandas is a software library written for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free 17 software released under the three-clause BSD license.

2. Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

3. Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms

4. Folium

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the Leaflet.js library. Manipulate your data in Python, then visualize it in on a Leaflet map via Folium.

5. Skfuzzy

This package implements many useful tools for projects involving fuzzy logic, also known as grey logic. It is pure Python and depends only on NumPy, SciPy, and NetworkX. There are many basic fuzzy logic functions in the base namespace, as well as a Pythonic, object-oriented system for fuzzy control systems in the skfuzzy.control submodule.

6. Preprocessing

- Time_Extractor.py
- Distribution_calculator.py

7. Data Cleaning

- Preprocessing.py

8. Data Segmentation

- **Dual Sliding Window**

Two fixed sized windows are created adjacent to each other. The distribution of data of both windows are then compared if the difference exceeds the limit specified by the user, then the anchor is placed between the windows.

- **Sliding Window**

The anchor is placed when the difference created by including one more data point is more than the predefined threshold.

- **Top-Down**

The anchor is placed when the difference between the variance of subsequent segments is more than a user-defined threshold. The segments are computed in the Divide and Conquer manner.

- **Bottom-Up**

The anchor is removed when the difference between means of consecutive segments is less than the user-specified threshold.

9. Fuzzy Inference System

- FuzzySystem.py (under Development)

10. Data Analysis

- Performance.py (under Development)

Use Cases

- Our project can be used in transportation science where the focus is on measuring daily travel patterns of individuals or group of individuals. The service provider can manage and monitor the transport more efficiently. This project also enable them to project the budget for their next tenure.
- The motivation for transport mode detection is the growing need in a different kind of MaaS (Mobility as a Service) based services that require reliable information about the recognized transportation mode in order to serve their own customers with the right product offering.
- Less and less human intervention will be needed since the “Intelligent System” can learn and decide upon the action to be taken with each varying data input for the prime result. Thus reducing the chances of errors and solution of the problem of efficiency reduction with the time.
- Insurance companies and ride-sharing service providers are interested in the driving behaviour of their customers and drivers. It also helps users to recommend them the ride they prefer by monitoring there daily mode of commute.
- It can be used as modules within Geographic Information System (GIS), namely in mapping, telecom and network services, accident and hotspot analysis, urban planning, and many more applications thereby proving the robustness of these algorithms.
- It can be used in Defence Sector for monitoring and keeping track of the vehicles in a high alert region.
- It can be implemented in monitoring and controlling probable levels of pollution in city by monitoring the inflow and outflow of vehicles from the city road network. Thus also help in Urban Planning and road Network management.
- Traffic can be managed efficiently using these techniques. This project report can be used to divert heavy vehicles from traffic prone road networks.

Dependencies

- The accuracy of projects depends on the large standard datasets and data sets with high sampling rate.
- Datasets with more features led to more accuracy.
- Performance depends on the hardware used to log the GPS data.
- Data to be worked upon should be strictly pre - processed, without irrelevant information.
- Accuracy of data depends upon the hardware or software used to collect and preprocess it.
- Accuracy depends highly on environmental factors.
- Sometimes, discrepancy in data is caused due to external factors, like traffic etc.

Limitations

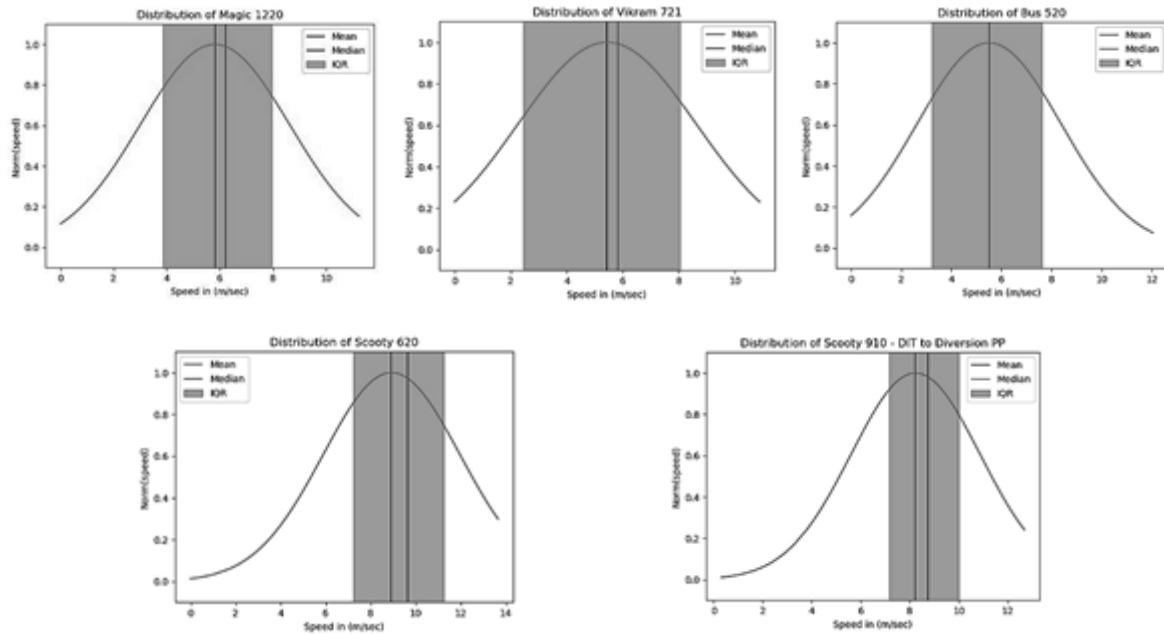
- Due to fact, that application might be deployed online, it might go on to become highly expensive as far as time management is concerned, with increase in data points.
- Speed based segmentation is highly based on the geography of trajectory involved.
- Sometimes, the algorithm might fail due to mechanical inefficiency of the concerned machine.

Conclusion

During this project phase we found many conclusions and all of them are listed below:

1. The distribution of speed is very similar in all of the four wheelers among themselves while all of the two wheelers reflects the same distribution.

Distribution of Speed



2. The speed is necessary but not sufficient for the identification of the transport mode, we need multiple features for the correct identification of transport mode detection. As the decision boundary that separates two different classes of transport mode is not very crisp and hence speed is not sufficient.
3. The speed of transport mode depends on location as heavily populated places tend to have lower speed thresholds as compared to less populated areas. Road networks with narrower road, heavy traffic or sharp turns generally have less average speed as compared to their counterparts.
4. The speed of transport also depends on the time of transit, as times of Office hours, or school hours, etc. starts or ends is much slower than other times as congestion increases.

5. For Mahalanobis Distance :

$$de(t) = NaN$$

When both samples are identical in nature thus we simply slide the window without saving anchor.

$$de(t) = Inf$$

When both samples are not identical and one sample has an identical set of data that means variance is zero so, that will result in infinity.

Future Development

The current limitation of our project is the scarcity of the input variable, this project is currently dependent on the time series data of speed of the object at any instant. Thus the error in measurement of speed will greatly influence the result of the algorithms. So, for further development, the GPS Logger can be integrated with Gyroscope Sensor and Accelerometer of the phones. It helps to monitor the real-time movement of the object and provide more features to improve the accuracy of the output. The alignment and 3D acceleration will help in differentiating between the turns, stops and other unexpected anomalies (potholes, speed breakers, rumble strips etc.).

Another addition could be done in preprocessing by using the concept of Data Imputation or Map Matching approach to improve the results of GPS Logged data and hence greatly influence the results. Data Imputation will help us to improve the regularity of the data thus speed can be cross-checked with the computed speed using $\text{Speed} = \text{Distance} / \text{Time}$. It will also help to identify the acceleration at any instant of time. Acceleration of heavy vehicles and commercial vehicle is less as compared to light or private vehicles. Sometimes due to external factors like High rise buildings, Clouds, Heavy Traffic, Forest covers etc. the accuracy of GPS drops thus the reading of data is not correct. Thus to counter this problem, the map matching approach can be used. Map Matching will reduce the erratic behaviour of the GPS data and make it more reasonable by aligning the data with the Base map/Terrain or Road Network in Urban Environment.

The project can be trained as the Deep Neural Network and then can be developed into a standalone mobile/desktop application with GPS/Accelerometer/Gyroscopic Sensor integrated within them to make it full-fledged, ready to use application that directly presents the result to the user in a user-defined format(like .gpx , .kml , .csv etc.). The application can be further developed to present results in Real Time.

It can be transformed into a module that can embed on other applications like GIS applications, Navigation applications, etc.

Key Terms

Trajectory

The trajectory is the path followed by an object (here, a person) in motion. Here, a trajectory is the set of points that refer to the position of the person at any instant of time.

Trajectory τ is the set of sequence of points P_i , such that

$$\tau = [P_i : P_i = \{ x_i, y_i, \langle z_i \rangle, T_i \}]$$

where, $T_i < T_{i+1}$

x_i and y_i are the coordinates of the object in Geographics coordinate system at any instant T_i

Gaussian Membership Function

Gaussian Membership Function defined by a central value m and a standard deviation $k > 0$.

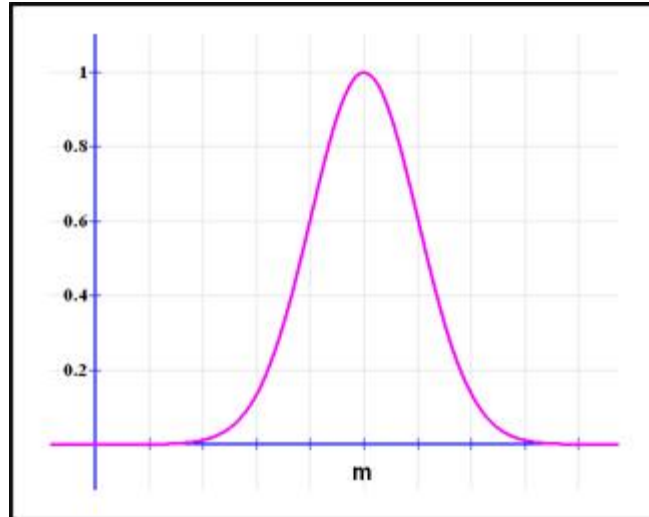
This function returns the value between 0 and 1. The closer the value is to mean m the output will be closer to 1.

$$\mu_A(x) = e^{-\frac{(x-m)^2}{2k^2}}$$

where, m is mean of distribution and

k is standard deviation of distribution

Figure 7 : Gaussian Function



GNSS

GNSS stands for Global navigation satellite system. It is the generic term for satellite navigation systems that provide autonomous geo-spatial positioning with global coverage. For Example, GPS (Global Positioning Service), GLONASS (Globanaya Navigazionnaya Sputnikovaya Sistema) (Russian version of the American GPS), and China's BEIDOU. India is also developing the IRNSS(Indian Regional Navigation Satellite System) with an operational name of NAVIC(Navigation with Indian Constellation).

GIS

GIS stands for Geographic Information System. It is a system that is designed to capture, store, manipulate, analyse, manage and visualize the spatial or geographic data. For Example ArcGIS, QGIS, GRASS GIS are most popular.

Figure 8 : Trajectory Data Presented on Google Maps



References

1. Algorithmic Methods for Segmentation of Time Series

In this paper authors reviewed the various algorithms for Segmentation and analysed their comparative performances.

Authors : Miodrag Lovrić, PhD Marina Milanović, MSc Milan Stamenković
UDC 004.65.021:519.246.8

2. Automatic P-Phase Picking Based on Local-Maxima Distribution

In this paper, we referenced the idea of Mahalanobis Distance for comparasion of two distributions.

Authors : Costas Panagiotakis Eleni Kokinou Filippas Vallianatos
IEEE Transactions On Geoscience And Remote Sensing, Vol. 46, No. 8, August 2008

3. Automated Urban Travel Interpretation: A Bottom-up Approach for Trajectory Segmentation

In this paper, authors demonstrated the use of segmentation in geospatial data.

Authors : Rahul Deb Das Stephan Winter

DOI : 10.3390/s16111962

4. Mamdani Fuzzy Inference Method

In this page, author explained the Mamdani Fuzzy Inference System and its use case.

Author : Sanjay Krishnankutty Alonso

Link :

http://www.dma.fi.upm.es/recursos/aplicaciones/logica_borrosa/web/fuzzy_inferencia/main_en.htm

Abbreviation

GNSS : Global Navigation Satellite System

GPS : Global Positioning System

NaN : Not a Number

Inf : Infinity

MaaS : Mobility as a Service

FIS : Fuzzy Inference System

List of Figures

1. Flow Chart for Top Down Algorithm
2. Flow Chart for Bottom Up Algorithm
3. Flow Chart for Sliding Window Algorithm
4. Flow Chart for SWAB Algorithm
5. Block Diagram of Fuzzy Inference System
6. Distribution of Speed
7. Gaussian Function
8. Trajectory data presented on Google Maps