Mobile Phones as Sensors in the Determination of the Components of a Trip Chain without User Input

Nelson R. Gómez-Torres, PhD[1], and Didier M. Valdés-Díaz, PhD[2]
[1]Universidad del Turabo, Puerto Rico, ngomez42@suagm.edu
[2]Universidad de Puerto Rico at Mayagüez, Puerto Rico, didier.valdes@upr.edu

Abstract*– Given the high market penetration of GPS-enabled cell phones, these units can be used as sensors to obtain an enormous amount of information with the potential to improve the availability of data needed to implement the Dynamic Traffic Assignment (DTA) algorithms and procedures used to maximize the network's capacity utilization. Dynamic Origins-Destinations and Dynamic Mode Identification are two aspects of DTA that can benefit from the data gathered. The algorithms and procedures proposed in this paper consist of using GPS-enabled mobile phones as sensors (probes) to determine origins, destinations and modal components of trips without user input. The main algorithm is called the Mode Identification Algorithm (MIDA), designed to identify origins and destinations of the following modes: pedestrians, motorized vehicles and heavy rail. MIDA does not identify trips on buses and it was developed to identify heavy rail trips only on single route networks. MIDA is divided into two components, Identification of Heavy Rail (IDHR) under the limitation previously described and Identification of Stop, Walk and Motorized vehicles (IDSWAM). The IDHR component relies heavily on positioning, while the IDSWAM component is a fuzzy algorithm that relies on speed, direction and consistency. The algorithm was tested with data gathered on the field. MIDA was tested in its identification of several modes, and it showed to be capable of doing so. Testing MIDA with data from volunteers showed an error of less than 6% in the identification of modes over the time period of the tests.

***Keywords*— GPS Tracking, Mode Identification, Origin-Destination Matrices, and Fuzzy Logic**

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2015.1.1.153
ISBN: 13 978-0-9822896-8-6
ISSN: 2414-6668

Mobile Phones as Sensors in the Determination of the Components of a Trip Chain without User Input[*]

Nelson R. Gómez-Torres, PhD[1], and Didier M. Valdés-Díaz, PhD[2]
[1]Universidad del Turabo, Puerto Rico, ngomez42@suagm.edu
[2]Universidad de Puerto Rico at Mayagüez, Puerto Rico, didier.valdes@upr.edu

Abstract*– Given the high market penetration of GPS-enabled cell phones, these units can be used as sensors to obtain an enormous amount of information with the potential to improve the availability of data needed to implement the Dynamic Traffic Assignment (DTA) algorithms and procedures used to maximize the network's capacity utilization. Dynamic Origins-Destinations and Dynamic Mode Identification are two aspects of DTA that can benefit from the data gathered. The algorithms and procedures proposed in this paper consist of using GPS-enabled mobile phones as sensors (probes) to determine origins, destinations and modal components of trips without user input. The main algorithm is called the Mode Identification Algorithm (MIDA), designed to identify origins and destinations of the following modes: pedestrians, motorized vehicles and heavy rail. MIDA does not identify trips on buses and it was developed to identify heavy rail trips only on single route networks. MIDA is divided into two components, Identification of Heavy Rail (IDHR) under the limitation previously described and Identification of Stop, Walk and Motorized vehicles (IDSWAM). The IDHR component relies heavily on positioning, while the IDSWAM component is a fuzzy algorithm that relies on speed, direction and consistency. The algorithm was tested with data gathered on the field. MIDA was tested in its identification of several modes, and it showed to be capable of doing so. Testing MIDA with data from volunteers showed an error of less than 6% in the identification of modes over the time period of the tests.

Keywords—GPS Tracking, Mode Identification, Origin-Destination Matrices, and Fuzzy Logic

I. Introduction

The question posed in the title "Data, data, data – Where's the data?" by Tate-Glass et al. [1] motivated this paper. Tate-Glass et al. declared that research for Dynamic Traffic Assignment is active, but the data to apply it in real world situations is still under development or nonexistent. The solution proposed herein is the use of mobile phones as sensors (probes) to determine stops (origin and destination of a tour, or complementary stops) and mode components of a trip without user input. A new methodology was developed to use the data gathered using GPS-enabled mobile phones. Specifically, an algorithm that automatically identifies heavy rail, motorized vehicles, walking and stops components from mobile phone GPS data was designed and evaluated.

GPS capable mobile phones are able to gather and send GPS data of their positions in real time. Hellinga et al. [2] sets out that determining traffic conditions from positioning data requires five steps: map matching, path identification, probe filtering, travel time allocation and travel time aggregation. Map matching and path identification steps consist of determining the position of a vehicle or traveler in the transportation network and the possible path utilized to change positions. GPS data is a collection of position and times that may be gathered at a particular rate to identify a path taken. Both of these steps can be done with available technology/methods. The challenge lies in probe filtering and travel time allocation.

Probe filtering consists of determining the transportation mode being used. The data have to be analyzed to establish the transportation mode. Speed and routes are the principal variables to observe. When several links are used between data points, we need to estimate the travel time on each link. Travel time allocation refers to this estimation. Methods to estimate the travel time of each link between data points are also included in this step. In addition, when a change in travel mode occurs, we need to separate the travel time in each mode. This is also related to the time interval between points.

The solution proposed herein is the use of GPS-enabled mobile phones (also called cell phones) as sensors (probes) to determine stops (origin and destination of a tour, or complementary stops) and mode components of a trip without user input. Previous research focused on obtaining the average speeds of vehicles [3], relied on user input to determine modal split [4] or relied on accelerometers to determine whether the user was walking, running, biking or riding a vehicle [5]. Recent developments show advances in the area of automatic mode estimation. In the case of González et al. [6] their neural network algorithm works only for single mode trips with the users informing the critical points of the trip. Another case is the study by Zhang et al. [7] that integrated GPS data from users, itinerary of users and GPS data from buses to identify bus trips.

To determine stops (origin and destination of a tour, or complementary stops) and mode components of a trip without user input this paper uses the Mode Identification Algorithm (MIDA). MIDA is an algorithm designed by Gómez-Torres [8] utilized to identify the origins and destination of trips, and several transportation modes. The modes identified by MIDA are pedestrians (Walk), motorized vehicles (Car) and heavy rail (HR). From the research done, to identify buses, GPS data directly from the buses seems to be required [8].

[*] This study was performed with the financial aid of the National Science Foundation (NSF) through its IGERT (Integrative Graduate Education and Research Traineeship) program, the Civil Engineering and Surveying Department of the University of Puerto Rico at Mayagüez and the Program of Civil Engineering in the Department of Engineering and Informatics Technology of Caribbean University at Bayamón..

13th LACCEI Annual International Conference: "Engineering Education Facing the Grand Challenges, What Are We Doing?"
July 29-31, 2015, Santo Domingo, Dominican Republic **ISBN:** 13 978-0-9822896-8-6 **ISSN:** 2414-6668
DOI: http://dx.doi.org/10.18687/LACCEI2015.1.1.153

According to the Electronic Privacy Information Center [9], individuals must opt-in (not be entered by default) and be informed that data (in this case GPS data) is being collected. The data must be unidentified and protected, not only from criminal use, but also from misuse by those collecting it. Self-regulation is common for companies dealing with private personal data. However, comprehensive legislation must be enacted to legally protect the privacy of individuals, instead of self-regulation. Privacy issues must be evaluated deeply, but this paper will not deal with those issues.

This paper will show the evaluation of the MIDA algorithm with data gathered in Puerto Rico. The data was gathered by the principal researcher and by five (5) volunteers. The data was gathered while driving cars, walking, riding a single route heavy rail system and during stops.

II. The Data and Its Sources

In recent years, studies on the use of mobile phones for traffic studies are becoming more common due to the characteristics and availability of mobile phones. The characteristics that seem to be capable of helping to make traffic studies are:

1) Sample size: USA population was about 307 million people by July 2009 [10]. According to the Cellular Telecommunications Industry Association (CTIA) there are 270.3 million of wireless subscribers in the USA alone. They indicate that the wireless penetration is 87% of the total USA population [11].

2) Development in location strategies: Mobile phone providers in the U.S. are increasingly using GPS to detect the location of their mobile phones if a call to 911 is made. This responds to the FCC's E9-1-1 rule that require the mobile phone providers to be able to locate their mobile phones with an accuracy of at least 50 to 300 meters [12].

3) Usage pattern: Most mobile phone users carry them everywhere and do not turn the mobile phone off during trips.

The data was gathered with three (3) different mobile phones a Nokia E71, a Nokia 5800 and a Pharos Traveler 619. All of these mobile phones have integrated GPS and were connected to the T-Mobile network of Puerto Rico. The data gathered with the mobile phones consisted in locations (latitude and longitude) and times. Each data point consists of a 3-tuple (time, latitude, longitude). Each data point was gathered at the fastest rate allowed by those mobile phones, which turn out to be between 1 and 2 seconds. The data used for this paper was gathered in Puerto Rico's San Juan Metropolitan Area. A GPS application for mobile phones was utilized to gather the data and was saved in the phone itself. It is also worthy to point out that the trips where done with reality in mind, but trying to travel near the heavy rail route in other transportation modes to challenge the algorithm as much as possible.

Several transportation modes were tested in this study (Walk, Car, Bus and HR). Origins and destinations (Stops) were tested too. The data was taken in the form of designed multimodal trips and "real world" data gathered with the help of volunteers.

The multimodal trips were randomly designed to have changes from all modes to all modes. These trips are not representative (typical) trips. The multimodal trips were designed to have all modes and were intentionally designed to be near the route of the heavy rail system (Tren Urbano) of the San Juan Metropolitan Area (SJMA) of Puerto Rico. Four (4) multimodal trips were designed, two (2) for development of the algorithm and two (2) for internal validation.

The "real world" data was gathered with the help of five (5) volunteers. The volunteers were asked to record GPS data on their mobile phones and keep a log of all the changes in mode made by them during one week.

A. Selecting Data Collection Rate

Speed is one of the most important parameters to identify the transportation mode used. The speeds in this section were calculated with 2 points, where each point contains the information of latitude, longitude and time. To calculate speeds, the distance between points was calculated using the geodesic distance of the earth ellipsoid [13]. Elevations were ignored in the calculations of speed. The speed is calculated with the following equation:

$$S_1 = \frac{\text{Geodesic\_Distance\_Between}(x_1, x_2)}{\text{Time\_Between}(x_1, x_2)} = \mathrm{f}(x_1, x_2) \qquad \textbf{(1)}$$

Ignoring the elevations becomes a systematic error, but its magnitude is very low. Considering a road segment with a slope of 10% (which is a rather high percentage), a vehicle moving at 100.5 km/h (as measured in an instant) will appear to be moving at 100 km/h if the height is ignored. Also with the current GPS technology available in mobile phones the errors in elevations are high; as a result of both of these factors it is better to ignore the elevations.

Also, we wanted to determine the rate that should be used to gather data without losing important information. To that purpose Figure 1 and 2 were constructed. The first data point (or Data Point 1) will be called DP_1 and the second DP_2, therefore the n data point will be called DP_n. This test was performed on May 24, 2010. In Figure 1, the curve marked as 1 denotes speeds calculated with each data point, where the first speed marked is calculated with the 2-tuple (DP_1, DP_2) and will be denoted as S_1. The second speed marked, S_2, is calculated with (DP_2, DP_3), therefore in general the following relation applies:

$$\mathrm{f}(DP_n, DP_{n+1}) = S_n; \forall n \qquad \textbf{(2)}$$

Around data point 815 (Figure 1 and Figure 2) an unusually high speed (120 mph) was found and it was related to a data point with an unusually high error.

In Figure 1 the curve marked as 10 denotes speeds calculated with each data point where the first speed marked is calculated with the 2-tuple (DP_1, DP_{11}) and will be denoted as S_1. The second speed marked, S_2, is calculated with (DP_2, DP_{12}), therefore in general the following applies:

$$\mathrm{f}(DP_n, DP_{n+10}) = S_n; \forall n \tag{3}$$

Therefore, in Figure 1, the relations for the curves marked as 30 and 60 appear in equations (4) and (5).

$$\mathrm{f}(DP_n, DP_{n+30}) = S_n; \forall n \tag{4}$$

$$\mathrm{f}(DP_n, DP_{n+60}) = S_n; \forall n \tag{5}$$

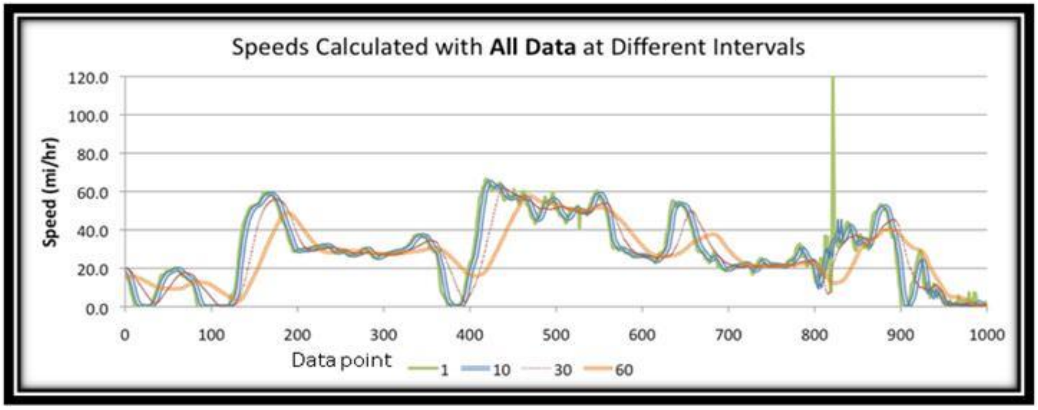

Figure 1 Speeds calculated using all the data

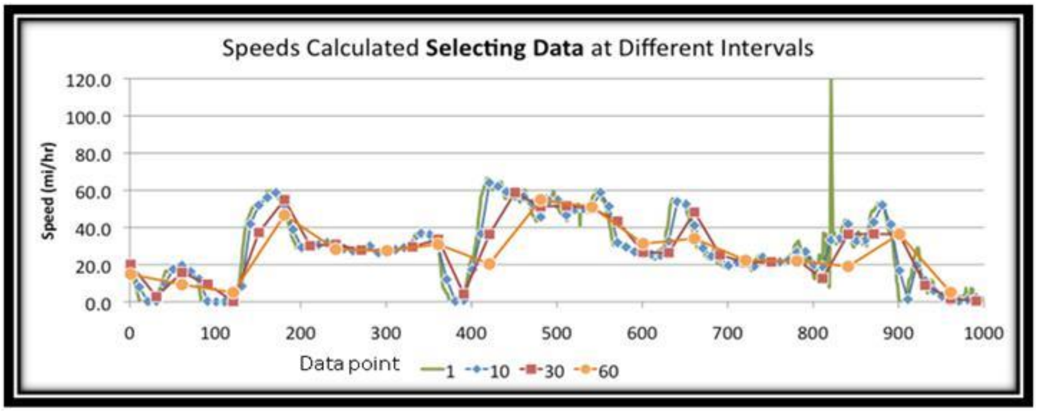

Figure 2 Speeds calculated with a selection of data

For Figure 2, the relations are similar, but not all data points were used. The curve marked as 1 is exactly the same as the curve marked 1 in Figure 1. The first speed, S_1, of the curve marked as 10 is calculated with the 2-tuple (DP_1, DP_{11}), the second speed, S_2, with the 2-tuple (DP_{11}, DP_{21}); therefore the following relation applies:

$$\mathrm{f}(DP_n, DP_{n+10}) = S_i; n = 1 + 10(i-1); \forall i \tag{6}$$

Therefore, in Figure 2, the relations for the curves marked as 30 and 60 appear in equations (7) and (8).

$$\mathrm{f}(DP_n, DP_{n+30}) = S_i; n = 1 + 30(i-1); \forall i \tag{7}$$

$$\mathrm{f}(DP_n, DP_{n+60}) = S_i; n = 1 + 60(i-1); \forall i \tag{8}$$

Comparing Figure 1 and Figure 2, we can see that at different intervals not all the data is needed. The shapes of both speed charts show that we were able to select greater intervals without losing critical information. In the Figure 2, the curve marked as 10 (skipping 10 data points) seemed to be adequate (by visual inspection); therefore, gathering data at an interval of between 10 and 20 seconds seemed to be appropriate (remember that each data point is gathered every 1 to 2 seconds). Even gathering data at an interval of 30 data points (every 30 to 60 seconds) is possible without losing most of the details. Also, considering the Nyquist-Shannon sampling theorem makes it difficult to justify a sampling rate of more than 60 seconds since it is possible for a modal component of a trip to be less than 2 minutes.

III. THE MODE IDENTIFICATION ALGORITHM (MIDA)

This section deals with the procedures and ideas applied to create the Mode Identification Algorithm (MIDA). MIDA is an algorithm designed to identify the modal split of a trip with GPS information gathered from mobile phones (or any other GPS enabled device). MIDA intends to classify between stops, walk, motorized vehicles, and heavy rail. MIDA has two components, IDHR and IDSWAM. IDHR is the Identification of Heavy Rail and IDSWAM is the Identification of Stops, Walking and Motorized vehicles.

A. Identification of Heavy Rail

The component of MIDA designed to identify the use of heavy rail in a trip (from now on IDHR) will be shown in this section. The IDHR component relies on GPS data (gathered from mobile phones) and polygons representing the heavy rail system (in this case the Tren Urbano of Puerto Rico). Even though the polygons were done with the Tren Urbano system in mind, the principles applied seem to be appropriate to similar systems. The polygons and the principles to make them will be shown in the next subsection, followed by a subsection explaining the IDHR component.

A.1. Representation of the Heavy Rail System

The Tren Urbano system is a heavy rail system that provides transportation in the San Juan Metropolitan Area in Puerto Rico. The system consists of sixteen (16) stations with two (2) of those stations underground. Two (2) segments of route are aboveground and one (1) segment underground, therefore there are two (2) transitional areas (from aboveground to underground). Figure 3 shows examples of all the types of polygons designed to identify heavy rail.

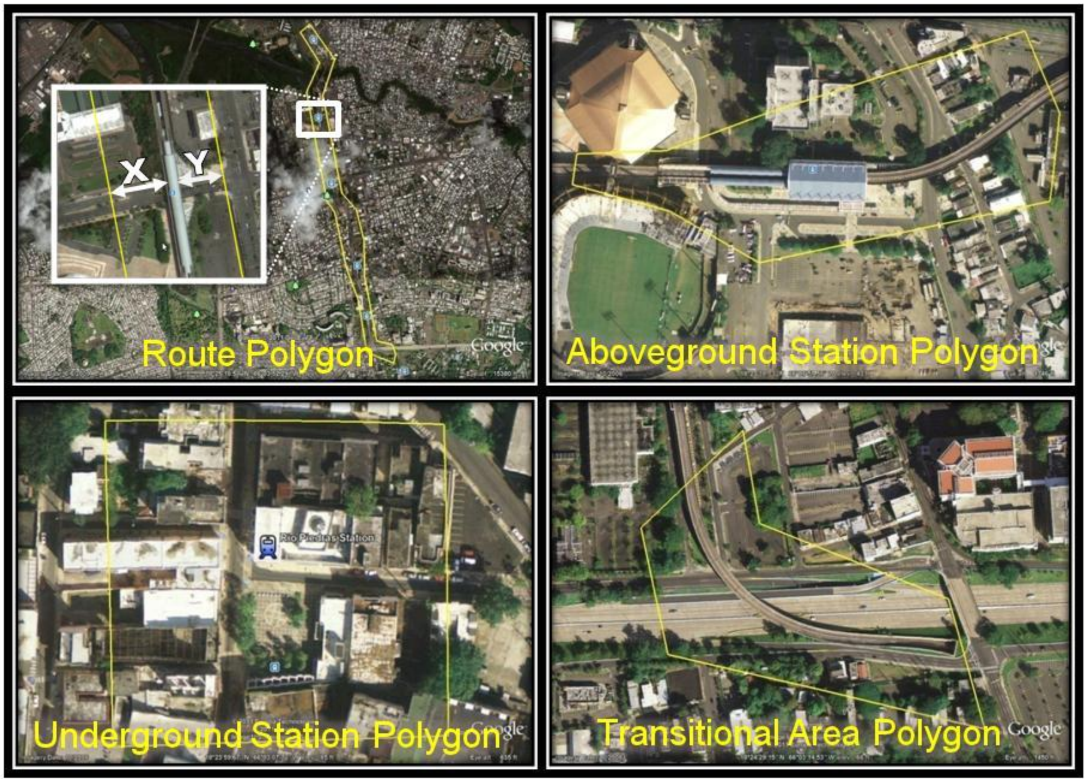
X
Y
Route Polygon
Aboveground Station Polygon
Rio Piedras Station
Underground Station Polygon
Transitional Area Polygon
Google

Figure 3 Types of Polygons

Designing the polygons require experience and understanding of the precision error of the GPS. GPS data needs to be observed on the map to be able to shape the polygons. The size and shape of the polygons depend on the position of the station (or route), nearby buildings, nearby roads, type of station and GPS error. The researcher tried to make the polygon as small as possible, while taking into consideration the precision error of GPS. For IDHR to be effective, careful consideration must be applied in the design of the polygons. The next subsection explains the IDHR component of MIDA.

A.2. IDHR Algorithm

The IDHR component of the MIDA algorithm is a function of GPS data (latitude, longitude and time) and the polygons discussed in the previous section. The concept of IDHR consists on identifying each heavy rail segment of a trip. A heavy rail segment is defined as traveling in a train between two consecutive stations. Figure 4 shows a simplified flowchart of the IDHR component of MIDA. When the GPS data show that a station is reached, the IDHR component verifies if it travelled through the route polygon from another station.

The flowchart in Figure 4 is composed of six (6) sections. Each of those sections does a specific job within the IDHR algorithm. Information on the details of each section follows:

- Section 1 – This section tests GPS data to find if it is located inside a station polygon (primary station). This section may end the algorithm or continue to section 2.
- Section 2 – This section searches GPS information back in the database until another station is found. To continue searching back GPS data cannot leave routes or transitional areas. This section may go to section 6, instead of section 3, if another station is not found. Another station must be found (without exiting the boundaries) to be able to establish HR mode occurred.
- Section 3 – This section declares that Mode is heavy rail (Mode = HR), after the appropriate conditions are met. The "T" represents the quantity of points required to add at least 1.25 minutes. This section always goes to section 4.
- Section 4 – This section checks if the secondary station is an underground station. For underground stations the next section is section 6. For aboveground stations the next section is section 5. If the secondary station is an underground station, then the last data point found inside the underground station polygon will be marked as the beginning of the HR leg of the trip. If the secondary station is an aboveground station, the first data point found inside its polygon is considered the beginning of the HR leg of the trip.
- Section 5 – In this section all GPS data that remains inside the secondary station is declared as heavy rail. This section always goes to section 6.
- Section 6 – This section searches for the next position that is outside the primary station. This section may terminate the algorithm or go to section

There are rules applied in the algorithm that help to avoid over and under identification. First as soon as an initial station is reached heavy rail mode begins, but ending "T" data points after reaching the last station (see section 3 of Figure 4). That "T" provides time for the user to exit the station and ending the heavy rail mode. "T" can be adjusted for each station individually. The adjustment can be done by observing the average time it takes users to exit the station. Therefore, evaluating each station individually with data from the general public is required for the final implementation. For the purpose of this study 1.25 minutes was selected as an average for "T" and was not tailored to each station.

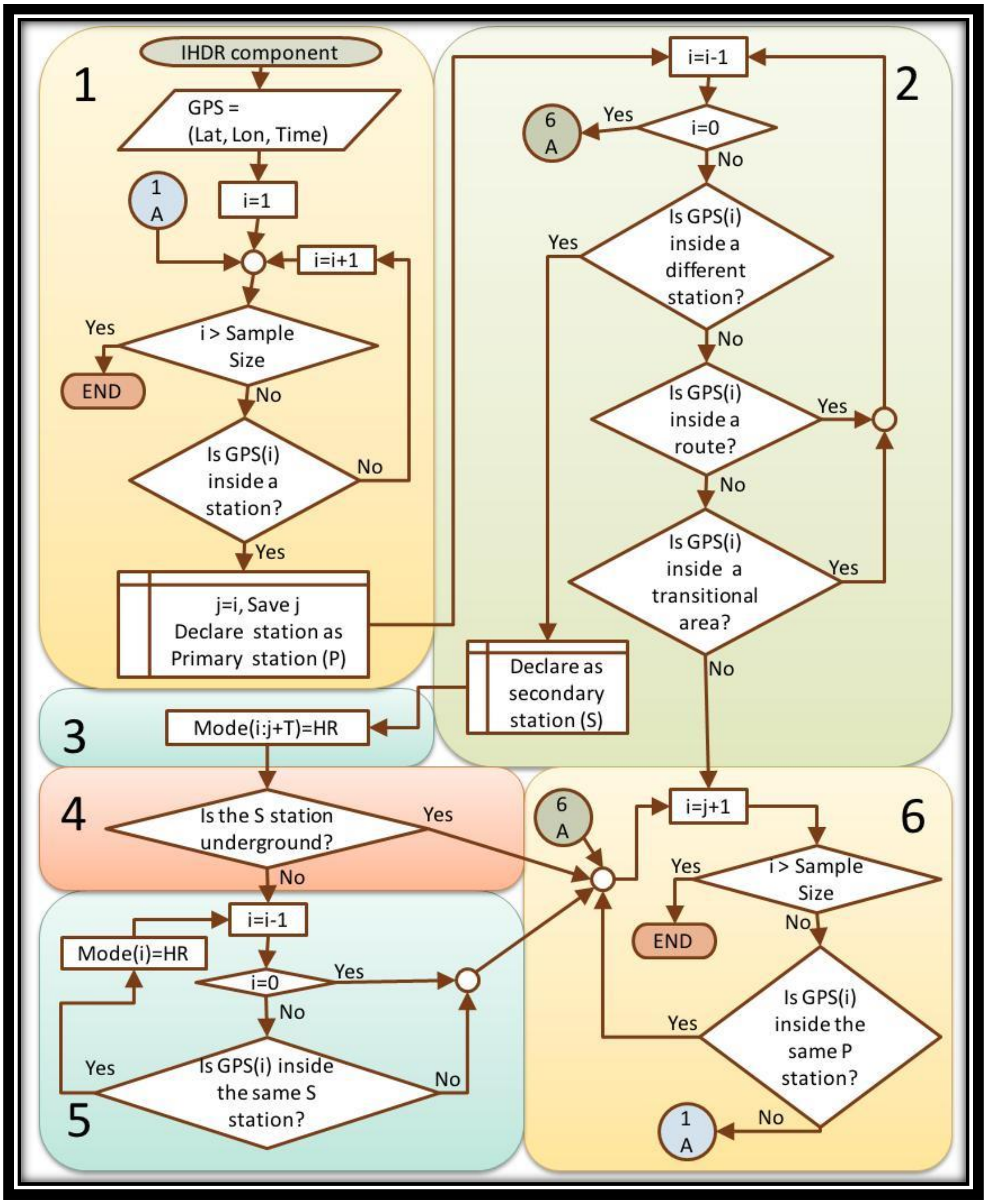

Figure 4 Simplified flowchart of the IDHR component

The second rule is that the first data point found in an underground station (USTU) becomes the start of heavy rail mode (see section 4 of Figure 4). This is done because there is no GPS connectivity inside an underground station. Notice that the first data point found is the last time when the mobile phone GPS is available before entering an underground station.

The third rule is that each leg is dealt with separately, therefore if one data point is outside of the appropriate polygons (see section 2 of Figure 4), only part of the trip will be missing. If a trip requires several legs, the legs are unified automatically by the IDHR component of the MIDA algorithm.

A.3. Identification of Stops, Walk and Motorized Vehicles

The difference between stop, walking and motorized vehicles (in terms of GPS data) is the travelling speed. Therefore, in order to identify them, the researcher needed to evaluate them together. The identification of stop and walking (without motorized vehicles) was attempted, but it failed when vehicles travelled at low speeds. Therefore all three modes were identified at the same time with the approach described in this section.

As seen in Figure 5, GPS data is not precise enough to ensure that a person at rest (stopped) will appear to be travelling at exactly 0 mph. At the same time walking speed does not exceeds 5.7 mph (9.1 km/h) and the lowest possible speed can be practically 0 mph, being the lowest comfortable speed for any group of age/gender 2.85 mph (4.6 km/h) [14] Motorized vehicles speed range from 0 mph (0 km/h) to more than 70 mph (113 km/h). Hence, the three modes overlap at low speeds. At the same time, Figure 5 shows that a higher speeds the effect in the velocity vector of the GPS precision error is reduced.

The IDSWAM component of MIDA relies on fuzzy logic. Boolean logic deals with true or false statements, but fuzzy logic deals with degrees of truth and is utilized in complex decisions. Stops, walk and MV have similar speeds in the low end of the spectrum. Also, the precision error may increase the calculated speeds of stops and walk. Those facts start showing the picture of a complex decision.

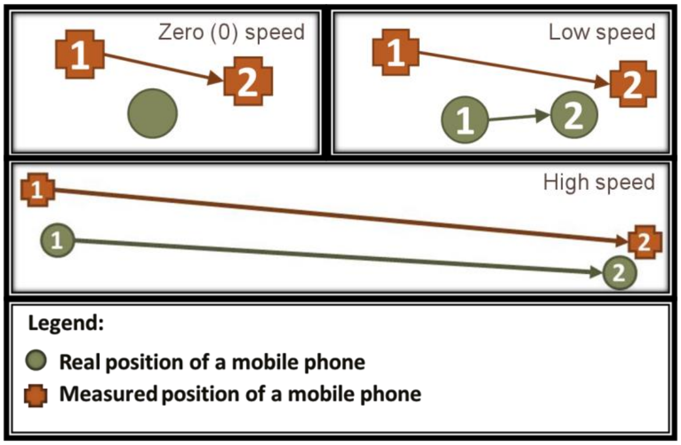
Zero (0) speed
1
2
Low speed
1
2
1
2
High speed
1
1
2
2
Legend:
● Real position of a mobile phone
■ Measured position of a mobile phone

Figure 5 GPS precision error

Figure 6 shows the flowchart of the IDSWAM component of MIDA. After applying a series of fuzzy rules to the GPS data, the results are the memberships of each data point in terms of stops, walk and MV. Then those memberships pass through the defuzzification process. The defuzzification process consists of averaging the memberships of 5 data points and selecting the classification with the highest membership. In order to improve consistency, a factor based on continuous classification is given to the memberships. The last step is the stabilization process, where the most frequent classification of 5 data points is declared. The last step is intended to aid the IDSWAM component to provide steady classifications.

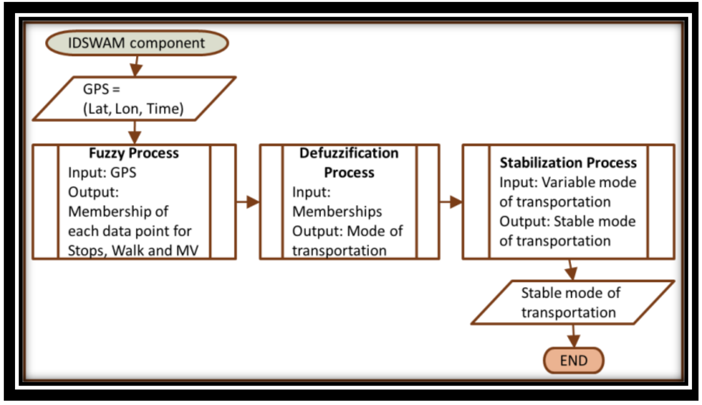

Figure 6 Simplified flowchart of IDSWAM

One of the factors that aid in the identification is the direction of the velocity vector. The direction during a stop (if data is available), tends to be erratic (see Figure 5 again). In the case of the direction of MVs, it tends to change slowly. Still, when MVs are waiting at intersections their perceived direction can change abruptly (precision error).

A mobile phone registered at high speeds for a long time, can be identified as an MV. Also, a mobile phone registered at low speeds with erratic velocity and direction for a long time, can be identified as being stopped. When trying to use Boolean logic one of the problems is setting the values. How high should the speed be? How erratic should the direction be? This is when fuzzy logic becomes a useful tool.

A set of fuzzy rules relating the speed (in km/h) and the change in direction (in degrees) were done to design the membership functions for Stop, Walk and Motorized Vehicle (MV) as shown in Figure 7, Figure 8 and Figure 9, respectively.

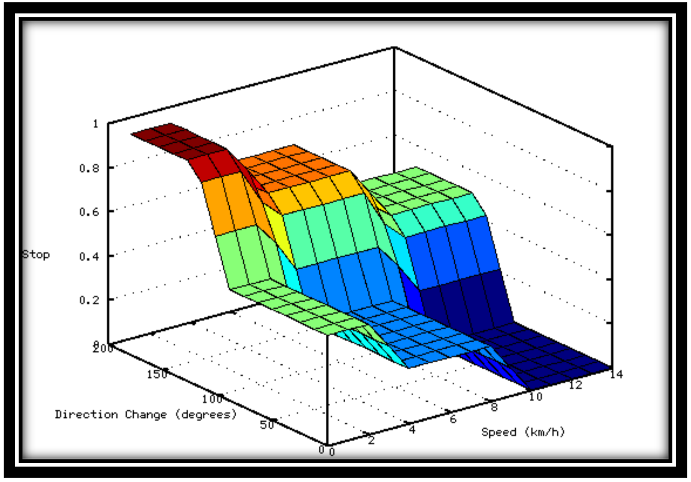

Figure 7 Stop membership function

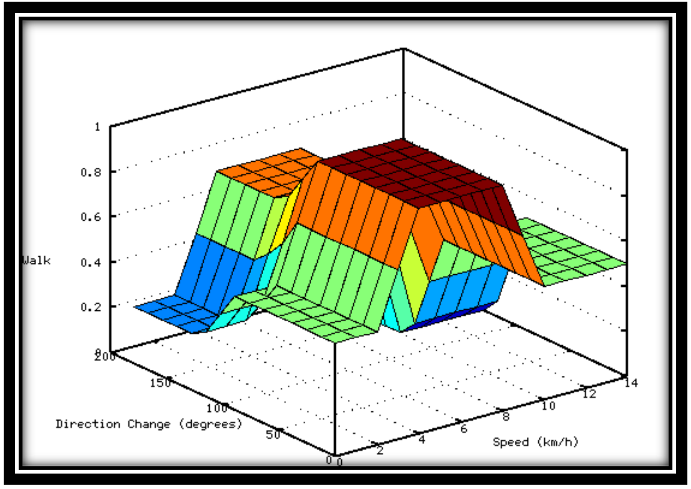

Figure 8 Walk membership function

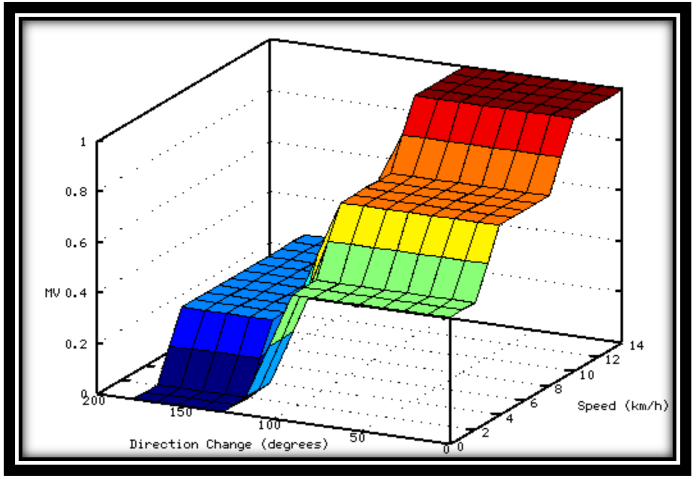

Figure 9 Motorized Vehicle (Car) membership function

IV. COMPUTATION OF ERROR

To understand the capacity of MIDA to correctly identify modes, five measurements of errors were used. These measurements are the following:

- Missed Identification (MID) Error – This error is the difference between observed time in a mode and the correctly identified time in that mode, divided by the observed time in that mode (see equation 9). MID error shows the proportion of time a mode is not identified in terms of all the time the mode was observed. In the plots this error is portrayed in black.

$$MID_{mode}=(Obs\ Time_{mode}–Correct\ ID\ Time_{mode})\ /Obs\ Time_{mode}\ (9)$$

- Over Identification (OID) Error – This error is the difference between all the time a mode is identified and the correctly identified time in that mode, divided by the all the time that mode is identified (see equation 10). OID error shows the proportion of time a mode is incorrectly identified in terms of all the time the mode was identified. In the plots this error is portrayed in red.

$$OID_{mode}=(All\ ID\ Time_{mode}\ –Correct\ ID\ Time_{mode}\ )/All\ ID\ Time_{mode}\ (10)$$

- General Missed Identification (GMID) Error – This error is the difference between observed time in a mode and the correctly identified time in that mode, divided by the complete time of the study (see equation 11). GMID error shows the proportion of time a mode is not identified in terms of the complete time of the study.

$$GMID_{mode}=(Obs\ Time_{mode}\ –Correct\ ID\ Time_{mode})/Complete\ Time\ (11)$$

- General Over Identification (GOID) Error – This error is the difference between all the time a mode is identified and the correctly identified time in that mode, divided by the complete time of the study (see equation 12). GOID error shows the proportion of time a mode is incorrectly identified in terms of the complete time of the study.

$$GOID_{mode}=(All\ ID\ Time_{mode}–Correct\ ID\ Time_{mode})/Complete\ Time\quad (12)$$

- General Error (GE) – This error is the sum of the GOID error (or GMID error) for all modes (see equation 13). GE shows the proportion of time any mode is incorrectly identified in terms of the complete time of the study. Notice that the sum of GOID or GMID provides the same results. This is due to the fact that in the same multimodal trip, all the instances that were identified incorrectly can be accounted as a Missed Identification (MID) of the observed (real) mode or an Over Identification (OID) of another mode.

$$GE = \sum GOID_{mode} = \sum GMID_{mode};\ for\ all\ modes\ (13)$$

The summary of the results obtained using MIDA of the multimodal trips (not "real world") are shown in Table **1**. The General Error (GE) is 16.3%. The GE, while useful to identify the level of the error for a single multimodal trip, it is more useful when dealing with data from volunteers or "real world" data. The GE is more useful with "real world" data because it represents the level of error expected from the widespread use of MIDA.

It is important to notice that the multimodal trips done by the researchers had more transferences from mode to mode than ordinary trips (as designed). Also, the multimodal trips were designed to be difficult to identify.

In terms of MID, OID and GMID, the percentage of error of Walk is the highest. This is also related to the transferences as well as the inherent difficulty of identifying walking. In the case of Stops (the second highest in OID and the highest of GOID), the error can be related to the fact that other modes include stops. Still, MID_{STOP}, MID_{CAR}, and MID_{HR} were under 19%, while OID_{CAR} and OID_{HR} were under 11%.

Table 1 Summary of MIDA results

Complete Data Set				
Mode	Missed Identification (MID %)	Over Identification (OID %)	General Missed Identification (GMID %)	General Over Identification (GOID %)
Stop	15.08	27.9	3.61	8.8
Walk	27.36	27.84	4.86	4.31
Car	18.55	10.88	4.45	2.25
HR	13.99	4.6	3.39	0.94
			General Error =	**16.3**

The results obtained using MIDA with data from volunteers will be provided for each volunteer. The five (5) volunteers that gathered data were identified with a number between 1 and 5. All of them live in Puerto Rico.

Volunteers 1 to 3 spend most of their time in the San Juan Metropolitan Area, while volunteers 4 and 5 spend most of their time in the West region of Puerto Rico. Four of the volunteers have a job. Four of them own a car.

Volunteers were between 25 and 35 years old and all of them own a mobile phone. Volunteers 1 and 2 used mobile devices with iOS (the OS designed by Apple Inc.), volunteer 3 used a Nokia E71 (Symbian OS) and volunteers 4 and 5 used mobile devices with Android (the OS designed by Google). From the comments of the volunteers, battery life of the Android devices seems to be longer, when compared to the other devices.

It is important to note that the General Error (GE) is the error for the whole day of data provided by the volunteers and none of them had a GE over 6%.

Volunteers stated that it was difficult to remember to record each change in mode. Volunteer 3 made more mistakes, recording changes in mode, than the rest of the volunteers. Volunteers 4 and 5 were noticeably better having only 10 mistakes between both of them.

Sometimes the volunteers remembered to report a change in mode a few minutes after the fact. Those estimations tended to be off, by more than 3 minutes. In one instance, volunteer 2 estimated a change in mode occurred 15 minutes before, but after revising the data in Google Earth the change occurred 30 minutes before. The most extreme case was with volunteer 3, in which a mode change estimated to have occurred 30 minutes before, was in fact one (1) hour before.

V. CONCLUSIONS

The main objective of this study was to develop a model for the estimation modal split using mobile phone without user-inputs. The Mode Identification Algorithm (MIDA) was developed to determine origins and destinations with the modal components of trips from mobile phone's GPS data. MIDA identifies origins and destinations (Stops), pedestrians (Walk), motorized vehicles (Car) and single route heavy rail (HR). MIDA does not identify trips on buses (Bus).

The identification of Stops (origins and destinations) had the most errors with the data gathered on the **multimodal** trips. Transfers between modes tend to have higher errors, therefore the error related to Stops were higher than with "real world" data. Data from volunteers showed a lower error related to Stops. A precision error associated to GPS technology was utilized to help identify Stops. But, at the same time errors identifying Heavy Rail were caused by the precision error.

The identification of Walk was more difficult due to the fact that when a transfer occurs, usually walking also occurs. Short walks occur frequently during the day. For most people those short walks are related to parking location. Identifying Car required a stabilization process to avoid misidentification (MID_{car}). Car may travel at very low speed or be stopped. The stabilization process at the same time increased the misidentification of short stops (MID_{stop}). The identification of Heavy Rail (HR) had a smaller error. This is due to the fact that Heavy Rail has its own route and stations.

REFERENCES

[1] M. Tate-Glass, R. Bostrum and G. Witt, "Data, data, data – Where's the data?," *Transportation in the New Millennium,* p. 7, 2000.

[2] B. Hellinga, P. Izadpanah, H. Takada and L. Fu, "Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments," *Transportation Research Part C: Emerging Technologies,* vol. 16, no. 6, pp. 768-782, 2008.

[3] Mobile Millennium. (2009). . (July, 2009), "The Mobile Millennium project," 2009. [Online]. Available: http://traffic.berkeley.edu/theproject.html. [Accessed July 2009].

[4] P. Winters, S. Barbeau and N. Georggi, "Smart phone application to influence travel behavior (trac-it phase 3)," Florida DOT, Florida, USA, 2008.

[5] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks,* vol. 6, no. 2, pp. 1-27, 2010.

[6] P. González, J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. Georggi and R. Perez, "Automating mode detection for travel behaviour analysis by using global positioning systemsenabled mobile phones and neural networks," *Intelligent Transport Systems, IET,* vol. 4, no. 1, pp. 37-49, 2010.

[7] L. Zhang, S. Gupta, J. Li, K. Zhou and W. Zhang, "Path2Go: Context-aware services for mobile real-time multimodal traveler information," in *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Washington, DC, 2011.

[8] N. Gomez-Torres, "Mobile Phones as Sensors in the Estimation of Modal Split without User Input," University of Puerto Rico at Mayagüez , Mayagüez, PR, 2012.

[9] EPIC, Electronic Privacy Information Center, "Public Opinion on Privacy," 2012. [Online]. Available: http://epic.org/privacy/survey/. [Accessed April 2012].

[10] U.S. Census Bureau, "Population," 2009. [Online]. Available: http://census.gov/main/www/popclock.html. [Accessed July 2009].

[11] CTIA, Cellular Telecommunications Industry Association, "Wireless Quick Facts," 2009. [Online]. Available: http://www.ctia.org/advocacy/index.cfm/AID/10323. [Accessed July 2009].

[12] FCC, Federal Communications Commission, "Enhanced 9-1-1 - Wireless Services," 2009. [Online]. Available: HTTP://www.fcc.gov/pshs/services/911-services/enhanced911/Welcome.html. [Accessed January 2009].

[13] S. K. Roy, Fundamentals of Surveying, Prentice- Hall of India Learning Pvt. Ltd., 2004.

[14] R. Bohannon, "Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants," *Age and Ageing,* vol. 26, pp. 15-19, 1997.