

# Using Smartphones and Sensor Technologies to Automate the Collection of Travel Data

**Tamer Abdulazim, M.Sc.\***

Phone: 416-273-6367

Email: [tamer.abdulazim@utoronto.ca](mailto:tamer.abdulazim@utoronto.ca)

**Hossam Abdelgawad, Ph.D.\***

Postdoctoral Fellow

Tel. 416-978-5049

Email: [hossam.abdelgawad@alumni.utoronto.ca](mailto:hossam.abdelgawad@alumni.utoronto.ca)

**Khandker M. Nurul Habib, PhD, P.Eng\***

Phone: 416-946-8027

Email: [Khandker.nurulhabib@utoronto.ca](mailto:Khandker.nurulhabib@utoronto.ca)

**Baher Abdulhai, PhD, PEng\***

Phone: 416-946-5036

Email: [baher.abdulhai@utoronto.ca](mailto:baher.abdulhai@utoronto.ca)

**\* Department of Civil Engineering**

**University of Toronto**

35 St. George St. GB105

Toronto, ON M5S 1A4

Fax: 416-978-5054

Number of Words: 6, 175 (text) + 1250 (figures and table) = 7, 425

Submission Date: March, 2013

*Forthcoming in Transportation Research Record*

## Abstract

This paper presents a data collection framework and its prototype application for personal activity-travel survey through the exploitations of smartphone sensors. The core components of the framework run on smartphones, backed by cloud-based services for data storage, information dissemination and online decision support. The framework employs machine learning techniques to automatically infer activity types and travel modes with minimum interruptions for the respondents. There are three main components of the framework: 1) a 24 hours location data collection, 2) a dynamic land use database, and 3) a transportation mode identification component. The location logger is based on the smartphone network and it can run for 24 hours with minimum impact on smartphone battery and equally applicable in places where GPS is available or not. The land use information is continuously updated from internet location services such as Foursquare. Transportation mode identification module is able to distinguish six different modes with 98.85% accuracy. Prototype application is conducted in the city of Toronto and results clearly indicate the viability of this framework.

## INTRODUCTION

---

Understanding the behaviour of transportation system users is crucial for implementing effective control strategies to alleviate congestion, improve transportation level-of-service and achieve better quality of life. However, traditional travel data collection methods are not respondent-friendly because they often require extra effort from survey participants to record and recall their activities, i.e., keep activity-travel diary. As a result, a multi-day personal travel survey is often run for a few days or at most weeks to limit the survey burden on the respondents. Such restriction makes traditional travel survey methods inappropriate for a long period data collection, which is necessary to capture variations in activity-travel behaviour that may not occur within a short period (e.g. recreational trips). Lack of accurate data of activity-travel patterns limits travel demand modelling capability. It also restricts the ability of the decision makers to understand the influence of various transportation policies on activity-travel behaviour. Recently, there has been an increasing interest in advancing data collection techniques by incorporating emerging technologies (1) (2) (3) (4) (5) to reduce survey burden and accurately collect location information, which are main concerns for seeking help of emerging technologies, e.g. GPS and smartphones. This paper is therefore geared to achieve this goal by investigating the applications of smartphone sensing technologies for continuous spatio-temporal activity-travel data collection.

This study is motivated by rapid advancements in smartphone capabilities, for example, having more sensors and processing power. We anticipate that such technological advancement would strongly impact survey methods and provide a more “respondent-friendly” survey design. Smartphone sensors can be categorized into three groups according to their application in travel data collection as follows:

- Motion sensors (6):
  - *Accelerometer*, measures the device linear acceleration,
  - *Gyroscope*, measures the angular rate of change (i.e. rotation velocity),
  - *Magnetometer* (i.e. compass), measures magnetic field strength.
- Location sensors:
  - *GPS* which is commonly used in outdoor settings,
  - *Network-based location services* which uses cellular network and Wi-Fi to determine the location (i.e. via triangulation)
- Ambient sensors:
  - *light sensor*,
  - *microphone*,
  - *proximity sensor*, which detects nearby objects and can indicate when the phone is near the user’s ear (e.g. during a call).

This research focuses on collecting data from travellers in a respondent-centric approach by considering their interests and capacity. Unfortunately, traditional data collection techniques that directly interact with participants are limited by factors such as respondent’s memory, cognition and interest to participate, which pose a challenge to the quality of collected data. These challenges accentuate the need for a user-friendly method of data collection that seamlessly integrates into travellers’ daily lives. The paper proposes such methodology for activity-travel data collection.

The remainder of this paper is structured into three sections as follows: the first section provides a literature review on the current data collection methods, with more emphasis on the applications of emerging technologies (e.g. GPS and smartphones). Second section discusses main components of the proposed travel data collection framework and how explains how they can be employed for practical data collection projects. The framework is then applied to three data collection experiments using smartphones and sensors to demonstrate the functionalities and capabilities of the framework. Finally, we summarize the paper by offering our conclusions and recommendations for future work.

## LITERATURE REVIEW

---

Finding innovative ways of interaction with transportation users to collect travel data is an active and growing research area, which generally aims to use emerging technologies to enhance data quality and reduce respondents' burden. In this paper, related studies are categorized into two groups as follows:

1. Applications of smartphones and modern technologies in travel surveys, particularly for collecting location data;
2. Inferring trip information from the data with emphasis on studies that automate transportation mode identification.

### *Data Collection Modes*

A number of studies has shown the potential of collecting travel data using new technologies (e.g. smartphones and GPS) to replace or supplement the traditional travel survey methods (e.g. travel diary) in order to enhance data quality and reduce survey burden (7) (8) (9). A recent review of applications of smartphones in data collection (1), argues that these innovative devices have a strong potential to advance transportation data collection and enhance system operation as well as long-term transportation planning. GPS-assisted data collection has been intensively investigated in the literature. An experiment was conducted using GPS for household travel survey showed that GPS data are more accurate than traditional computer-assisted telephone interviews (CATI) in terms of reported trip rate; also, the results indicated that GPS can reduce the non-response rate of young population but it might decrease the participation of minority population (8). Comparisons between GPS data and travel diary were attempted in a few recent studies such as (10) (2), in which the authors suggested that GPS should be used in household travel surveys, but special care should be given to data processing algorithms which extract trips from GPS logs. Also, factors such as GPS cold start (insufficient signal), forgetting to carry the device and elders population should be carefully considered when using GPS-based surveys. GPS traces were also used to identify active transportation modes and to assess health impacts of having a nearby transit option (9). For long-term planning purposes, GPS was used to capture travel behaviour changes overtime (7). Further, it was found as a plausible technology to evaluate the effectiveness of travel demand programs and policies (11). GPS-assisted survey can be followed by a prompted survey to collect further data (12).

In addition to collecting travel behaviour data at the micro level, GPS devices and GPS-enabled smartphones have been used to collect macro data for an entire city or a roadway network; such data could be used to monitor traffic, and to observe overall transportation dynamics and movement patterns in a city. For example, Mobile Millennium is a recent project that demonstrated the benefit of GPS-enabled smartphones in real-time traffic monitoring that is currently implemented in San Francisco and the Bay Area (13). WikiCity is another recent project that captures, in real-time, city dynamics and movement patterns of the people (14), which is currently implemented in Rome, Copenhagen and Amsterdam to visualize real-time people and transportation movement, also areas of intense crowding (e.g. massive gathering event).

In summary, the potential of GPS and smartphone applications on the transportation system cannot be overstated, with tangible benefits for travel behaviour modellers, transportation planners, system operators and decision makers. There are still a few technical challenges of using GPS-only data collection such as signal disturbance, in-door navigation and tunnels (e.g. subway), high power consumption that can limit collecting 24 hours location data. Moreover, for a dedicated GPS device (i.e. not a GPS-enabled smartphone) there are additional challenges as people will have to carry an extra device. Also, it is not commonly possible to run a customized application on GPS devices (e.g. to process data or display personalized information), which restricts the communication with travellers and force data processing to be executed in centralized back-end servers that might encounter scalability issues especially in large-scale data collection projects.

### *Inferring Travel Activity*

Generally, the goal of travel surveys is to collect trip information attributes such as location, start and end time, purpose, transportation mode and route. These attributes can then be correlated with travellers' socio-economics attributes such as age, income, employment and gender to build behavioural models. To automate the data collection process, some of these attributes should be extracted from the raw data (e.g. GPS traces) to a certain degree of accuracy. Two main approaches have been applied in the literature for data processing: algorithmic and heuristic, and learning-based approaches (i.e. machine learning and data mining). The algorithmic and heuristic approach typically applies predefined deterministic procedures or rules to extract trip attributes; unlike the learning-based approach, in which such rules are learned automatically from data. For example, to determine transportation mode, a heuristic approach may use a rule such as "if the average speed below 5 km/hr, then the mode is Walking", while a learning-based approach will infer such rules from data.

Algorithms and heuristic approaches are useful to automatically extract some trip attributes from raw data and to expedite the data post-processing. Despite the terminological difference between them, both approaches are often combined together; thus for the context of this paper they are considered to the same. Heuristic rules about speed and duration of stay at a location were used to automatically identify transportation mode and trip purpose/activity from GPS traces in combination with GIS database that contains network information (15) (16) (17). These rules are commonly derived from domain knowledge about the transportation system, land use information of a study region, or from previous related studies. This approach is not only ad

hoc but also difficult to validate; additionally such rules are often tightly coupled to the environment they were derived from (e.g. transportation system of specific region) and probably cannot be transferred or generalized. Moreover, developing a generic algorithm to classify activities based on collected data is a tedious task and generally not feasible when fine classification is required, in which case a learning-based approach would be more appropriate.

In the learning-based approach the relation between predictors (i.e. independent variables) and the target/response variable (e.g. transportation mode, activity classification) is learned by applying machine learning techniques to a training dataset that contains both predictors and target variable. A learning-based approach was proposed to automatically predict some of the activity attributes (e.g. location and duration) and fill the survey automatically, which help in reducing respondents' burden (18). For transportation mode detection, speed and acceleration values have been frequently used as predictors. Such data sometimes were either collected from a dedicated GPS device (19) or using a smartphone-based GPS in which case a pre-processing algorithm was implemented in the mobile device (20). Classification using Multilayer Perceptron Neural Networks (MLP-NN) was implemented in (19) (20), while Support Vector Machine (SVM) was the technique used in (21). Both techniques have advantages and disadvantages. For NN, the training time, to fit the data, is normally longer than SVM; because NN solves a non-convex optimization problem while SVM solves a convex optimization problem which can be optimized much faster with efficient algorithms. However, NN provides a probabilistic output (i.e. probability of the training case belonging to each possible target classification) which is useful if this output is processed by further algorithm to produce the final classification decision (e.g. incorporate risk factor associated with misclassification or other domain knowledge). On the contrary, SVM outputs the final classification (22). Recently, a hybrid-approach was implemented in (23) to combine learning-based and rule-based to identify transportation mode at the trip level.

The above studies classify a limited number of transportation modes which might not be sufficient for travel behaviour modelling. For example, In (20) only walk, auto and bus were classified; walk, run, bike and auto in (21); walk, auto, street car and bus in (19); finally, in (23) stand, walk, bike and motorized were classified then heuristics were applied to further classify the motorized mode to drive, bus or train.

Different features can be used as input for the learning algorithm. For example, acceleration values can be used directly as input to the learning algorithm (19) (20) or after a feature extraction process such as Fast Fourier Transform (FFT), mean and variance (21) (23). The first approach depends on the variability in the magnitude of acceleration values to distinguish between transportation modes. In other words, the learning is performed in time domain in which there are dependencies between data points. The second approach transforms the data into a feature domain (e.g. frequency domain) where the learning take place. On one hand, for similar modes the variability in the frequency domain might not be sufficient for classification. On the other hand, analysis in the time domain is more complicated because the dependency between values violates the assumption that data points are independent and identically distributed (which is assumed by many learning algorithms including NN and SVM). The Learning-based approach has been proven to be effective for identifying transportation mode; however, more research is required to fully utilize this approach to allow for a fine

classification of transportation modes and activities.

Review of current literature shows that GPS-based data collection has been intensively investigated, but only a limited number of studies have incorporated smartphones' sensors (e.g. accelerometer). Also, the classification of activity-travel data has been limited to the identification of transportation mode, route, and coarse activities (e.g. work, home, school) while more fine classifications of transportation modes and activities have not been yet addressed. Further, many of the classification techniques applied in the literature suffer from the following limitations: 1) lack of rigorous techniques that can be validated (i.e. unlike heuristic rules), 2) classification that depends intensively on specific system information (i.e. certain road network or transit route), and 3) data collection that heavily depends on GPS which is difficult to be practically used for 24 hours due to the high power consumption and loss of signal in indoor settings including subways.

This paper therefore introduces an integrated framework that uses smartphones and sensors technologies to pervasively collect travel data. We developed the framework to achieve three essential requirements:

1. Integrate various smartphone sensors into the collection travel data to increase its quality and quantity;
2. Automate the identification of travel activities and transportation modes, to a fine level of detail, in order to reduce respondents' burden; and
3. Have a battery-friendly operation to practically collect 24 hours travel data for long-term travel surveys.

The next section discusses the tools and techniques used to achieve the above requirements.

## THE TRAVEL DATA COLLECTION FRAMEWORK

Automating the activity-travel data collection process is a challenging task due to the diversity and complexity of travellers' behaviour. However, we envision this process to be broken down into three non-trivial core tasks 1) continuously capture location data, 2) discovering the nearby land use characteristics, and 3) identify the transportation mode. FIGURE 1 shows a schematic diagram of the framework and data sources. The smartphone application, which implements the framework, continuously collects location data and when the location does not change for a configurable threshold (e.g. 5 minutes), it indicates an activity starts. If the location is a known location (visited before), the application will just record the start of a visit to this location, otherwise it will run the land use discovery module to collect information about nearby location that can be used to classify the current new location where the user at. After obtaining the best classification of the current location, the user will be prompted to verify or correct the current location, which will be stored afterwards as known location. When the application recognizes that the location is constantly changing, it will start the transportation mode detection module to classify the current mode of travel. Also, GPS can be used to collect accurate route information if the route is new.

### *Capturing Location Data*

While GPS is the most known location sensor, location can be sufficiently determined using network-based triangulation with cellular network and Wi-Fi, which is battery-friendly and seems to provide good accuracy in dense urban areas and for locations that are frequently visited (e.g. work, school and grocery store). Network-based location services are available via many providers such as (24) (25). These providers typically have a large database that contains numerous geocoded Wi-Fi networks and cellular towers which is used later to determine the smartphone location.

Network-based location services often require the smartphone to have Internet connection to determine the location. To reduce the dependency on Internet, our location logger module implements the same concept by maintaining a small-scale database of “landmarks” such as cell id and Wi-Fi network for frequently visited locations. For example, when the individual is at work, information about cell id, available Wi-Fi network and Bluetooth devices (e.g. Bluetooth printer) are stored along with GPS coordinates; afterwards, just by detecting these landmarks the location will be recognized and the associated GPS coordinates can be used even if there is no GPS signal or Internet connection. Further, by observing known landmarks such as Wi-Fi networks, the application might be able to indicate the arrival and departure time from known locations; for example, the time of detecting a known Wi-Fi network can be an indication of arrival time to this location and similarly when Wi-Fi connection is lost might indicate departure time. Moreover, at some public places (e.g. café and restaurants) Wi-Fi network might have the business name of that place which would help in determining the activity. Network-based location is normally less accurate than GPS for in outdoor environment and does not provide speed. However it is still an attractive alternative or complementary to GPS because they are: 1) battery-friendly, 2) and work in indoor settings where no GPS signal; yet, they have not been fully utilized in transportation research (26). In the experiment setup section, we will discuss more details on how we incorporated network-bases location in data collection.

### *Discovering Land Use Characteristics*

The output of the location logger is the current smartphone coordinates along with the accuracy of the location. In order to infer traveller’s activity, more information about the current location is required particularly the land use characteristics including residential buildings, schools, parks and shops. The land use discovery component is responsible for collecting such information given the current coordinates and accuracy. Accuracy is important for network-based location because it can vary from about 10 meters to 1000+ meters depending on the existence of landmark such as Wi-Fi networks. Consequently, the land use discovery needs to be aware of the accuracy to adapt the location search range accordingly i.e. discover locations within the accuracy circle. It is noteworthy that the objective of this component is not to identify the exact location of the smartphone (e.g. which building), but to discover the general characteristics of the location (e.g. is it residential or shopping area) which is helpful to infer the type of activities at that location. For example, if a person is located within a business district between 9am-5pm it might indicate that the trip purpose is work, which will be assured if this trip is repeated on daily basis during the weekdays. In summary, this module receives a coordinate from the location logger and provides a general description of the land use characteristics.



Generally, land use information is obtained from geographic information system (GIS) maps which can be publically or commercially available. Such maps might contain high level information such as residential buildings, parks, hospitals or detailed information to the level of business name and classification. Unless this data is freely available, purchasing such maps may add a non-trivial cost to the data collection project which varies according to the map level of details. Another limitation is that GIS maps typically contain static information that has been collected at a certain period of time and need to be frequently updated to reflect land use changes. Also, they does not provide sufficient information about the popularity of a location i.e. frequency of visitors or real-time information about places (e.g. an ongoing event at certain location). Such information may appear to be unnecessary; however it is very useful in inferring travel activity. For example, in a given neighborhood if the distribution and classification of places are known (e.g. how many restaurant, residential, shops) we can predict the type of activities that can be carried out in this location. In addition, if the visitors' statistics are known at each place we can learn which location is more likely to attract travellers.

The land use discovery module depends on location-based social services such as Foursquare (27) to obtain land use characteristics. Unlike GIS maps that are typically hosted in a closed environment, Foursquare's data are mostly generated by users which make it continuously updated and contain global information (i.e. places worldwide). Rich land use information can be obtained from such services, for example, places can be categorized as homes, offices, schools and shops up to a detailed classification such as college gyms, kids' shops and hiking trails. Such detailed classifications would definitely contribute to recognizing the activity. Additionally, this service provides real-time information on the number of users that are currently at a certain place which suggests the relative popularity of the place; similarly, it can detect public gatherings and massive events that might cause a serious disturbance to the transportation network or attract more travellers. Another example that works closely with Foursquare is the YellowPages (28), which not only provides information such as business name and address, but also the business classification which again can be used to infer activity. All these services are typically accessible via Application Programming Interface (APIs) that can be invoked from a smartphone application. Foursquare data can be integrated with YellowPages data to cross-validate business locations and to enhance the overall data quality. More details on how land use data were obtained from Foursquare are provided in the experiment setup section.

### *Identifying Transportation Mode*

The previous two modules mainly contribute to recognizing the trip purpose, i.e. the activity that derives transportation demand, while this module contributes to identifying the mode of travel. Many of the studies we reviewed here have proposed different solutions to automatically infer transportation mode. In this framework, the mode detection component uses smartphone sensor and does not depend on GPS or network information (e.g. transit routes). This approach was motivated by the observation that every transportation mode seems to have a unique motion pattern. For example, bus is likely to have different acceleration and deceleration pattern than private vehicles. Also, subway might have less sharp turns comparing to bus or car. Smartphones seems to be an appropriate platform to capture motion pattern of transportation modes because they are equipped with different motion sensors such as accelerometer, gyroscope and compass. Also, people typically carry their smartphones and they

do not have to carry an extra device (e.g. GPS).

For this module to operate it requires a model that captures motion pattern of different modes. Such model can be obtained by training a machine learning classifier with labelled dataset that contains sensor data along with the transportation mode. The initial training process can be done offline (i.e. not within the smartphone application), however because the motion pattern can be affected by how people carry their smartphone (e.g. in a bag or inside pocket). Thus, the framework supports online training, in which the initial trained model will be updated independently for each individual to adapt to their motion pattern. The online training can be done within the smartphone application, thanks to the increasing processing power of smartphones. This feature contributes significantly to the value of this framework for practical data collection projects, as it allow the framework (including the smartphone application) to be transferred to different system (e.g. another region) without necessary re-train the model with new dataset. We conducted an experiment to validate the accuracy of the mode detection module, which is presented in the next section.

## EXPERIMENT SETUP AND RESULTS

---

This section presents the setup and results of three experiments to demonstrate the functionalities of the framework, which fulfills the requirements stated earlier.

### *Collecting location data*

We developed an android application that continuously collects location data using network location service provided by Google (25). Six participants installed the application on their smartphone. The application stores data in a local database on the smartphone and then automatically transmitted to a pre-configured server. The application starts automatically when the smartphone boots and keep running in the background. Every five minutes the application wake up and conduct the following:

- Scan nearby Wi-Fi, then checks if the smartphone can see the same (or subset) of the Wi-Fi networks since the previous scan;
  - If yes, consider location is the same and exit;
  - If no, retrieve the current location via a network location provider and compare it to the last known location; if the distance is within 25 meters consider location is the same and exit. Otherwise update the current location.
- If location is the same, update the end time of the present activity to the current time; otherwise add a record for new activity session and set its time to current time (since location changed it is considered as a new activity session).

Collected data from a long-term prototype survey are presented in TABLE 1 which shows a total of 549 days of location data were collected using only network location provider. The objective of this prototype survey is to assess the feasibility of 24 hours location data collection with minimum impact on battery life. Interestingly five out of six the participants reported no significant change in battery life after installing the application. The participant who reported decay in battery life because Wi-Fi remains active, own a phone that is known to have bug when trying to switch off Wi-Fi from the application, which we handled in updated version. Noteworthy, the application collected data within Canada as well as USA.

Location accuracy varies as follows: 1) for participants who live and work in dense urban areas (e.g. Downtown Toronto), the location accuracy was about 100 meters; 2) and for those in relatively low-density areas, the location accuracy was about 600 meters; 3) and it reached 4000+ in low-density cities outside Ontario. Such insights should be cautiously generalized because of the small sample size participated in this experiment. However, our goal in this prototype survey was to only evaluate the effectiveness of network-based location and the results match our expectation that in rural areas and low-density areas where number of Wi-Fi networks are limited the accuracy will deteriorate because it will mainly calculate the location using cellular network, which can have large coverage area (over 1 km).

To validate collected location data, an android application was developed to visualize the location data. The user inputs a specific day then specify activity duration threshold, which is used to filter locations and display only locations where duration of stay (in minutes) greater than the activity threshold. Finally, when the user click “Show Activities” button, the map will be centered on the activity location and will display activity start time and duration. The user can navigate through all the activities that have been conducted that day and easily verify if the application correctly captures location, arrival time and duration of stay. A screenshot of this application is presented in FIGURE 2-a. During the data validation process we discovered that when the smartphone run out of battery the location data might be incorrect (e.g. report longer activity duration). To correct this behaviour, the application log the time before the smartphone shutdown and once it start again, all this period will be marked as ‘unknown’ and presented to the user to verify activity location and duration.

Clearly, collecting long-term location data is proven to be feasible with network-based location especially for activities in dense area where accuracy is relatively high. To the best of the authors’ knowledge, such long-term travel survey (range from 16 days to 9 months) has not been achieved before. Beside network-based location the framework supports GPS location providers as well, such hybrid approach is useful to balance between accuracy and battery life for practical long-term travel data collection which we will use in our next survey wave that targets larger sample size.

### *Acquiring Land Use Information*

Location-based social network services such as Foursquare exposes their data and services for developers as APIs that can be invoked by other applications. Foursquare provides a service that returns the list of all venues or points of interest around a given location and within a certain radius. In this experiment, we extracted all postal codes within the city of Toronto (a total of 49,262 postal codes) from a GIS map (29). Then a Java application was developed to process this database and sends the coordinates of each each postal code to Foursquare to retrieve list of nearby venues within 500 meters radius and store unique venues in a local database. The database has two main tables: 1) venue table which includes 62,493 distinct venue records for the city of Toronto; 2) categories table that contains 393 classifications that can describe a venue (e.g. Residential Building, Restaurant, and Gym). Foursquare has a three-level hierarchal tree to classify a venue. For example, the distribution of location in Toronto, using only the top level classification, as follows: Shop & Service (21%), Travel & Transport (6%), Great Outdoors (5%), Arts & Entertainment (4%), Professional & Other Places (19%), College & University

(3%), Food (14%), Nightlife Spots (3%) and Residence (25%). Furthermore, Toronto seems to be well covered by Foursquare database; to validate this, we grouped all the places by forward sortation area (FSA) which is almost equal to a neighbourhood and found that the average number of locations per FSA is 613, which seems to be appropriate to get the distribution of land use per zone. Such statistics reveal interesting information about the quality of land use information that can be obtained from Foursquare.

Such rich source of land use data is not limited to basic venue attributes such as name, location and classification, which might be found in static GIS maps, but it contains dynamic and real-time information such as check-in statistics including how many users checked in this place and even the number of currently present users. Moreover, Foursquare provides historical data on the daily pattern of check-ins at a specific location, which might reveal significant insights about the attraction power of a certain location or neighborhood change throughout the day; subsequently the travel demand to this area. Therefore, we anticipate many applications of such dynamic land use data source including activity recognition, location choice modeling and context-aware travel data collection.

Additionally, we anticipate possible challenges that might arise when depending on a user-generated data source that might not be verified. We propose three solutions to verify and enhance the accuracy of Foursquare land use information:

- 1- Many businesses claim the ownership of venues as a result Foursquare can assign a field with each venue to indicate if it is verified or no, this can be accessed via when querying venues;
- 2- For the unverified venues, data from multiple location-service providers such as YellowPages or social network services similar to Foursquare can be combined and cross-validated with each other;
- 3- GIS maps that contain detailed point of interest information can be used to complement, verify or correct Foursquare data depending on which data source is more likely to be accurate.

However it can be argued that user-generated data are regularly verified as users update incorrect venue information or report missing venues to Foursquare. Also, there are known success stories of user-generated data such as OpenStreetMap (30) and Wikimapia (31) which contains maps that are edited and updated by users. In short, there is a very promising potential for collecting dynamic land use information from Foursquare and similar services; which we are currently investigating.

### *Transportation Mode Identification*

To collect labelled motion pattern from smartphone sensors, a smartphone application was developed to allow the user to select the transportation mode they are about to take and then use the transportation mode normally. The application (shown in FIGURE 2-b) continuously records sensor data and tags it with the transportation mode the user selected. This labelled dataset was used to train different machine learning algorithms. The application records data from various sensors, but for this experiment only accelerometer, gyroscope and orientation sensors were used. The data were collected using Nexus S smartphone that has 3-axis

accelerometer, gyroscope and orientation (a total of 9 degrees of freedom). The accelerometer sensor measures the forces that are applied on the device's 3-axis. For example, if the phone is placed flat on a desk, the z-axis should measure the gravity force. It is clear that sensor readings are affected by the device orientation and that since gravity is a constant it will not help in predicting travel mode. Consequently, an orientation sensor was included to the features to consider the device orientation when processing other sensors readings. Also, a linear acceleration sensor in Android was used to subtract the gravity constant from accelerometer readings. Linear acceleration and orientation sensors were sampled at 15Hz rate while gyroscope, which is faster, was sampled at 100Hz rate.

Two data pre-processing steps were conducted: first, averaging the gyroscope data to have the same sampling rate as others sensors, and then combining all sensors into one file ordered by the timestamp in which each line has the 3-axis readings of accelerometer, gyroscope then orientation (total of 10 variables including the transportation mode). If one sensor reading was missing, which is rare, the entire record was eliminated to keep a balanced file and maintain the order. Second, every  $n$  records were combined into one record, where  $n$  is the window size. Different window sizes were examined (i.e. 7, 15, 20, 30 and 70) and 7 were selected using cross validation. With window of size 7, a total 133,886 records was obtained and then divided into 65% a training set and 35% a testing set. The window size is merely the number of records to be grouped not seconds; however giving the 15Hz frequency it was approximately equivalent to 0.5 second.

The final two datasets were used to train and evaluate the performance of different machine learning algorithms including NN, SVM and Random Forest. Giving the large dataset, NN and SVM required minimum of three hours of training time, while random forest classifier took a few minutes only. Thus we found random forest to be more plausible for this problem (32). Weka toolkit was used to train the classifiers (33) and obtained performance measures using the testing set. The classification accuracy of the random forest is **98.85%** which is remarkably high, without using GPS or additional information sources; also there was no need to apply any post-processing algorithm to smooth the predictions such as Viterbi algorithm (34).

## CONCLUSION AND FUTURE WORK

This paper presented an integrated framework for collecting detailed travel data while reducing respondents' burden. Results of the conducted experiments clearly indicate the benefits of the proposed framework over traditional travel diaries, which can be summarized as follows:

- Enable long-term data collection of personal travel which is necessary to capture seasonal activities.
- Utilizing smartphones reduces survey administration effort to distribute and collect data loggers (e.g. GPS). Distributing smartphone application can be done much easier than physical devices.
- Smartphone enables two-way communication with travellers, which means instead of having a prompt call survey to collect further data; it can be collected directly using the proposed framework which runs on the smartphone.
- Relying on user-generated public data and maps contribute to reduce the data collection project cost in case it requires proprietary land use dataset or maps.
- Beside data collection, smartphones can be used to influence the travel pattern of travellers by providing personalized real-time information and feedback (unlike

GPS).

- Employing machine learning techniques allow the framework to be transferred to different transportation system or deployed in another region.

Major part of our future work is to implement the activity recognition module to determine the purpose of trips. Although the framework has most of the required data to infer activities such as land use information, it is a challenging task as people can visit the same location for different purposes. We are investigating the application of Bayes network for this inference problem. Further, the proposed framework is currently being tuned to be used in a real data collection project in Toronto with larger sample size.

## References

1. Vautin, D. A., and J. L. Walker. Transportation Impacts of Information Provision & Data Collection via Smartphones. in *Transportation Research Board 90th Annual Meeting*, Washington, D.C, 2011.
2. Bricka, S., S. Sen, R. Paleti, and C. R. Bhat. An Analysis of the Factors Influencing Differences in Survey-Reported and GPS-Recorded Trips. in *Annual Transportation Research Board Meeting*, 2011.
3. Wolf, J., R. Guensler, and W. Bachman. Elimination of the Travel Diary: An Experiment to Derive Trip Purpose From GPS Travel Data. in *Transportation Research Board 80th Annual Meeting*, Washington, D.C, 2001.
4. Wolf, J. Applications of New Technologies in Travel Surveys. in *International Conference on Transport Survey Quality and Innovation*, Costa Rica, 2004.
5. Roorda, M. J., A. Shalaby, and S. Saneinejad. Comprehensive Transportation Data Collection: Case Study in the Greater Golden Horseshoe, Canada. *Journal of Urban Planning and Development*, Vol. vol. 137, no. no. 2, 2011, pp. 193-203.
6. SensorWiki. <http://www.sensorwiki.org/doku.php/sensors/introduction>. Accessed June 2011.
7. Stopher, P., N. Swann, and C. FitzGerald. Using an Odometer and a GPS Panel to Evaluate Travel Behaviour Changes. in *The 11th TRB National Planning Applications*, Daytona Beach, FL, 2007.
8. Bricka, S., J. P. Zmud, J. L. Wolf, and J. Freedman. Household Travel Surveys with GPS: An Experiment. In *Transportation Research Record: Journal of the Transportation Research Board*, no. No. 2105, 2009, pp. 51-56.
9. Lee, M., A. Fucci, P. Lorenc, and W. Bachman. Using GPS Data Collected in Household Travel Surveys to Assess Physical Activity. in *Transportation Research Board 91st Annual Meeting*, 2012.
10. Stopher, P., and L. Shen. An In-Depth Comparison of GPS and Diary Records. in *Transportation Research Board 90th Annual Meeting*, 2011.
11. Stopher, P., Y. Zhang, J. Zhang, and B. Halling. Results of an evaluation of TravelSmart in South Australia. in *Proceedings of the 32nd Australasian transport research forum (ATRF)*, 2009.
12. Li, Z.J. a.A.S.S. Web-based GIS System for Prompted Recall of GPS-assisted Personal Travel Surveys: System Development and Experimental Study. in *Proceedings of the 87th Annual Meeting of the Transportation Research Board*, Washington, D.C, 2008.
13. Mobile Millennium. <http://traffic.berkeley.edu>. Accessed June 2011.
14. Biderman, A., F. Calabrese, K. Kloeckl, C. Ratti, B. Resch, and A. Vaccari. wikicity [\*]. *senseable.mit.edu*, <http://senseable.mit.edu/wikicity/>. Accessed June 2010.
15. Tsui, S.Y.A., and A.S. Shalaby. An Enhanced System for Link and Mode Identification for GPS-based Personal Travel Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, 2006, pp. 38-45.
16. Stopher, P., C. FitzGerald, and J. Zhang. Deducing Mode and Purpose from GPS Data – Case Studies. in *Transportation Research Board 87th Annual Meeting*, 2008.
17. Chung, E.-H., and A.S. Shalaby. Development of a Trip Reconstruction Tool for GPS-Based

- Personal Travel Surveys. *Journal of Transportation Planning and Technology*, Vol. vol. 28, no. no. 5, 2005, pp. 381-401.
18. Auld, J., C. Williams, A. K. Mohammadian, and P. Nelson. An Automated GPS-Based Prompted Recall Survey with Learning Algorithms. *International Journal of Transportation Research*, Vol. vol. 1, no. no. 1, 2009, pp. 59-79.
  19. Byon, Y.-J., B. Abdulhai, and A. Shalaby. Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices. *Journal of Intelligent Transportation Systems*, Vol. vol. 13, no. no. 4, 2009, pp. 161-170.
  20. Gonzalez, P.A., J.S. Weinstein, S.J. Barbeau, M.A. Labrador, P.L. Winters, N.L. Georggi, and R. Perez. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *Intelligent Transport Systems, IET*, Vol. vol. 4, no. no. 1, 2010, pp. 37-49.
  21. Nham, B., K. Siangliulue, and S. Yeung. Predicting Mode of Transport from iPhone Accelerometer Data. *stanford.edu*, 2008. <http://www.stanford.edu/class/cs229/proj2008/NhamSiangliulueYeung-PredictingModeOfTransportFromIphoneAccelerometerData.pdf>. Accessed June 2010.
  22. Bishop, C. M. Pattern Recognition and Machine Learning, in *Pattern Recognition and Machine Learning*. NJ, USA: Springer-Verlag New York, Inc, 2006.
  23. Parlak, S., J. Jariyasunant, and R. Sengupta. Using Smartphones to Perform Transportation Mode Determination at the Trip Level. in *The 91st Transportation Research Board Annual Meeting*, Washington, D.C, 2012.
  24. Windows Location Provider. <http://msdn.microsoft.com/en-us/library/windows/apps/hh464919.aspx>. Accessed July 2012.
  25. Android Training: Location Strategies. *Android Training*, <http://developer.android.com/guide/topics/location/strategies.html>. Accessed July 2012.
  26. Stopher, P. R. Collecting and Processing Data from Mobile Technologie, in *Transport Survey Methods: Keeping Up With a Changing World*, Bonnel, P., M. Lee-Gosselin, J. Zmud, and J.-L. Madre.: Emerald Group Publishing, 2009, pp. 361-291.
  27. Foursquare. <https://developer.foursquare.com/>. Accessed May 2011.
  28. YellowAPI. [http://www.yellowapi.com/?locale=en\\_CA](http://www.yellowapi.com/?locale=en_CA). Accessed June 2011.
  29. (2011, Aug.) Platinum Postal Suite: CanMap Multiple Enhanced Postal Code.
  30. Open Street Map. <http://www.openstreetmap.org/>. Accessed Nov. 2012.
  31. Wikimapia. <http://wikimapia.org/>. Accessed Nov. 15, 2012.
  32. Breiman, L. Random forests. *Machine Learning*, Vol. vol. 45, 2001, pp. 5–32.
  33. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, Vol. 11, no. 1, 2009.
  34. Viterbi algorithm. *Wikipedia*, [http://en.wikipedia.org/wiki/Viterbi\\_algorithm](http://en.wikipedia.org/wiki/Viterbi_algorithm). Accessed June 2011.
  35. NextBus. <http://www.nextbus.com/homepage/>. Accessed Nov. 1, 2012.



## List of Figures

FIGURE 1 Framework Main Components and Data Sources

FIGURE 2: Screenshots of Android applications

FIGURE 3: Example of accelerometer pattern of the x-axis for different transportation modes

shows acceleration pattern of different transportation modes, which clearly confirms that the mode can be classified from motion pattern. A detailed confusion matrix and performance measures (obtained from the independent testing set) are summarized in TABLE 2.

Beside general mode classification, it is possible identify the exact transit route using the NextBus web service (35), which provides real-time information on transit vehicles' movement for many cities across North America. The NextBus service is already integrated in the framework.

## CONCLUSION AND FUTURE WORK

---

This paper presented an integrated framework for collecting detailed travel data while reducing respondents' burden. Results of the conducted experiments clearly indicate the benefits of the proposed framework over traditional travel diaries, which can be summarized as follows:

- Enable long-term data collection of personal travel which is necessary to capture seasonal activities.
- Utilizing smartphones reduces survey administration effort to distribute and collect data loggers (e.g. GPS). Distributing smartphone application can be done much easier than physical devices.
- Smartphone enables two-way communication with travellers, which means instead of having a prompt call survey to collect further data; it can be collected directly using the proposed framework which runs on the smartphone.
- Relying on user-generated public data and maps contribute to reduce the data collection project cost in case it requires proprietary land use dataset or maps.
- Beside data collection, smartphones can be used to influence the travel pattern of travellers by providing personalized real-time information and feedback (unlike GPS).
- Employing machine learning techniques allow the framework to be transferred to different transportation system or deployed in another region.

Major part of our future work is to implement the activity recognition module to determine the purpose of trips. Although the framework has most of the required data to infer activities such as land use information, it is a challenging task as people can visit the same location for different purposes. We are investigating the application of Bayes network for this inference problem. Further, the proposed framework is currently being tuned to be used in a real data collection project in Toronto with larger sample size.

## References

1. Vautin, D. A., and J. L. Walker. Transportation Impacts of Information Provision & Data Collection via Smartphones. in *Transportation Research Board 90th Annual Meeting*, Washington, D.C, 2011.
2. Bricka, S., S. Sen, R. Paleti, and C. R. Bhat. An Analysis of the Factors Influencing Differences in Survey-Reported and GPS-Recorded Trips. in *Annual Transportation Research Board Meeting*, 2011.
3. Wolf, J., R. Guensler, and W. Bachman. Elimination of the Travel Diary: An Experiment to Derive Trip Purpose From GPS Travel Data. in *Transportation Research Board 80th Annual Meeting*, Washington, D.C, 2001.
4. Wolf, J. Applications of New Technologies in Travel Surveys. in *International Conference on Transport Survey Quality and Innovation*, Costa Rica, 2004.
5. Roorda, M. J., A. Shalaby, and S. Saneinejad. Comprehensive Transportation Data Collection: Case Study in the Greater Golden Horseshoe, Canada. *Journal of Urban Planning and Development*, Vol. vol. 137, no. no. 2, 2011, pp. 193-203.
6. SensorWiki. <http://www.sensorwiki.org/doku.php/sensors/introduction>. Accessed June 2011.
7. Stopher, P., N. Swann, and C. FitzGerald. Using an Odometer and a GPS Panel to Evaluate Travel Behaviour Changes. in *The 11th TRB National Planning Applications*, Daytona Beach, FL, 2007.
8. Bricka, S., J. P. Zmud, J. L. Wolf, and J. Freedman. Household Travel Surveys with GPS: An Experiment. In *Transportation Research Record: Journal of the Transportation Research Board*, no. No. 2105, 2009, pp. 51-56.
9. Lee, M., A. Fucci, P. Lorenc, and W. Bachman. Using GPS Data Collected in Household Travel Surveys to Assess Physical Activity. in *Transportation Research Board 91st Annual Meeting*, 2012.
10. Stopher, P., and L. Shen. An In-Depth Comparison of GPS and Diary Records. in *Transportation Research Board 90th Annual Meeting*, 2011.
11. Stopher, P., Y. Zhang, J. Zhang, and B. Halling. Results of an evaluation of TravelSmart in South Australia. in *Proceedings of the 32nd Australasian transport research forum (ATRF)*, 2009.
12. Li, Z.J. a.A.S.S. Web-based GIS System for Prompted Recall of GPS-assisted Personal Travel Surveys: System Development and Experimental Study. in *Proceedings of the 87th Annual Meeting of the Transportation Research Board*, Washington, D.C, 2008.
13. Mobile Millennium. <http://traffic.berkeley.edu>. Accessed June 2011.
14. Biderman, A., F. Calabrese, K. Kloeckl, C. Ratti, B. Resch, and A. Vaccari. wikicity [\*]. *senseable.mit.edu*, <http://senseable.mit.edu/wikicity/>. Accessed June 2010.
15. Tsui, S.Y.A., and A.S. Shalaby. An Enhanced System for Link and Mode Identification for GPS-based Personal Travel Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, 2006, pp. 38-45.
16. Stopher, P., C. FitzGerald, and J. Zhang. Deducing Mode and Purpose from GPS Data – Case Studies. in *Transportation Research Board 87th Annual Meeting*, 2008.
17. Chung, E.-H., and A.S. Shalaby. Development of a Trip Reconstruction Tool for GPS-Based

- Personal Travel Surveys. *Journal of Transportation Planning and Technology*, Vol. vol. 28, no. no. 5, 2005, pp. 381-401.
18. Auld, J., C. Williams, A. K. Mohammadian, and P. Nelson. An Automated GPS-Based Prompted Recall Survey with Learning Algorithms. *International Journal of Transportation Research*, Vol. vol. 1, no. no. 1, 2009, pp. 59-79.
  19. Byon, Y.-J., B. Abdulhai, and A. Shalaby. Real-Time Transportation Mode Detection via Tracking Global Positioning System Mobile Devices. *Journal of Intelligent Transportation Systems*, Vol. vol. 13 , no. no. 4, 2009, pp. 161-170.
  20. Gonzalez, P.A., J.S. Weinstein, S.J. Barbeau, M.A. Labrador, P.L. Winters, N.L. Georggi, and R. Perez. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *Intelligent Transport Systems, IET*, Vol. vol. 4, no. no. 1, 2010, pp. 37-49.
  21. Nham, B., K. Siangliulue, and S. Yeung. Predicting Mode of Transport from iPhone Accelerometer Data. *stanford.edu*, 2008. <http://www.stanford.edu/class/cs229/proj2008/NhamSiangliulueYeung-PredictingModeOfTransportFromIphoneAccelerometerData.pdf>. Accessed June 2010.
  22. Bishop, C. M. Pattern Recognition and Machine Learning, in *Pattern Recognition and Machine Learning*. NJ, USA: Springer-Verlag New York, Inc, 2006.
  23. Parlak, S., J. Jariyasunant, and R. Sengupta. Using Smartphones to Perform Transportation Mode Determination at the Trip Level. in *The 91st Transportation Research Board Annual Meeting*, Washington, D.C, 2012.
  24. Windows Location Provider. <http://msdn.microsoft.com/en-us/library/windows/apps/hh464919.aspx>. Accessed July 2012.
  25. Android Training: Location Strategies. *Android Training*, <http://developer.android.com/guide/topics/location/strategies.html>. Accessed July 2012.
  26. Stopher, P. R. Collecting and Processing Data from Mobile Technologie, in *Transport Survey Methods: Keeping Up With a Changing World*, Bonnel, P., M. Lee-Gosselin, J. Zmud, and J.-L. Madre.: Emerald Group Publishing, 2009, pp. 361-291.
  27. Foursquare. <https://developer.foursquare.com/>. Accessed May 2011.
  28. YellowAPI. [http://www.yellowapi.com/?locale=en\\_CA](http://www.yellowapi.com/?locale=en_CA). Accessed June 2011.
  29. (2011, Aug.) Platinum Postal Suite: CanMap Multiple Enhanced Postal Code.
  30. Open Street Map. <http://www.openstreetmap.org/>. Accessed Nov. 2012.
  31. Wikimapia. <http://wikimapia.org/>. Accessed Nov. 15, 2012.
  32. Breiman, L. Random forests. *Machine Learning*, Vol. vol. 45, 2001, pp. 5–32.
  33. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, Vol. 11, no. 1, 2009.
  34. Viterbi algorithm. *Wikipedia*, [http://en.wikipedia.org/wiki/Viterbi\\_algorithm](http://en.wikipedia.org/wiki/Viterbi_algorithm). Accessed June 2011.
  35. NextBus. <http://www.nextbus.com/homepage/>. Accessed Nov. 1, 2012.

## **List of Figures**

FIGURE 1 Framework Main Components and Data Sources

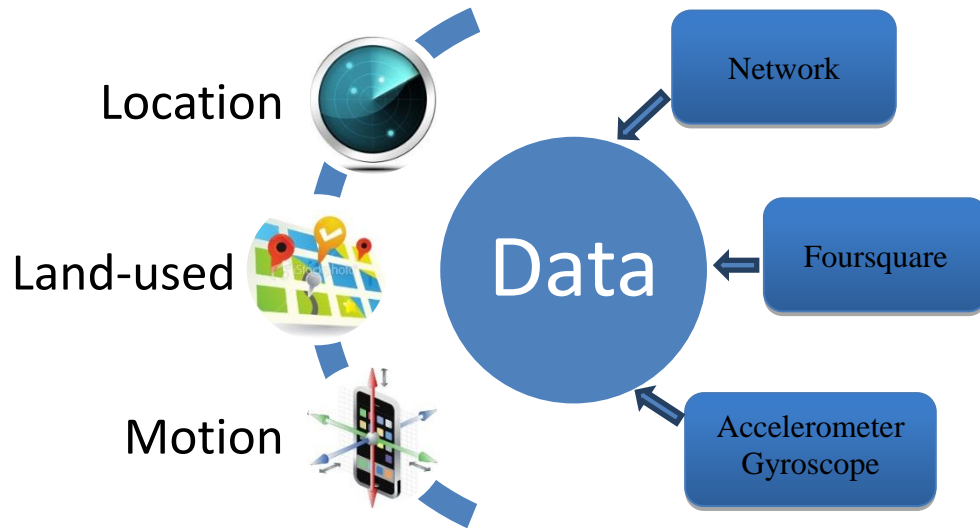
FIGURE 2: Screenshots of Android applications

FIGURE 3: Example of accelerometer pattern of the x-axis for different transportation modes

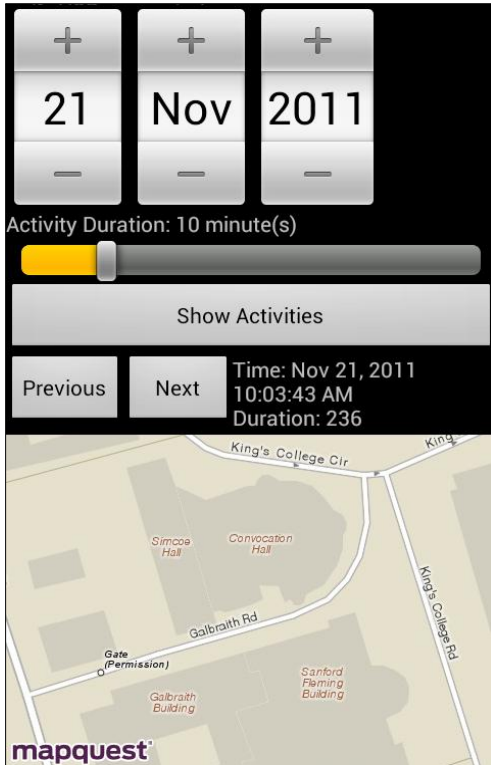
## **List of Tables**

TABLE 1: Summary of location data collected using network provider

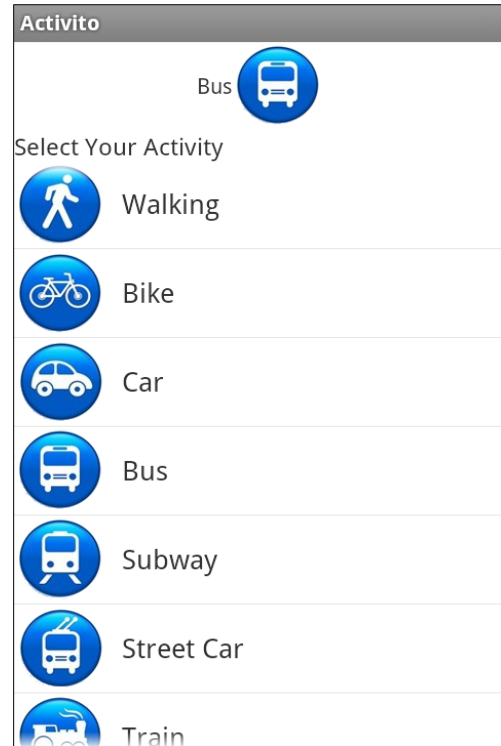
TABLE 2: Confusion Matrix and Performance Measures of the Mode Classification using Random Forests



*FIGURE 1 Framework Main Components and Data Sources*



(a)



(b)

FIGURE 2: Screenshots of Android applications  
(a) Application to verify the accuracy of network-based location data  
(b) Application to collect motion patterns to train the mode identification algorithm

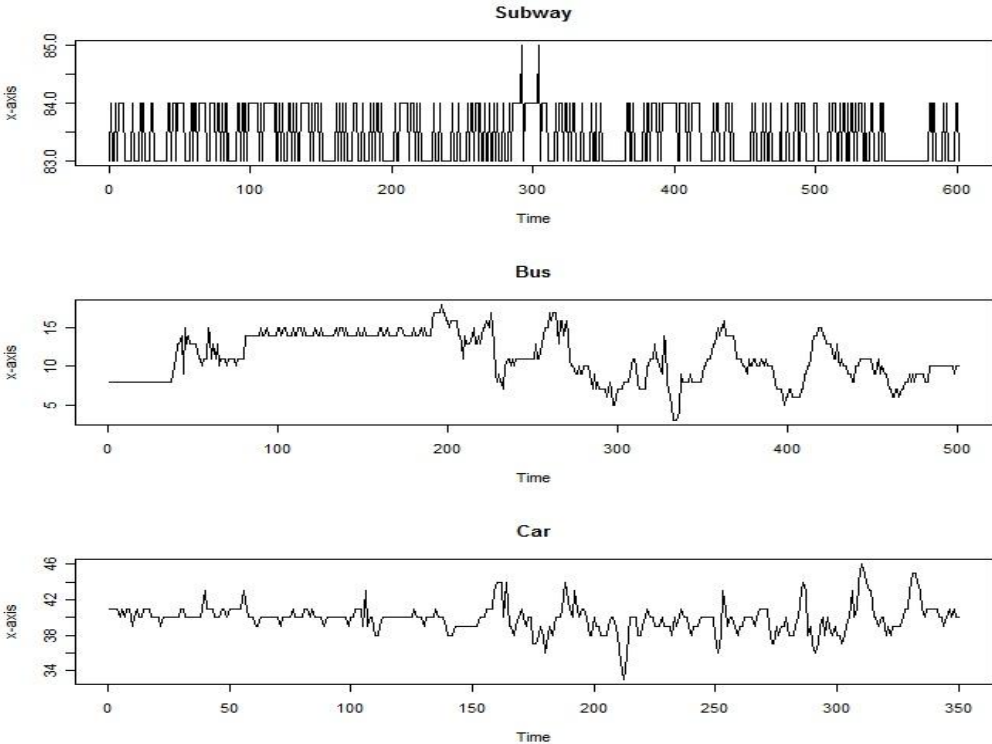


FIGURE 3: Example of accelerometer pattern of the x-axis for different transportation modes

*TABLE 1: Summary of location data collected using network provider*


Participant Id	Start date	Last data update	No. of days with location data <sup>**</sup>	Best location accuracy <sup>*</sup>	Average accuracy <sup>*</sup>	Worse accuracy <sup>*</sup>
1	Oct 22 2011	Nov 11 2011	21	20	99	1242
2	Nov 02 2011	Jan 04 2012	60	20	119	1579
3	Oct 23 2011	Apr 21 2012	156	20	293	4031
4	Oct 22 2011	Jul 25 2012	278	20	302	3026
5	Nov 04 2011	Jan 02 2012	18	25	627	1427
6	May 5 2012	May 14 2012	16	38	2043	4855
<b>Summary</b>	-	-	549	20	549	4855

<sup>\*</sup> Location accuracy in meters

<sup>\*\*</sup> Some days are missing which might be due to international travel or switching off the smartphone



*TABLE 2: Confusion Matrix and Performance Measures of the Mode Classification using Random Forests*

Classified As 	Bus	Subway	Car	Bike	Running	Walking
Bus	<u>2767</u>	0	3	21	0	21
Subway	3	<u>8417</u>	26	36	3	51
Car	0	45	<u>11153</u>	24	6	27
Bike	4	8	0	<u>12937</u>	8	20
Running	0	3	6	24	<u>2053</u>	30
Walking	0	51	18	92	3	<u>8614</u>