International Journal of Transportation Science and Technology 8 (2019) 82–100

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

journal homepage: www.elsevier.com/locate/ijtst

An automated approach from GPS traces to complete trip information

Ali Yazdizadeh [a,*], Zachary Patterson [a], Bilal Farooq [b]

[a] Department of Geography, Planning and Environment, Concordia University, Montreal, Canada
[b] Department of Civil Engineering, Ryerson University, Toronto, Canada

ARTICLE INFO

Article history:
Received 7 February 2018
Received in revised form 16 June 2018
Accepted 17 August 2018
Available online 13 October 2018

Keywords:
Smartphone
Household travel survey
Machine learning
Social network
Activity pattern

ABSTRACT

Recent advances in communication technologies have enabled researchers to collect travel data based on ubiquitous and location-aware smartphones. These advances hold out the promise of allowing the automatic detection of the critical aspects (mode of transport, purpose, etc.) of people's trips. Until now, efforts have concentrated on one aspect of trips (e.g. mode) at a time. Such methods have typically been developed on small data sets, often with data collected by researchers themselves and not in large-scale real-world data collection initiatives. This research develops a machine learning-based framework to identify complete trip information based on smartphone location data as well as online data from GTFS (General Transit Feed Specification) and Foursquare data. The framework has the potential to be integrated with smartphone travel surveys to produce all trip characteristics traditionally collected through household travel surveys. We use data from a recent, large-scale smartphone travel survey in Montréal, Canada. The collected smartphone data, augmented with GTFS and Foursquare data are used to train and validate three random forest models to predict mode of transport, transit itinerary as well as trip purpose (activity). According to cross-validation analysis, the random forest models show prediction accuracies of 87%, 81% and 71% for mode, transit itinerary and activity, respectively. The results compare favorably with previous studies, especially when taking the large, real-world nature of the dataset into account. Furthermore, the cross validation results show that the machine learning-based framework is an effective and automated tool to support trip information extraction for large-scale smartphone travel surveys, which have the potential to be a reliable and efficient (in terms of cost and human resources) data collection technique.

1. Introduction

Travel demand forecasting and modelling are central tools in the analysis of transportation plans, projects, and policies in urban areas. During the last two decades, there have been rapid advances in the collection of data used in transportation modelling. In particular, the use of new technologies such as location-aware smartphones to collect data have become increasingly common (Cambridge Systematics, 2012; Montini et al., 2015; Ta et al., 2016). These advances have enabled researchers to implement travel surveys with smartphone applications (Cottrill et al., 2013; Ta et al., 2016) that gather

* Corresponding author.
E-mail address: ali.yazdizadeh@mail.concordia.ca (A. Yazdizadeh).

https://doi.org/10.1016/j.ijtst.2018.08.003
2046-0430/

respondent traces with lower costs, when compared with traditional travel surveys such as face-to-face interviews (Wolf, 2000), Computer-Assisted Telephone Interviews (CATI) (Stopher et al., 2009), or mail-back paper surveys (Cambridge Systematics, 2012). Moreover, traditional surveys suffer from other shortcomings such as trip under-reporting (Wolf et al., 2003; Pearson, 2004), time inaccuracies and origin-destination location errors, which generally caused by respondent fatigue or forgetfulness.

A great deal of research in this field has concentrated on methods of inferring trip characteristics (e.g. mode) from passively collected data. When this has been done, it often has considered one type of trip characteristic at a time, and has primarily used small or researcher-collected smartphone data; not on large-scale, real-world data collection efforts. The goal of this study is to show the potential of automatically detecting "complete" trip information based on data from a large-scale samrtphone travel survey. We refer to "complete" trip information as including the following characteristics for a trip: geographic origin/destination, start and end time, mode, itinerary (route), as well as purpose the destination of a trip. These are the characteristics typically required in trip-based transportation modelling (Willumsen et al., 2011).

In the literature, methods of origin/destination and mode detection have been studied extensively (Gong et al., 2014; Nitsche et al., 2014; Zhang et al., 2010; Wu et al., 2011). By contrast, activity detection and itinerary inference have received less attention still (Montini et al., 2014; Gong et al., 2014; Zahabi et al., 2017), and the field is open to further research.

We believe that one reason for the lack of research on the two latter characteristics is their dependence on information apart from simple GPS traces of travelers. For example, transit itinerary detection depends on transit schedule and route information. Activity detection relies on comprehensive data from existing land-use as well as popular places around trip destinations (Oliveira et al., 2014; Ermagun et al., 2017). Recently, due to the widespread use of location-base social networks like Foursquare and Yelp (Ermagun et al., 2017; Gong et al., 2014) a comprehensive list of popular activities in urban areas is freely accessible and has the potential to be used in activity detection. Furthermore, General Transit Feed Specification (GTFS) data, which defines a common format to share public transportation schedules and associated geographic information (Catala et al., 2011), enables us to deploy detailed transit schedules and related locational information to infer transit itinerary from raw GPS data.

Apart from the data required, there is also the question of methodological approaches to infer information from smartphone and other data sources. Recently, there has been widespread use of machine learning classifiers, such as Neural Networks (NN) or Random Forests (RF), particularly in mode detection (Gonzalez et al., 2010; Nitsche et al., 2014). Such success in mode detection encourages us to take advantage of such classifiers in transit itinerary and activity detection steps as well. In this study, we have developed a series of machine learning (Random Forest Breiman, 2001) models (See Fig. 1), to automatically detect mode, transit itinerary and activity from smartphone travel survey data. Also, regarding trip/segment detection we have used a rule-based algorithm developed by Patterson and Fitzsimmons (2016).

The overall goal of this study is to show how it is possible to derive all critical trip information gathered by traditional travel survey methods such as CATI or mail-back surveys but through the processing of a large-scale smartphone travel survey data. This research has been conducted using data collected by the MTL Trajet (Patterson and Fitzsimmons, 2017), smartphone Travel Survey App. MTL Trajet was an instance of the smartphone travel survey app, DataMobile (Patterson and Fitzsimmons, 2016). Datamobile was recently renamed Itinerum when it was built out into a travel survey platform in 2017 (Patterson, 2017). MTL Trajet was released as part of a large-scale pilot study on the 17th of October 2016 in a study that lasted 30 days. The location data collected from MTL Trajet is shown on a map of Greater Montreal in Fig. 2. Also, a typical trip trajectory (sequence of GPS points) from MTL Trajet is shown in this figure.

The paper is organized as follows: a background section describes the literature related to GPS-based travel surveys and the algorithms used to process and derive information from them. The Methodology Section describes the methodology and algorithms used to detect the above mentioned "complete" trip information. The Data preparation section introduces the data collected in the MTL Trajet app during the study, as well as additional data used in the analysis. Afterwards, the model estimation section describes the estimated RF models and their attributes. The following section describes the prediction accuracy of the RF models based on the cross-validation results and compares these results with previous studies. Finally, we present our conclusions in the last section.

2. Related background

Urban travel surveys were first introduced in the 1950s, where paper-based face-to-face interviews were conducted to elicit household trip information (Shen and Stopher, 2014). Disadvantages related to this method, such as labour and time costs caused them to be replaced by mail surveys in the 1960s (Wolf, 2000). The main drawback of mail surveys was their low response rates, which gave rise to Computer Assisted Telephone Interviews (CATI) as well as Computer-Assisted Self-Interviews (CASI) (Stopher, 2009). More recently, web surveys, which can be considered as a practice of CASI methods, have been used. However, both CATI and CASI suffer from non-response and misreporting (Shen and Stopher, 2014).

Advances in Information and Communication Technologies (ICT) and the shortcomings of traditional travel survey methods brought about automated travel surveys based on GPS technology, beginning in the late 1990s (Shen and Stopher, 2014). In the beginning, GPS surveys were launched as a supplement to traditional surveys to increase their accuracy. However, the potential for GPS surveys in gathering more accurate spatio-temportal characteristics of trips eventually caused them to be considered as a potential alternative to traditional surveys. Furthermore, by using prompted-recall surveys (Zhao et al.,

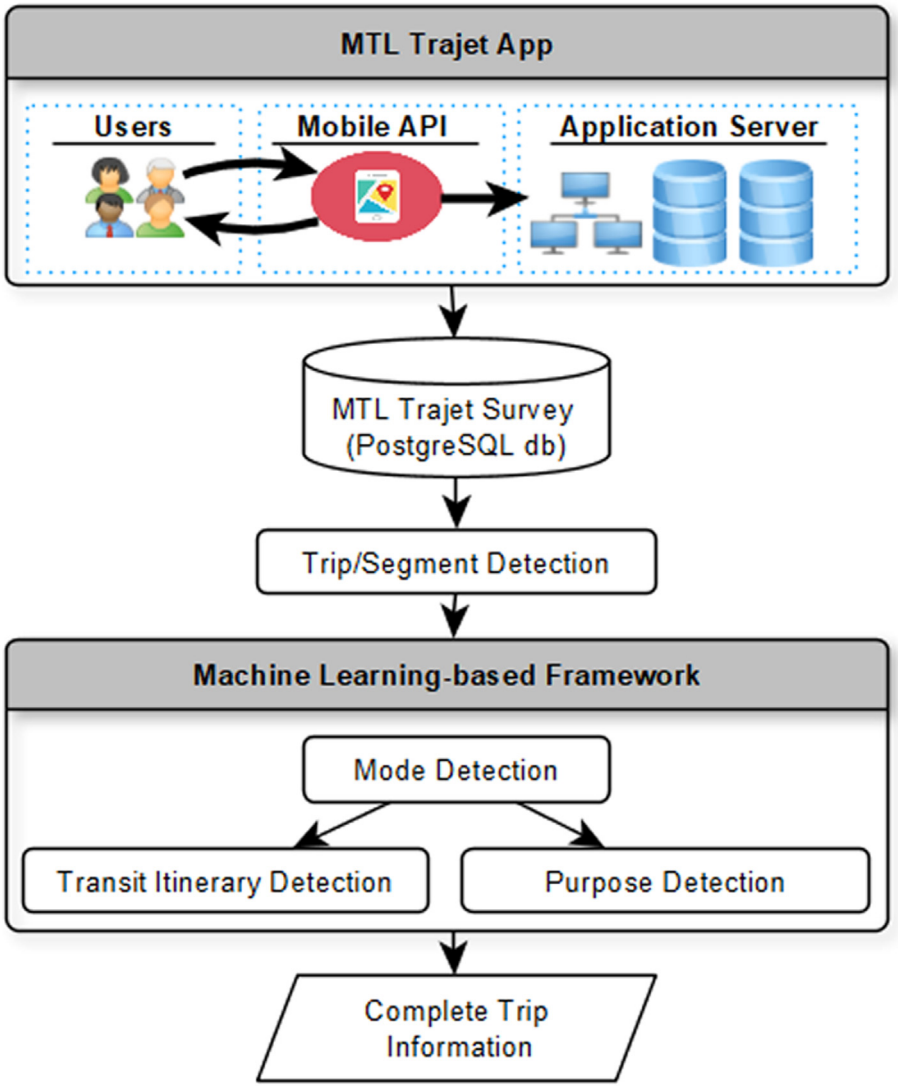

Fig. 1. MTL trajet app and machine learning-based framework to derive complete trip information.

2015), which enabled travelers to validate their trips later on a website, GPS travel surveys grew in popularity. Nevertheless, even prompted-recall surveys face similar problems as traditional survey methods, such as the forgetfulness of travellers when validating their trips after the fact. Some studies, such as Xiao et al. (2016), have implemented surveyor-intervened prompted recall surveys to enhance the accuracy of validations. In such surveys, the surveyors ask respondents by telephone to recall the details of their trips. However, this comes at the expense of rising the surveying cost. Recently, the trip validating step has been carried out in almost real-time by developing in-app prompts (Greene et al., 2016; Patterson, 2017; Patterson and Fitzsimmons, 2017) that ask respondents to validate their trips as they go.

With respect to the theoretical concepts of machine learning algorithms, there is a considerable amount of literature in computer science as well as previously published transportation studies. Witten et al. (2016) have clearly explained the machine learning algorithms such as decision trees, random forest, support vector machines, etc., and their application in different fields of study. The book written by Goodfellow et al. (2016) is a comprehensive reference to neural network algorithms and their applications. Regrading the transportation studies, Ghasri et al. (2017) have explained in detail the decision tree and random forest classifiers. Dabiri and Heaslip (2018) and Gonzalez et al. (2010) have described neural networks models and their application in transportation mode detection.

In this section, we review studies using GPS-based travel surveys that also developed algorithm(s) for detecting trip mode, transit itinerary or activity. Tables 1 and 2 report different approaches applied for trip/segment and mode detection, transit itinerary as well as activity detection, and their prediction accuracy.

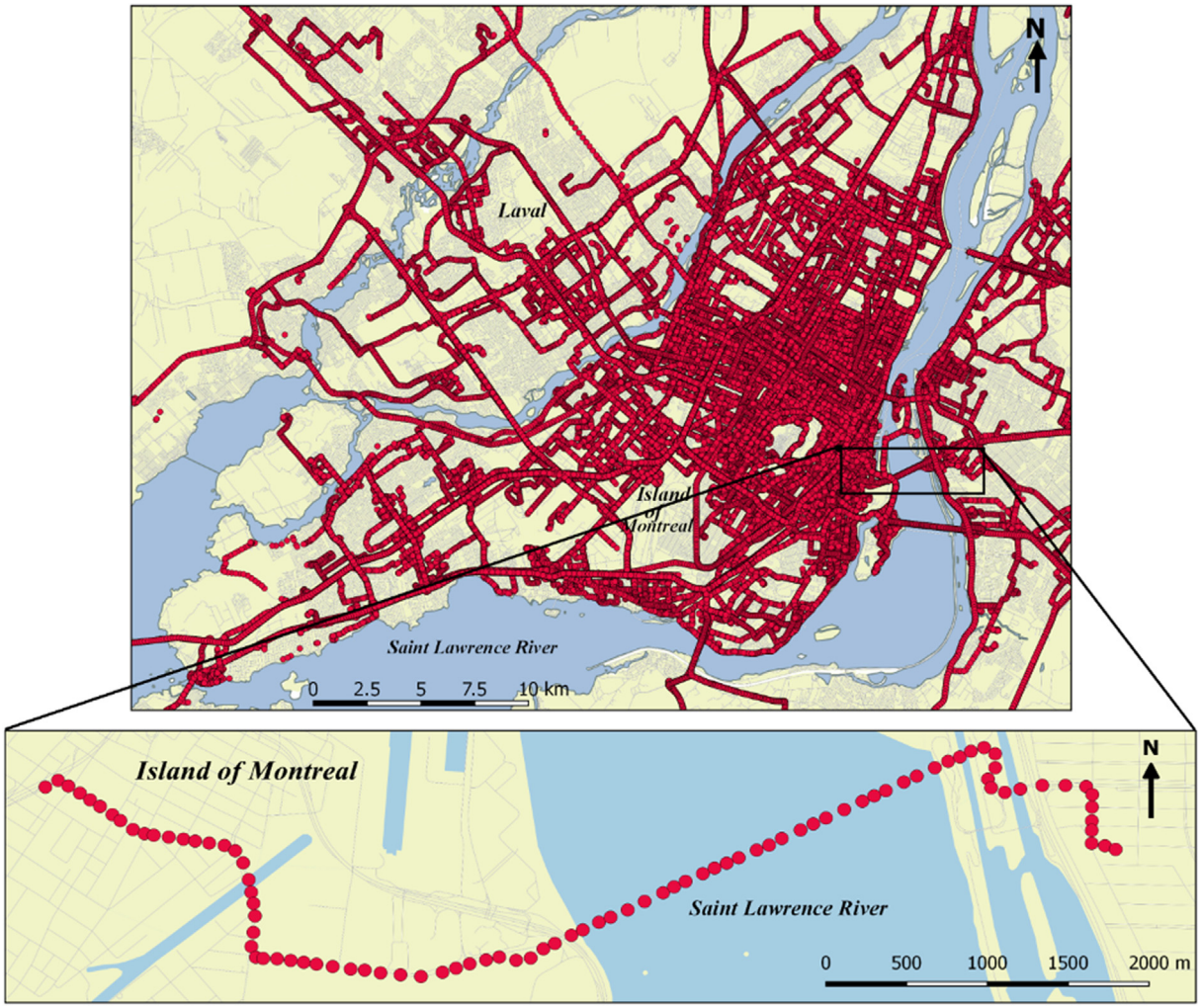

Fig. 2. Location data collected in the MTL trajet study.

2.1. Trip/segment detection

With respect to detecting trip/segment, several studies have used rule-based algorithms, as shown in Table 1. Recently, Zhou et al. (2017) have developed a Random Forest model to identify the trips ends using large-scale smartphone-based GPS tracking data. Their model achieved an accuracy of 96.17% for the identification of trip ends. With respect to the rule-based algorithms, dwell time between GPS points has been considered as the most prominent criterion for detecting the segment/trips, ranging from 1 to 3 min. Additional rules have also been applied for trip/segment detection (Shen and Stopher, 2014), such as point density (Schuessler and Axhausen, 2105), latitude and longitude change (Stopher et al., 2009) or speed threshold and amplitude of the accelerometer signals (Nitsche et al., 2014).

Zhou et al. (2017) have used several attributes as input for the random forest classifier, namely (1) local attributes such as time difference, instantaneous speed and acceleration, (2) global extreme value attributes like total duration and total duration, (3) speed-related attributes such as average speed and standard deviation of the instantaneous speed (4) acceleration-related attributes like the average absolute acceleration and the largest absolute acceleration, (5) tracking points clustering attributes, such as the largest distance between any two points within one neighborhood point set and (6) heading change attributes, like the average heading change in the neighborhood point set. They also proposed the concept of neighborhood point set to describe the temporally continuous points near a specific GPS tracking point.

2.2. Mode detection

Mode detection, now well examined in the literature has been done using various approaches. Rule-based algorithms (Stopher, 2009; Bohte and Maat, 2009), fuzzy systems (Schuessler and Axhausen, 2105; Biljecki, 2010), machine learning

Table 1
Studies on trip/segment and mode detection.

Author/s	Method	Attributes	Validation Accuracy
Trip/Segment Detection			
Stopher (2009)	RBA*	Dwell time	n/a
Wolf et al. (1768)	RBA	Dwell time	n/a
Schuessler and Axhausen (2105)	RBA	Dwell time, distance between points	n/a
Biljecki (2010)	RBA	Dwell time, distance between points	91%
Gonzalez et al. (2010)	Manually segmented by the cell phone user	Speed, acceleration, data quality, travel distance	n/a
Shafique and Hato (2015)	RBA	Dwell time	n/a
Nitsche et al. (2014)	RBA	Speed, high amplitudes accelerometer signal	n/a
Dalumpines and Scott (2017)	RBA	Dwell time and Speed	n/a
Xiao et al. (2015)	RBA	Dwell time, Critical length, critical distance	96%
Patterson and Fitzsimmons (2016)	RBA	Dwell time, distance to metro stops, speed	n/a
Zhou et al. (2017)	Random Forest	24 attributes including: instantaneous speed, acceleration, average absolute acceleration standard deviation of the instantaneous speed average heading change etc.	96.17%
Mode detection			
Stopher (2009)	RBA	Speed, GIS, car/bike ownership	95%
Bohte and Maat (2009)	RBA	Average and maximum speed, GIS land use data	n/a
Schuessler and Axhausen (2105)	Fuzzy-logic approach	Speed, Acceleration	n/a
Biljecki (2010)	Fuzzy System	Proximity of trajectories to nearest network, speed	n/a
Gonzalez et al. (2010)	Neural Network	Maximum speed, Average speed, Maximum acceleration, Average acceleration,Total Distance	91.23%
Feng and Timmermans (2013)	Bayesian Belief Networks	Speed, Accelerate, Car/Bike/Motor-cycle ownership, data quality	96%
Shafique and Hato (2016)	Random Forest	Acceleration and standard deviation, skewness, kurtosis, maximum and average acceleration	99.6%
Nitsche et al. (2014)	Ensemble of probabilistic and Discrete Hidden Markov Model (DHMM)	5th, 50th and 95th percentile of speed, accelerations, decelerations, direction change, standard deviation of the highfrequency accelerometer magnitudes, and power Spectrum of the accelerometer signal for Frequencies	Range from 65% to 95%
Dalumpines and Scott (2017)	Multinomial Logit	Median speed, Median change in heading, total travel time	90%
Xiao et al. (2015)	Basyesian Network	Travel distance, average speed, average absolut acceleration, 95% percentile speed, low speed rate, and average heading change	92.74%
Eftekhari and Ghatee (2016)	RBA	acceleration data from gyroscope and accelerometer sensors	95.2%
Bantis and Haworth (2017)	Basyesian Network	Speed and socio-demographic data	Range from 82% to 90%
Endo et al. (2016)	Deep Neural Networks	Time series of speed,head change, time interval, distance of the GPS points	Range from 84.8%
Dabiri and Heaslip (2018)	Convolutional Neural Networks	Speed, acceleration, jerk (the rate of change in the acceleration) and bearing rate (rate of change in the heading direction)	Range from 84.8%

* RBA = Rule-Based Algorithm.

classifiers (Feng and Timmermans, 2013, 2010, 2016, 2014, 2015) and discrete choice models (Dalumpines and Scott, 2017) have been implemented. Mode detection algorithms have been applied on various data sources including raw GPS trajectories (Zheng et al., 2008; Dabiri and Heaslip, 2018; Rezaie et al., 2017), smartphones accelerometers (Eftekhari and Ghatee, 2016), call detail records (CDRs) from smatrphones (Lin et al., 2017), and smartphone's GSM data (Sohn et al., 2006). Moreover, some of the mode detection studies has exploited multiple data sources to improve the prediction accuracy of classifiers. Stenneth et al. (2011) integrated the GPS and GIS information for developing a mode detection algorithm. Other studies (Nitsche et al., 2014; Eftekhari and Ghatee, 2016) have used gyroscope, rotation vector, and magnetometer data to improve the mode detection accuracy. In addition to the smartphone data, traveller's socio-demographics can result in higher prediction accuracy of classifiers (Bantis and Haworth, 2017). A comprehensive and comparative review of existing studies on travel mode detection have been presented in the paper by Wu et al. (2016).

Zheng et al. (2008) developed a framework to automatically infer mode of transport from GPS trajectories. They have applied a rule-based segmentation algorithm to split a trip into segments with distinct modes of transport. Afterwards, various features such as mean and variance of the speed, top three speeds and accelerations have been utilized to develop

Table 2
Studies on trip activity detection and transit itinerary inference.

Author/s	Method	Attributes	Validation Accuracy
Activity Detection			
Stopher (2009)	RBA*	GIS land-use data, home and workplace/ school addresses, address of the two most frequently used grocery stores	over 60%
Wolf et al. (1768)	RBA	GIS land-use data	69%
Bohte and Maat (2009)	RBA	GIS land-use data, home and workplace/ school addresses	43%
Wu et al. (2011)	Random Forest	Distance, speed, acceleration	97% for indoor, 84% for in-vehicle travel, and lower for other
Lu and Liu (2012)	Decision tree	Socio-demographics, land use, temporal information,previous & next trip attributes	73.4%
Pereira et al. (2013)	Historical data matching rules	Activity duration,POI, socio-demographics, work hours travel time	n/a
Zhu et al. (2014)	Support Vector Machine	Demographic Features: gender, age, occupation spatial Features: Foursquare venue category and tips temporal Features: duration stayed at a destination	75%
Oliveira et al. (2014)	Decision tree	Land use, temporal information, socio-demographics	65%
Kim et al. (2015)	Random Forest	Point of interest, age, gender	75.5%
Xiao et al. (2016)	Neural Networks	Polygon-based information and point of interest	96.5%
Zhang et al. (2017)	Sequential model-based clustering method	Visiting frequency, most frequently-visited locations, distance between visited locations, relation between a location and its surrounding	n/a
Ermagun et al. (2017)	Random Forest	Travel mode, previous activity type, trip characteristic, Nearby place characteristics, socio-demographics	64.17%
Transit Itinerary Inference			
Zahabi et al. (2017)	RBA	Nearby bus routes (GTFS data), dwell time, speed	87%

* RBA = Rule-Based Algorithm.

classifiers, such as decision trees, to detect mode of transport. Sun and Ban (2013) extracted acceleration- and deceleration-based features to classify vehicle type by support vector machine (SVM) classifier.

Stenneth et al. (2011) developed five classification algorithms (Bayesian Net, Decision Tree, Random Forest, Nave Bayesian, and Multilayer Perceptron) to detect mode of transport. Their result demonstrated the superiority of Random Forest over other developed classifiers, in terms of prediction accuracy rate.

Recently, Xiao et al. (2017) developed tree-based ensemble classification algorithms that outperform traditional ones such as the decision tree. Endo et al. (2016) and Wang et al. (2017) developed Deep Neural Network (DNN) algorithms to detect mode of transport. Also, Dabiri and Heaslip (Dabiri and Heaslip, 2018) have used Convolutional Neural Network (CNN) to train a mode detection model.

2.3. Activity detection

Unlike mode detection, transit itinerary and activity detection have not received as much attention in the literature. With respect to activity detection, the methods in the existing literature have been categorized by Gong et al. (2014) into three broad categories: rule-based algorithms, statistical methods and machine learning methods. However, recent studies in the field tend to use machine learning approaches along with geo-tagged information from social media, such as FourSquare or Google Places (Ermagun et al., 2017). Rashidi et al. (2017) reviewed the studies on social media data and concluded that geo-tagged data possess a great potential to improve our knowledge in understanding activity behaviour. Geo-tagged social media data have been utilized by several studies from social science, computer science, and transportation science to extract meaningful activity behaviour patterns (Rashidi et al., 2017). The studies include activity recognition (Lian and Xie, 2011), activity choice patterns (Hasan and Ukkusuri, 2014), and understanding life-style behaviour from activity-location patterns (Hasan and Ukkusuri, 2015). Lee et al. (2016) utilized geo-tagged tweets to generate individual activity spaces based on minimum bounding geometry (convex hull). Hasan and Ukkusuri (2014) investigated Foursquare check-in data to infer individual weekly activity patterns using probabilistic topic models. However, as Rashidi et al. (2017) have mentioned, the capacity of social media platforms such as Facebook, Twitter, LinkedIn, Foursquare, and Yelp to provide information on household daily travel has been minimally examined (Rashidi et al., 2017). Zhang et al. (2017) proposed a sequential model-based

clustering method to investigate the potential of social media (Twitter) to realize the longitudinal household survey. They concluded that geo-tagged data (tweets) provide a sample of human activity space through a list of locations. Zhu et al. (2014) investigated the potential of location-based social networks to explain the travellers' behaviour. They trained a Support Vector Machine (SVM) algorithm that achieved over 75% accuracy in predicting trip purposes combining with the traditional travel survey. Also, Lee et al. (2016) reviewed studies on emerging data collection technologies for mode of transport and trip purpose prediction.

As demonstrated in Table 2, different attributes have been used in the literature to predict activity, such as information on land use types around each trip destination, related activity characteristics (e.g., duration of the activity at the destination, previous activity, activity start/end time), socio-demographics data (e.g., respondents age, occupation and income), and Point of Interest (POI) data.

Oliveira et al. (2014) developed a decision tree model to differentiate between 12 trip purposes. In their study, the decision tree model achieved a prediction accuracy of 65%. They also used a Nested Logit model that predict trip purposes with 60% accuracy. Wu et al. (2011) developed two models, a rule-based and a random forest model to classify indoor, outdoor static (i.e., when an individual is relatively stationary while outdoors), out-door walking, and in-vehicle travel activities. Both of the models successfully predicted indoor activities and in-vehicle travel with 96% and 88% accuracy, respectively. However, the both models were fairly successful in identifying outdoor static and walking points. Kim et al. (2015) developed a Random forest model to differentiate between 16 purposes. They used age and gender variables in the modelling process to improve the prediction performance of the model. Their Random Forest model can predict the trip purpose by 75.5% prediction accuracy.

2.4. Transit itinerary detection

With respect to transit itinerary detection, we only found the study by Zahabi et al. (2017) that implemented a rule-based algorithm on a GPS travel survey to detect transit itinerary. They used GTFS data to find a set of nearest active transit routes around each GPS point, and then inferred transit route by checking the history of all GPS points along a trip.

Based on this literature review, we note the following. First, studies typically consider only one trip characteristic at a time. Second, studies on mode and activity detection have been primarily based on small or researcher-collected smartphone data. For example, among 11 mode detection studies reviewed by Wu et al. (2016) the sample sizes are less than 45 persons or less than 114 trips. Third, we observe that there is relatively little literature on detecting trip activity, and even less on transit itinerary. As such in this paper, we develop a framework to predict trip mode, activity as well as transit itinerary; that is, all trip characteristics typically required in trip-based modelling approaches, using data from the same large-scale study. In the activity detection step, we use FourSquare data as a standardized and highly accessible source for data on nearby places surrounding a trip destination. Also, in transit itinerary detection step, we use OpenTripPlanner router (Opentripplanner bibliography, 2017) to generate all possible transit options between each trip origin and destination.

3. Data used

I all data sources were used to developing the models and algorithms to infer "complete" trip information from the MTL Trajet study:

1. The MTL Trajet survey
2. A Transit Itinerary survey
3. Land-use Data
4. Foursquare data
5. General Transit Feed Specification (GTFS) data
6. Bing Elevation data

A brief description of the above mentioned datasets is provide below.

3.1. MTL trajet survey

The purpose of the MTL Trajet study (Patterson and Fitzsimmons, 2017; Ville de Montreal et al., 2017) was to implement a large-scale pilot study by the City of Montreal to contribute to the development of their "Smart City" initiatives. The study was live from the 17th of October until the 17th of November 2016. In order to participate in the study, respondents needed to download the application, agree to the consent form, answer a few socio-demographic and daily travel-related questions and then allow the app to operate in the background of their phone. Respondents would be prompted after each stop to validate the mode and purpose of their trips. By the end of the study, there were 11,433 downloads of the Itinerum app, 4780 on Android and 6653 on iOS. Among them, 8033 users responded to the socio-demographic and survey questions. In the end, there were 7773 users for whom at least two data points were collected. These users provided data for 88,629 person days

for an average of 11.4 days of participation per person. With respect to validations, 6845 respondents validated at least one trip. Altogether, there were 131,777 validated trips, for an average of 19.2 validations per person.

3.2. Transit itinerary survey

The Transit Itinerary Inference (TII) data (Zahabi et al., 2017) was collected with the iOS version of the Datamobile app through a data collection initiative conducted at Concordia University in Montreal during July and August of 2016. The main goal was to collect validated transit itinerary data. There were 10 student surveyors who were asked to recreate transit trips identified randomly selected from the 2013 regional household survey in Montreal. The participants were asked to validate all public transit routes they used during the course of their trips. In the end, 599 validated transit segments were used in this analysis. We used this data set to train a Random Forest model to detect the transit itinerary.

3.3. Land-use data

Land-use information is one of the data sources with which the activities around a GPS coordinate can be inferred. The last updated Montreal land-use data was compiled by the provincial government ministry "MAMROT" (MAMROT, 2011) for 2011 for the "Montreal Metropolitan Community." This land-use data contains land-use characteristics of buildings. There are around 970 land-use types which have been classified into 23 categories. However, as a large part of land-use parcels are residential buildings, using only land-use data may cause the activity detection algorithm to be prone to classification error, due to myriad residential parcels around trip destinations. Thus, we sought other location based data sources, such as Foursquare, to use them as a complement to land-use data.

3.4. Foursquare data

Foursquare is an online location-based social network through which individuals can "connect" with the places they visit through "check-ins" using the Foursquare app. In general, a check-in specifies that a certain user has been present at a given venue. The check-ins are then associated with the venue as well as to other "friends" with the app (Cebelak, 2015). For this study, for each trip destination, we sent a request to the Foursquare API to search all venues within 250 meter of a trip destination. According to the online Foursquare API documentation (Foursquare for developers: Venue response, 2017), each request to Foursquare API returns maximum of 50 venues. For each venue, there are 35 fields of information, among which the following were used in the current study:

- Categories that indicate the Foursquare sub- or sub-sub-category to which the venue belongs
- Stats, which contains two pieces of useful information:
    1. *checkinsCount*, which is the total check-ins ever in a venue
    2. *usersCount*, which is the total users who have ever checked in a venue

As explained in the Foursquare API Documentation for Venue Categories (Foursquare for Developers: Venue Categories, 2017) Foursquare has categorized venues into the 10 top-level categories. Each top-level category has "sub-" and "sub-sub-" categories that result in a total of 910 categories. In this study, we have aggregated all *checkinsCount* for each top-level category, giving 10 *checkinsCount* for each trip destination. Also, the same procedure has been applied on *usersCounts*. This information is used in the random forest model to predict the trip purpose.

3.5. Bing elevation data

We used elevation data to approximate maximum and minimum slope of the earth along a trip. According to the documentation of Bing Elevation API (Get Elevations – MSDN, 2017), we can get elevations at equally-spaced locations along a path. By dividing the difference between the elevation of each two consecutive points by the direct distance between those points, we are able to approximate the slope of the earth between each pair of points. Then, the maximum and minimum slope along a trip were used in mode detection.

4. Research design and methodology

The framework developed to infer complete trip information is illustrated in Fig. 1. The MTL Trajet app was the data collection tool that gathered user location information and sent it to the MTL Trajet server. Afterwards, the data was transferred to a PostgreSQL database and cleaned through procedures explained below. Finally, the machine learning-based framework was developed to infer complete trip information. In this research, we have used one of the most common classifiers in machine learning, the random forest method, as it has shown to provide high prediction accuracy in many transportation classification situations (Shafique and Hato, 2015; Stenneth et al., 2011; Wu et al., 2016; Ghasri et al., 2017; Ermagun et al., 2017). Furthermore, since random forest is an ensemble learning approach, where predictions are made based on

multiple decision trees, it is less prone to over fitting (Ermagun et al., 2017; Ghasri et al., 2017). The research includes four major steps: (1) data preparation, (2) model selection and attribute description, (3) model estimation, and (4) model accuracy validation. In this section, the first two steps are discussed. First, data preparation procedures are discussed. Afterwards, as the Random Forest (RF) model is the core machine learning algorithm in this study, a brief explanation of random forest basics and related terminology is provided. Finally, the approaches used to derive complete trip information are explained.

4.1. Data preparation

The MTL Trajet dataset contains over 33 million location (primarily GPS) points. To detect trips and segments we used the rule-based trip-breaking algorithm developed in Patterson and Fitzsimmons (2016). The algorithm detects segments based on 3-min gaps in data while controlling for velocity and parameters relating to the public transit network (i.e. transit junctions and metro station location). Applying the trip-breaking algorithm on the MTL Trajet dataset resulted in 623,718 trips, of which 102,904 trips were validated by respondents.

Validated mode data was derived from the survey questions presented to respondents upon installation. In particular, respondents were asked the location of home, work and school, as well as the mode(s) of transport used for trips to these locations. Only trips from users who declared using only one mode option to travel between home and work or home and school were used. This procedure provided us with 10,518 validated trips. With respect to trip activity detection, six activity categories were used to predict trip purpose: "education," "health," "leisure," "shopping/errands," "return home" and "work." We used 102,904 prompt validated trips to train an RF model to detect the trip purpose.

4.2. Random forest model: basics and terminology

The RF method is based on a combination of decision tree predictors. The RF method first generates a series of training samples from the original training dataset with bootstrapping (Breiman, 2001). Then, about one-third of each bootstrap training sample is left out, called the Out-of-bag (OOB) set, and a decision tree is constructed on the remaining observations (the other two-thirds of the bootstrap sample is referred to as the InBag set) (Breiman, 2001). The newly constructed decision trees are generated with "random split selection" (Dietterich, 2000) where a random selection of attributes are used at each node to determine the split (Breiman, 2001). Finally, these decision trees will later "vote" to form the bagged predictor (Breiman, 2001). As RF model consists of multiple trees, the results are more stable than the individual decision tree models and less prone to overfitting (Breiman, 2001).

The OOB observations are used to make a natural test set for calculating the error of each tree, rather than using the cross-validation method to estimate the error of the RF (Archer and Kimes, 2008; Breiman, 2002). In addition, the OOB sets can be used to calculate variable importance. The importance of each variable in classification is crucial information for the model builder to know. Several ways are suggested in the literature to measure the variable importance (Breiman et al., 2003; Archer and Kimes, 2008). The best-known criteria is the "mean decrease in accuracy", that has been used in several studies in the field (Liaw et al., 2002; Ermagun et al., 2017; Archer and Kimes, 2008). The larger mean decrease of a variable, the more important that variable is deemed. The details of the calculation of the "mean decrease in accuracy" have been well explained by Breiman (2002) and Archer and Kimes (2008). For each tree in the random forest model, the importance of a variable V is defined as the decrease in the predictive accuracy on OOB data when the variable V is permuted. Overall variable importance is calculated as the average variable importance across all trees in the Random Forest. By definition, if variable V is not in a tree, its importance is set to zero.

Moreover, as Breiman and Cutler (2007) have explained, since the values of variable importance from tree to tree are independent, we can compute the standard error of each variable's importance. Finally, dividing a variable's importance by its standard error gives us a z-score that can be used to assign a significance level to each variable. Also, we can test the null hypothesis of zero importance for variable V and reject it where the calculated variable importance exceeds the -quantile of N(0,1).

There are two hyperparameters in growing a forest. First, the number of randomly selected attributes used for each split. Second, the number of trees grown in the forest. Liaw et al. (2002) suggested that for stable estimate of variable importance, a large number of trees is essential.

We have applied the RF model explained above to classify trip mode, activity as well as transit itinerary. All the RF models in current research were fitted using the randomForest package (Liaw et al., 2002) in R version 3.4.0 (R Core Team, 2014).

4.3. Mode detection

For this paper, a Random Forest (RF) model has been generated based on 10,518 validated trips from the MTL Trajet Survey. The mode of transport has been classified into five categories: walk, bike, car, public transit as well as car and public transit. Also, various attributes, summarized in Table 3, have been selected for generating the mode detection model. We have used trip characteristics such as 85th percentile speed, max/min acceleration and their difference, direct and cumulative distance. As traveler behavior varies on different days of the week and during different hours of a day, we also included time and day attributes related to a trip. Socio-demographic such as age, sex and occupation are the other attributes we considered in the RF model.

Table 3
Attributes used in mode detection analyses.

Attribute	Definition
Trip characteristics	
CUM_DIST	Cumulative distance (m) between O-D
DIR_DIST	Direct distance (m) between O-D
TR_TIME	Travel Time (min.) between O-D
AVG_SPEED	Average speed (km/h) between O-D
85_SPEED	85th percentile speed (km/h^2) between O-D
MAX_ACC	Maximum Acceleration (km/h^2) between O-D
MIN_ACC	Minimum Acceleration (km/h^2) between O-D
DIFF_ACC	Difference between Min. and Max. Acceleration (km/h^2) between O-D
SLOPE_MIN	Minimum slope between O-D
SLOPE_MAX	Maximum slope between O-D
MAX_TIME_POINTS	Max time interval (min) between each consecutive pair of GPS point
MAX_DIST_POINTS	Max distance (m) between each consecutive pair of GPS point
Time and Day characteristics	
TIME_DAY	Time of day from 0 to 24
DAY	1–7 for Monday to Sunday
Socio-demographics characteristics	
AGE	0: age between 16–24,1: 25–34, 2: 35–44, 3: 45–54, 4: 55–65, 5: 65+
GENDER	0:male, 1: female, 2: other/neither
OCCUPATION	0: full-time worker, 1: part-time worker, 2: Student, 3: Student and worker, 4: Retired 5: At home
AVG_PRICE_NEIGH	The average value of residential buildings around each individual's home (in 250 meters radius)
THURSDAY	1: If the trip is completed on Thursday/0: Otherwise
FRIDAY	1: If the trip is completed on Friday/0: Otherwise
Geographical and closeness characteristics	
GTFS_ORIGIN	Direct Distance between the origin and nearest public transit stop from GTFS data
GTFS_DESTIN	Direct Distance between the destination and nearest public transit stop from GTFS data
CBD_ORIGIN	1: if the origin is located in Montreal's CBD, 0: otherwise
CBD_DESTIN	1: if the destination is located in Montreal's CBD, 0: otherwise
AVG_PRICE_NEIGH	The average value of residential buildings around each individual's home (in 250 m radius)

4.4. Transit itinerary inference

Automatic detection of transit itineraries from smartphone travel surveys is valuable to transportation planners for analyzing transit planning scenarios (Zahabi et al., 2017). Our approach to infer transit itinerary is twofold: first, it finds all possible transit itineraries between each origin and destination. Second, by training a machine learning model, the actual transit itinerary (i.e. the one chosen by the travelers) is detected. From a graph theoretical perspective, finding possible transit itineraries between an OD requires a graph search algorithm, such as Dijkstra's Skiena, 1990, Bellman-Ford Bellman, 1958 or Multiobjective A^* (Stewart and White, 1991; Mandow et al., 2005) algorithm. Such algorithms search among all possible paths for the one that incurs the smallest cost (e.g. shortest travel time or minimum travel distance) (Stewart and White, 1991; Mandow et al., 2005). Detected itineraries are made up of a series of nodes and links from the transit and street networks. In this study, we have used the GTFS-based transit network, and the street network is taken from OpenStreetMap (OpenStreetMap contributors, 2017). The graph search algorithm used is the one available through OpenTripPlanner (OTP) (Opentripplanner bibliography, 2017), an open-source, multimodal trip planning software system. OTP uses the Multiobjective A^* (Stewart and White, 1991; Mandow et al., 2005) algorithm to find candidate transit itineraries (sequence of routes) between each origin and destination.

One approach to identifying transit itinerary among a set of possible alternatives is to match the trip trajectory and all candidate transit itineraries to find the itinerary chosen by the traveller. However, some problems arise with this approach. It is not always straightforward to identify the actual transit itinerary by matching the travel trajectories and all possible transit itineraries. First, GPS trajectories do not always perfectly match candidate transit itineraries due to errors in GPS recordings as well as signal loss in some urban areas or during underground trips. Second, when a transfer has occurred during a trip, there is a walking connection between the two transit routes which is not easy to detect, especially when transfers occur between bus and metro.

To solve such problems, some rules need to be applied. For example, a rule is required to specify the acceptable overlap (matching) percentage between the trip trajectory and alternative transit routes in order to say that the alternative route and the trip trajectory are the same. However, there is a wide spectrum of overlapping percentage values, from around 10% for metro trips where there is signal loss, or above 90% when the trip is done by bus without any transfer or signal loss.

Furthermore, we found that besides the overlapping percentage value, other attributes can help us to identify the chosen alternative, such as walking distance or waiting time during a transit trip. In addition, as the routing algorithm finds all

Table 4
Attributes used in transit itinerary inference analyses.

Attribute	Description
GPS tracks attributes	
GPS_AVG_SPEED	GPS tracks average speed
GPS_TIME	Time interval between the first and last GPS track of a trip
GPS_AVG_DIST	Average distance between consecutive GPS point
Attributes from OTP	
OTP_LEN	Itinerary length
OTP_TRANS_TIME	Total transit time of each returned (by OTP) itinerary
OTP_WALK_TIME	Total walking time of each returned (by OTP) itinerary
OTP_WAIT_TIME	Total waitingtime of each returned (by OTP) itinerary
OTP_TIME	Total travel time of each returned (by OTP) itinerary
OTP_TRANS	Number of transfers along each returned (by OTP) itinerary
OTP_WALK_DIST	Walking distance of each returned (by OTP) itinerary
OTP_ORDER	The order of itinerary returned by OTP
OTP_AVG_SPEED	Itinerary average speed
Attributes from GPS Tracks and OTP	
DIFF_LEN	Difference between GPS tracks length and itinerary length
OVL_PERC	Overlapping percentage of itinerary and GPS tracks

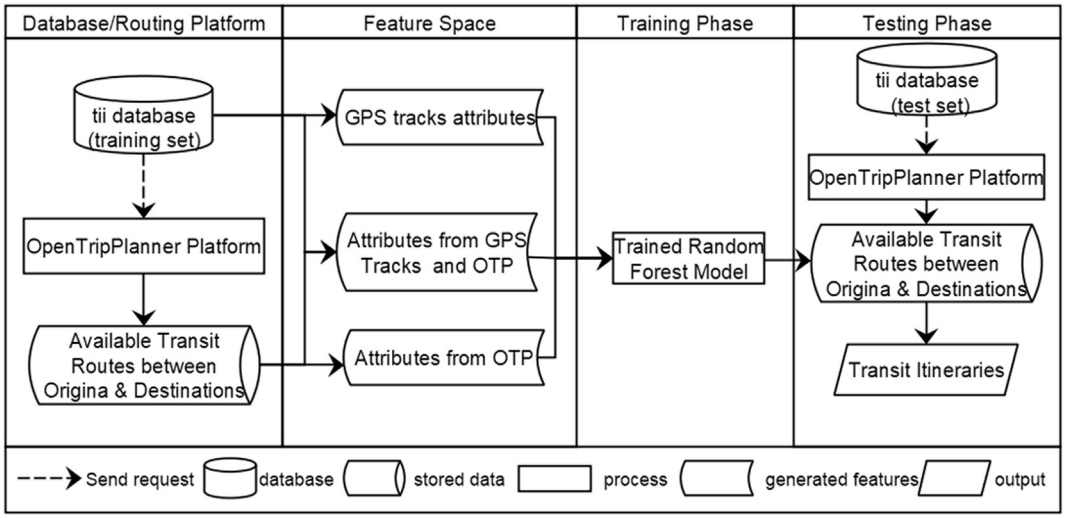

Fig. 3. Flow chart of transit itinerary inference model.

possible transit routes between each OD, sometimes it return unreasonable transit itineraries (e.g., far too long or circuitous) that can be easily rejected by setting a condition on the length or duration of transit itinerary. Hence, we decided to use a classifier, i.e. Random Forest, instead of a set of rules to identify transit itinerary. The attributes used in the RF model are shown in Table 4. The flow chart of the processing and estimation of the transit itinerary detection algorithm is shown in Fig. 3.

4.5. Activity detection

The flow chart of activity detection algorithm is shown in Fig. 4. The RF model is generated based on attributes shown in Table 5. The attributes come from three different data sources: MTL Trajet, land-use and Foursquare data. There are 61 attributes altogether, categorized into four major levels: (a) Socio-demographics, (b) Trip Characteristics, (c) Land-use, and (d) Foursquare.

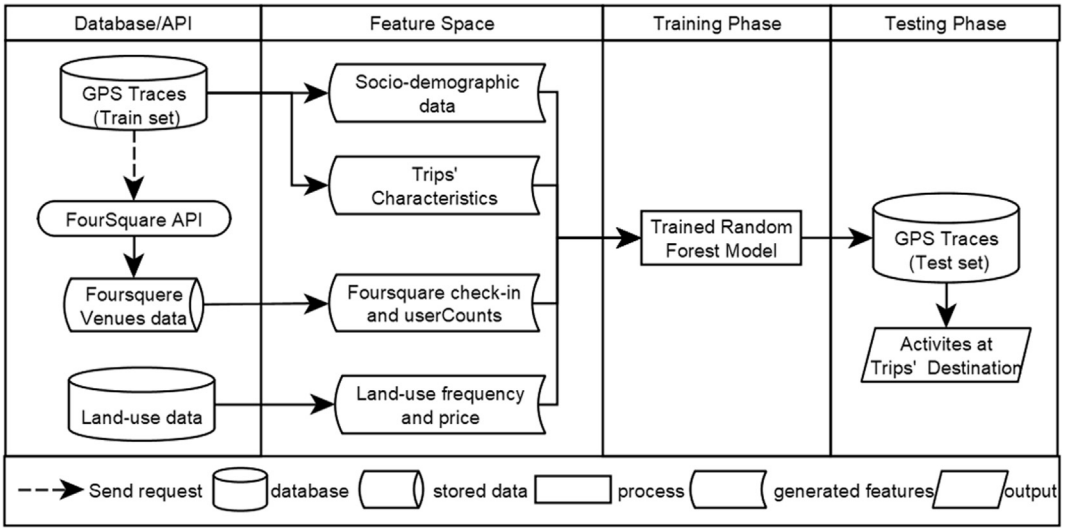

Fig. 4. Flow chart of activity detection model.

Table 5
Attributes used in activity detection analyses.

Attribute	Definition
Socio-demographics	
AGE	0: age between 16–24,1: 25–34, 2: 35–44, 3: 45–54, 4: 55–65, 5: 65+
AVG_PRICE_NEIGH	The average value of residential buildings around each individual's home
OCCUPATION	[l]0:full-time worker, 1: part-time worker, 2: Student,
3: Student and worker, 4: Retired 5: At home	
SEX	0:male, 1: female, 2: other/neither
Trip Characteristics	
DAY	1–7 for Monday through Sunday
CBD_ORIGIN	1: if the origin is located in Montreal's CBD, 0: otherwise
CBD_DESTIN	1: if the destination is located in Montreal's CBD, 0: otherwise
HOME_DEST	Direct distance between trip destination and individual home location
STUDY_DEST	Direct distance between trip destination and individual education location
WORK_DEST	Direct distance between trip destination and individual work location
HOME_ORG	Direct distance between trip origin and individual home location
STUDY_ORG	Direct distance between trip origin and individual education location
WORK_ORG	Direct distance between trip origin and individual education location
HOUR	Time of day from 0 to 24
MODE	Validated mode of transport via which the trip as been done
MTL_ORIGIN	1: if the origin is located in Montreal Island, 0: otherwise
MTL_DESTIN	1: if the destination is located in Montreal Island, 0: otherwise
TRAVEL_TIME	Total travel time of the trip
Land-use (number of land-use parcels in 250 meters around a trip destination)	
LU_*	23 different attributes each one shows the frequency of the corresponding land-use category
Foursquare (number of checkinCounts in 250 meters around a trip destination)	
CH_*	10 different attributes each one shows the checkinCounts for the corresponding Foursquare category
Foursquare (number of usersCounts in 250 meters around a trip destination)	
UC_*	10 different attributes each one shows the usersCounts for the corresponding Foursquare category

5. Model estimation

We use a serial approach to modelling the different trip characteristics.

5.1. Mode detection

As our basic/reference model, we started with the decision tree. However, as the prediction accuracy of decision tree models were around 70–80% and since the Random Forest model is an ensemble of decision trees and always achieves higher prediction accuracies, we continued our work on Random Forest models. To detect mode of transport, we generated Random

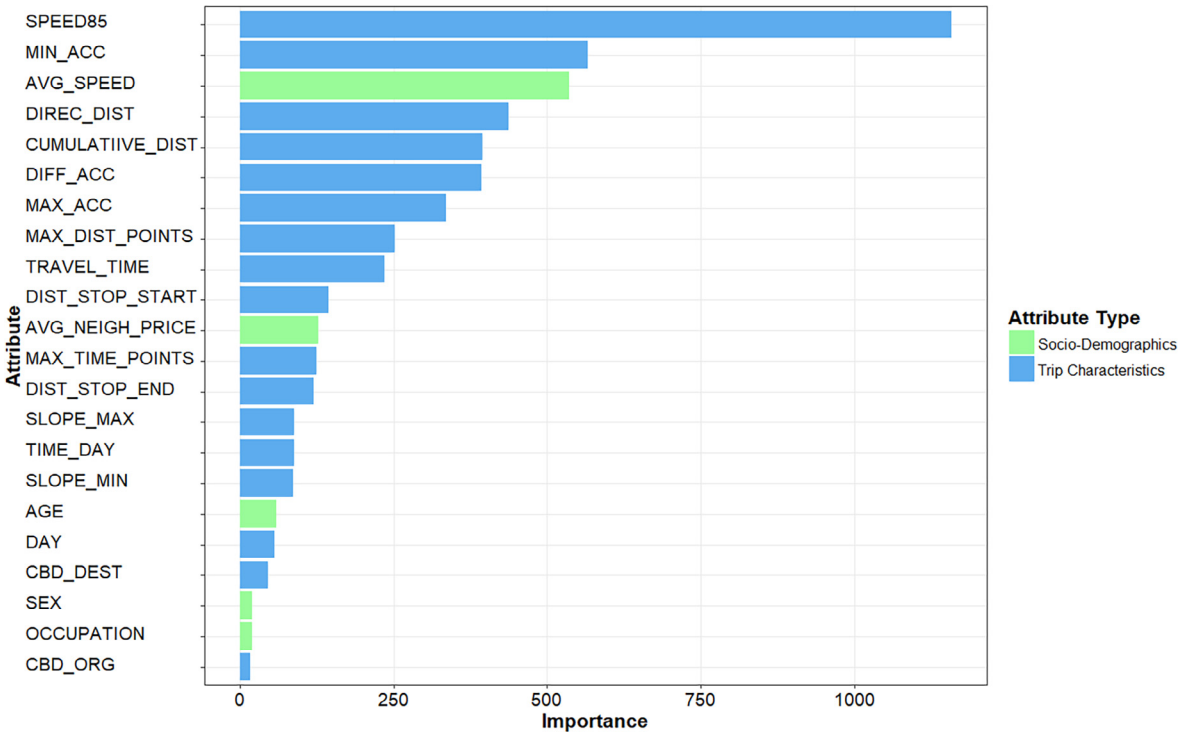

Fig. 5. Variable importance plot showing the mean decreases in predictive accuracy of mode detection model.

Forest models with different numbers of trees, ranging from 100 to 2000. We observed no change in variable importance or prediction accuracy for random forests with more than 1000 trees. Also, we allowed up to 8 attributes to be randomly sampled at each split based on previously published recommendations (Breiman, 2001; Liaw et al., 2002; Ermagun et al., 2017). Fig. 5 demonstrates the variable importance plot pointing to how important each variable in classifying trip mode. The variable importance values have been calculated by dividing the raw variable importance by the standard error, as explained in Section 4.2. Hence, we use the these values to test the null hypothesis of zero importance for the variables at 95 percent confidence level (i.e. $-z- \geqslant 1.96$). Obviously, for all the variables in the random forest we can reject the null hypothesis, as all the values in Fig. 5 exceed the z-score at 95 percent confidence value.

The most important variable, by far, is 85th percentile of speed. Average speed, direct distance, cumulative distance and "distance between trip origin and nearest transit stop" are the next most important variables. Among socio-demographics, average neighborhood price, as an indicator of income, is most important attribute.

5.2. Transit itinerary detection

We developed one binary RF classifier to detect true transit itinerary among a set of transit itineraries generated by OTP. The RF classifier was allowed to grow with 1000 trees, with 8 randomly sampled input attributes in each split. Fig. 6 demonstrates how important (as explained in Section 4.2) each attribute is to labeling the itineraries. The most important variable is overlap percentage between actual chosen itinerary and the itinerary produced by routing algorithm (OTP). The next most important variables are waiting time, in-vehicle transit time, walking distance and number of transfers along each transit itinerary, pointing to the fact that travelers tend to minimize the waiting and in-vehicle time, walking distance as well as number of transfers. Also, we tested the null hypothesis of zero importance for the variables in Fig. 6. As the importance of all the variables exceeds the z-score at 95 a percent confidence level, we can reject the null hypothesis.

5.3. Activity detection

To detect the activities at trip destinations a RF model was developed using the attributes listed in Table 3. The RF model was generated with the same values as the two previous RF models for hyperparameters. Fig. 7 shows the variable importance results of the RF model. Obviously, socio-demographic variables, i.e. age and average neighborhood price, as well as

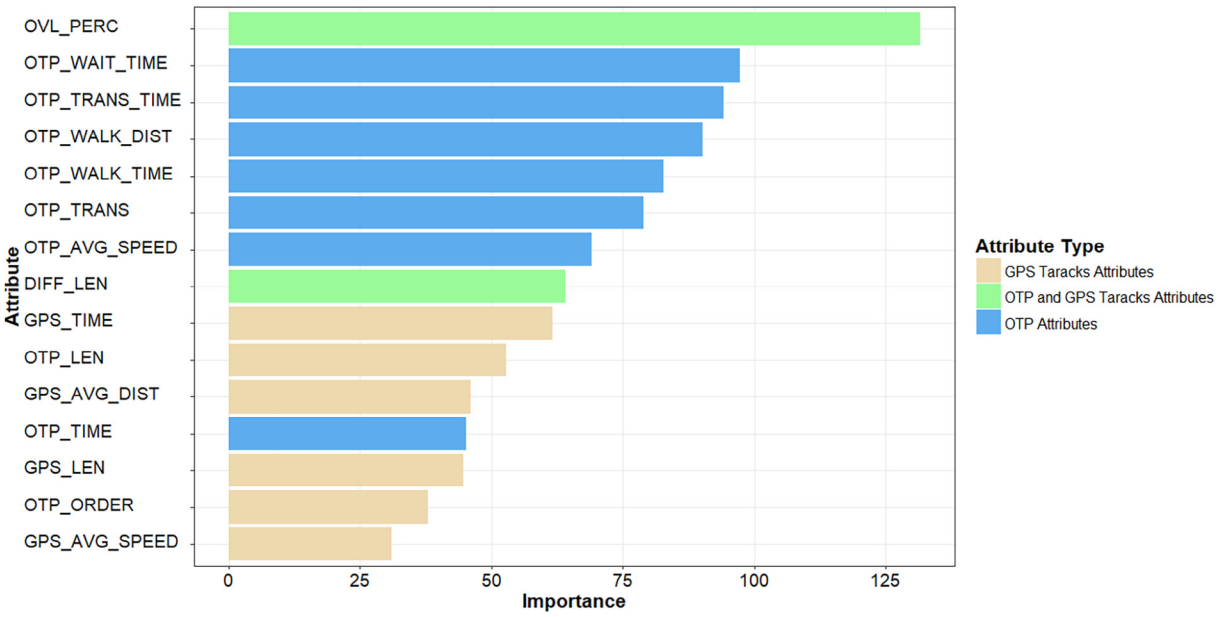

Fig. 6. Variable importance plot showing the mean decreases in predictive accuracy of transit itinerary inference.

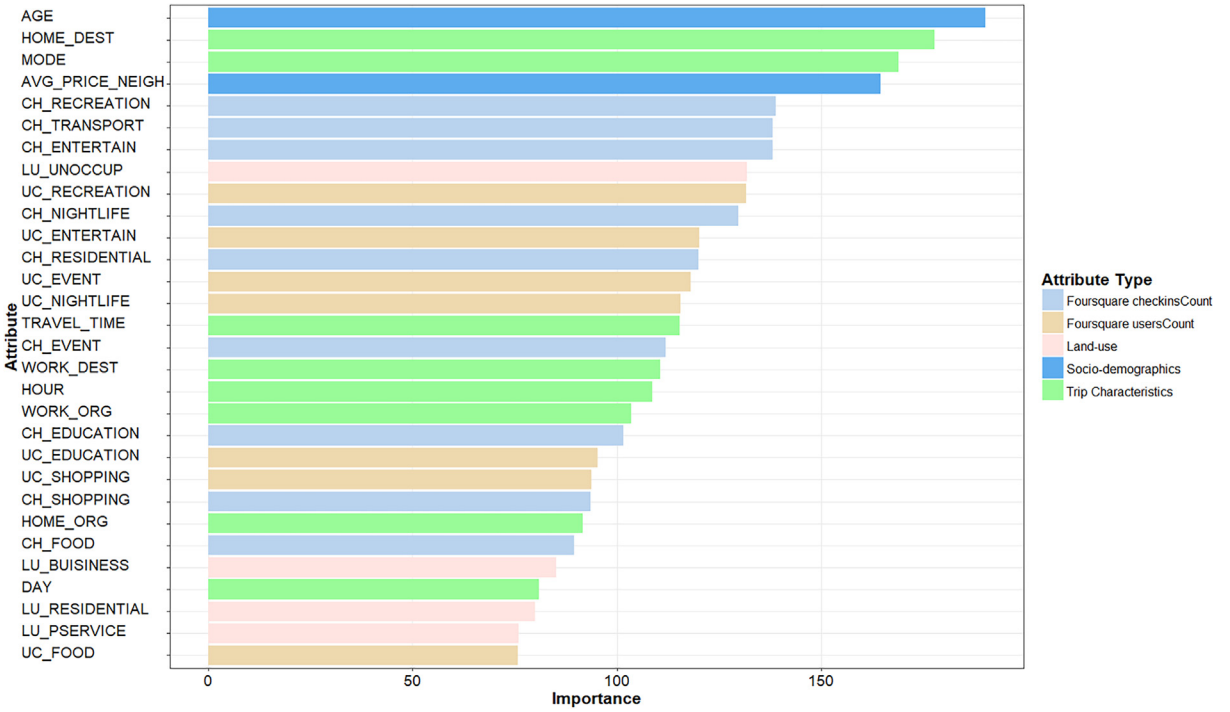

Fig. 7. Variable importance plot showing the mean decreases in predictive accuracy of activity detection model (the first 30 important variables).

mode and "distance between destination and individual's home" are among the most important variables. Foursquare checkinCounts for "recreation," "transport infrastructure" as well as "art and entertainment" follow in importance. The next important variables are unoccupied land-uses and Foursquare usersCounts for "recreation". Also, the null hypothesis of zero importance for the variables in Fig. 7 is rejected at the 95 percent confidence level.

In the next section, we have implemented a cross-validation procedure to validated the generated RF models.

6. Discussion

This section presents the predictive accuracy of the different models as well as comparing the results with previous studies in the literature.

6.1. Prediction accuracy assessment

To assess the predictive performance of the RF models, k-fold cross-validation was used, where k = 10, as commonly adopted in the literature (Ermagun et al., 2017; Gonzalez et al., 2010). The cumulative results of the 10-fold cross-validation for the RF models have been presented in a confusion matrix. Tables 6 to 8 show the analysis for the mode, transit itinerary and activity detection Random Forest models, respectively.

The confusion matrices used in this section were generated based on the methodology described by Aggarwal (2015). Each column of a confusion matrix shows the total actual number of observations for a given class over 10 iterations. In other words, each column indicates the total number of observations for a given class in the original dataset. On the other hand, each row is equal to the number of observations predicted by a prediction model to be of a given class (Ermagun et al., 2017). Each confusion matrix also presents two complementary measures to assess the per-class accuracy of the predictive model: precision (positive predictive value) and recall (sensitivity) (Aggarwal, 2015). The precision value for a given class i indicates the percentage of observations predicted by the model as class i that are actually of class i in the test dataset. Also, recall that the value for a given class i represents the percentage of actual observations of class i that have been predicted by the model as class i.

For illustration, looking at the very first column of Table 6 indicates that of the actual 1,176 walk trips the mode detection RF model predicted 933 trips as walk, 41 trips as bike, 108 trips as public transit, 92 trips as car and 2 trips as "car and public transit". This indicates that 79.34% of the actual walk trips have been predicted correctly by the model. Similarly, looking at the very first row of Table 6 we can see that the mode detection model predicted 1,127 trips to be walk trips, out of which 933 were truly walk, 41 trips were bike, 95 trips were public transit, 53 trips were car and 5 trips were "car and public transit." This corresponds to a 82.79% precision performance of the mode detection model for walking. Overall, the prediction performance of "car," "public transit" and "walk" trip modes is high both in terms of precision and recall. As we expected "Car and public transit" trip mode has the lowest precision and recall values, indicating this trip mode has similar attribute values to both "car" and "public transit." This caused these trips to be wrongly classified as car or public transit relatively often. Looking at the "Car and public transit" column, we see that of the 198 "Car and public transit" trips, 67 were correctly detected, While 83 trips were detected as car and 37 trips were detected as public transit, resulting in a low recall rate of 33.84%.

Table 6
Confusion matrix analysis for mode detection model.

Trip mode	Walk	Bike	Public transit	Car	Car and public transit	Precision (%)
Walk	933	41	95	53	5	82.79
Bike	41	527	31	48	6	80.70
Public transit	108	42	1907	146	83	83.42
Car	92	47	279	4660	37	91.10
Car and public transit	2	0	20	7	67	69.79
Recall (%)	79.34	80.21	81.78	94.83	33.84	

Table 7
Confusion matrix analysis for transit itinerary inference model.

Class	False Itinerary	True Itinerary	Precision (%)
False Itinerary	566	111	83.60
True Itinerary	111	488	81.47
Recall (%)	83.60	81.47	

Table 8
Confusion matrix analysis for activity detection model.

Activity	Education	Health	Leisure	Shopping/errands	Return home	Work	Precision (%)
Education	4426	219	485	478	241	212	73.02
Health	61	1902	74	47	42	50	87.41
Leisure	677	609	11,487	1386	955	971	71.41
Shopping/errands	1070	1009	3778	22,656	2360	2747	67.39
Return home	615	464	1965	2633	16,474	2050	68.07
Work	328	513	1020	1656	862	16,382	78.91
Recall (%)	61.67	40.33	61.07	78.51	78.69	73.09	

With respect to transit itinerary detection (Table 7), two classes, i.e. true and false itineraries, were validated. The third column of Table 7 shows the total number of transit trips in TII dataset. Among the 599 transit trip segments in the TII dataset, 488 of them were validated correctly resulting in an 81.47% recall rate. Also, there were 677 not chosen, or false itineraries, among which 566 itineraries were correctly classified as False.

The confusion matrix for the activity detection model is shown in Table 8. The highest prediction accuracy belongs to the purpose health, with about 87% accuracy. However, looking at the "health" column in Table 8 indicates that among 4,716 trips with health as purpose in the MTL Trajet dataset, only 1,902 trips were correctly classified as "health" trips. This low recall rate tells us that a great amount of health purpose trips are classified wrongly, dominantly as shopping trips, maybe due to the similarities between the attributes of these two activities or maybe because many health-related activities are accompanied by some shopping activities causing respondents to incorrectly validate the health purpose trips as shopping trips. The return home trip purpose has the highest recall rate in Table 8, meaning that 78.69% of return home trips in the dataset are correctly classified. The lowest prediction accuracy belongs to the shopping/errands activities, with 67% accuracy rate. This low prediction accuracy may be due to the similarity between shopping/errands trips and leisure trips, causing the RF model to predict shopping/errands as leisure trips. Also, the model has predicted a considerable number of return home trips as shopping/errand trips. This error may due to the fact that individuals usually do their daily shoppings on their way to home from work and at grocery stores close to home.

Summarizing the results, the mode, transit itinerary and activity detection RF models demonstrates an overall accuracy of 87%, 81% and 71%, respectively.

6.2. Comparison with the previous studies

Comparing prediction accuracy rates across studies needs to be done carefully because of differences in sample sizes, number of classes, and the quality of data across studies (Ermagun et al., 2017). As mentioned in Section 2, many studies in the literature have been based on small or researcher-collected smartphone data. In addition, the validation process can also effect the prediction accuracy of the models. Validation processes in smartphone travel surveys are usually done through an in-app prompt, a prompted-recall survey or a combination of them. Validating trip characteristics through a prompted-recall surveys can enhance the accuracy of reported trip characteristics and has a critical effect on the prediction accuracy of classifiers. The recall is usually implemented on the internet, when respondents are required to correct their trip characteristics if necessary (Xiao et al., 2016). Surveyor-intervened prompted recall surveys have also been used in some studies (Xiao et al., 2016). In such surveys, the surveyors ask respondents by telephone to recall the details of their trips.

Also, the number of categories varies across studies, ranging from very coarse (e.g. motorized/stationary Eftekhari and Ghatee, 2016) to relatively fine (e.g. walk/bike/bus/metro/car Dabiri and Heaslip, 2018). Such differences are also observed across activity detection studies, ranging from very general (e.g. indoor/outdoor Wu et al., 2011) to fine (e.g. work/education/shopping/eating out/recreation/Personal business/return home Ermagun et al., 2017). Such differences need to be considered when comparing the prediction accuracy of classifiers.

Furthermore, smartphone travel surveys varies regarding data from mobile phone sensors, such as GPS, accelerometer, gyroscope, rotation vector and magnetometer (Jahangiri and Rakha, 2014; Eftekhari and Ghatee, 2016) all of which have been used when detecting mode of transport (Feng and Timmermans, 2013; Dabiri and Heaslip, 2018). However, the larger the number of sensors that are used by an app, the greater the impact on battery life on respondent devices (Xiao et al., 2015). In addition, middle to low end smart-phones are not usually equipped with the all sensors. Hence, for large-scale applications, using few sensors is often optimal (Bantis and Haworth, 2017; Xiao et al., 2015; Dabiri and Heaslip, 2018).

Taking these considerations into account, our results compare well with other research in the literature. With respect to mode detection, we have chosen studies that have used data from GPS sensors and whose validation process is similar to the in-app prompt validation process of the MTL Trajet data set. Dabiri and Heaslip (2018) have reported test accuracies of 84.8% and 78.1%, on the GeoLife dataset (Zheng et al., 2011), for their Convolutional Neural Network and Random Forest models, respectively. Their best model, is able to predict walk, bike, bus, driving and train, with 81.6%, 90.3%, 80.7%, 86.6%, 92.3% precision, respectively. Also, the study of Zheng et al. (2008), which is the first solid mode detection framework from 2008, reported accuracies of 89%, 86%, 66% and 65% for the walk, driving, bus and bike modes, respectively. Their decision tree-based inference model, results in an overall accuracy of 76.2%. Both the aforementioned studies (Dabiri and Heaslip, 2018; Zheng et al., 2008) have used GPS data from mobile phones, similar to the MTL Trajet. Moroever, they have small sample sizes of 69 and 65 persons, respectively.

The random forest model in this study shows an overall accuracy of 87% and can predict walk, bike, public transit, car as well as car and public transit with 82.79%, 80.70%, 83.42%, 91.10% and 69.79% precision, respectively. The results demonstrates that overall accuracy of our random forest model is higher than the overall accuracy of both aforementioned comparable studies, although regarding bike trips our accuracy is lower than the corresponding accuracy in Dabiri and Heaslip's study and higher than that of Zheng et al. (2008).

Also, Gong et al. (2014) produce similar accuracy rates of 92% for car trips and 81% for bus trips, where our mode detection RF model predicts car trips with 90% and public transit trips with 83% accuracy rate.

Bantis and Haworth (2017) have developed several models, including a random forest model, to detect three modes of transport (walk, bus/car and train) as well as stationary points from GPS and accelerometer data as well as users

characteristics. Their training data contains trips from 5 individuals. They have proposed a hierarchical dynamic Bayesian network and compared the results against random forest, SVM, and Multilayer Perceptron classifiers. The accuracy of their proposed model is 90%, while it can predict Stationary status as well as Walk, Bus/Car and Rail trips with 94%, 65%, 88%, 91% precision, respectively. Although their categories are different from ours and the sample size of two studies are not similar, our random forest model predicted walking trips with higher precision, i.e. 65% vs. 82.79%. However, the overall prediction accuracy of our model is about 4% lower than their result, maybe due to lack of accelerometer data in our data set.

Eftekhari and Ghatee (2016) have applied different models to detect the movement and the stationary statuses in the motorized and non-motorized modes on a smartphone travel survey of 9 users. Their data set contains data from accelerometer, magnetometer and gyroscope sensors. Their best inference model can recognize the motorized mode with 95.2% accuracy. However, their result is not comparable with the results of this study or other studies with dissimilar categories.

We could not find any predicting accuracy for "car and public transit" trips in the literature. Our random forest model can predict this mode with 69.79% precision, which shows the more work is needed regarding detecting multimodal trips.

With respect to activity detection, the random forest model developed by Ermagun et al. (2017) produced an overall accuracy of 64% for predicting 5 trip purposes.They have reported the precision of 50.10%, 61.49%, 51.92%, 55.62% and 47.04% for predicting Eat out, Education, Personal business, Shopping, as well as Social, recreation, and community, respectively. Our random forest model is able to predict Education, Shopping/errand and Leisure with 73.02%, 67.39% and 71.41% precision, respectively and shows an overall accuracy of 71%. Also, the RF model generated by Oliveira et al. (2014) predicts activities with a 65% accuracy rate, which is lower than the result in this study.

Xiao et al. (2016) have applied artificial neural networks combined with particle swarm optimization on a surveyor-intervened prompted recall survey to detect 5 trip purposes (Home, Work/education, Eating out, Shopping, Social visit, Picking up/dropping off someone). Their sample size is 352 respondents and their proposed model shows the accuracy of 96.5%, which is higher from our model. We think that, besides the differences between modelling approaches, the high quality of their validated data, gathered by a surveyor-intervened prompted recall survey, contributes to the high prediction accuracy of their classifier.

Regarding transit itinerary detection, we found only one study in the literature, that of Zahabi et al. (2017), who reported an accuracy of 87% of for their rule-based transit itinerary detection algorithm. The study reported correctly predicted distance of transit trips while our random forest model uses trip segment as unit of analysis. As such, the results of the two studies are not perfectly comparable.

7. Conclusion

This research contributes to the literature on GPS-based travel diary surveys in four important ways: (1) it demonstrates the potential of deriving complete trip information (i.e. not only commonly inferred trip characteristics such as mode, but also purpose and transit itinerary) from smartphone travel surveys; (2) it shows that socio-demographic characteristics of travellers can play an important role in predicting the mode and activity from smartphone travel surveys; (3) it also contributes to the literature of transit itinerary inference by introducing a new approach to infer transit itinerary from GPS data; and finally (4) it shows the advantages of using other complementary data sources such as location-based social network APIs, like Foursquare, or GTFS data alongside GPS traces to develop more accurate predicting models.

Although there are still some hurdles to overcome, we believe that smartphone travel surveys have the potential to replace traditional travel surveys. Here, we try to address some of these hurdles. First, large-scale smartphone travel surveys may not produce the same quality of validated data as small or researcher-collected smartphone travel surveys. Prompted-recall surveys can improve the quality of gathered data by reducing the self-report errors, however, this comes at the expense of rising the surveying cost and more burden on the respondents. Actually, the need for labelled data is a requirement of supervised machine learning models used in the literature and the current study. Using semi-supervised or unsupervised models may help to reduce or eliminate the need for labelled data. Although there are some studies (Rezaie et al., 2017) investigating semi-supervised approaches to detect mode of transport, more research is needed to adequately evaluate these methods in this context. Second, detecting multimodal trips is usually harder than detecting uni-mode trips from GPS traces. Future studies should also investigate multimodal trip detection methods. Third, the data from traditional household travel surveys are also have used in activity-based models (Bowman and Ben-Akiva, 2001) which need the information about tours of travellers as well as the sequence of their activities along a day. While a lot of studies in the literature aim to detect trip characteristics and single activities, there are few studies, like Lin et al. (2017), aim to infer activity sequences as well as tours from smartphone travel surveys. Hence, more research is needed in the field to make the data from smartphone travel surveys applicable for activity-based modelling approaches.

The benefits related to smartphone travel surveys, such as lower costs and less user burden, provide encouragement for future research in the area to improve the prediction performance of models as well as overcome the above mentioned hurdles along the way. The results of current research demonstrate the potential of smartphone travel survey apps to be used as a stand-alone large-scale household travel survey in the future. Also, this paper shows that the trip characteristics extracted from GPS traces have the potential to be as complete as those of traditional survey methods, such as CATI or paper-based household travel surveys.

Acknowledgments

This research is based upon work supported by the Social Sciences and Humanities Research Council of Canada as well as "MonRso Mobilit" research contract with Ville de Montreal.

References

Aggarwal, C.C., 2015. Data Mining: The Textbook. Springer.
Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Statist. Data Anal. 52 (4), 2249–2260.
Bantis, T., Haworth, J., 2017. Who you are is how you travel: a framework for transportation mode detection using individual and environmental characteristics. Transp. Res. Part C: Emerg. Technol. 80, 286–309.
Bellman, R., 1958. On a routing problem. Q. Appl. Math. 16 (1), 87–90.
Biljecki, F., 2010. Automatic segmentation and classification of movement trajectories for transportation modes.
Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: a large-scale application in the netherlands. Transp. Res. Part C: Emerg. Technol. 17 (3), 285–297.
Bowman, J.L., Ben-Akiva, M.E., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transp. Res. Part A: Policy Pract. 35 (1), 1–28.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, L., 2002. Manual on Setting Up, Using, and Understanding Random Forests v3. 1. Statistics Department University of California, Berkeley, CA, USA 1.
Breiman, L., Cutler, A., 2007. Random Forests-classification Description. Department of Statistics, Berkeley 2.
Breiman, L., 2003. Manual on setting up, using, and understanding random forests v4. 0. URL ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf 11.
Cambridge Systematics, 2012. Travel demand forecasting: parameters and techniques, vol. 716. Transportation Research Board.
Catala, M., Dowling, S., Hayward, D., 2011. Expanding the google transit feed specification to support operations and planning. Tech. rep.
Cebelak, M.K., 2015. Transportation planning via location-based social networking data: exploring many-to-many connections. Ph.D. thesis.
Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., Zegras, P., 2013. Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore. Transp. Res. Record: J. Transp. Res. Board (2354), 59–67.
Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from gps trajectories using a convolutional neural network. Transp. Res. Part C: Emerg. Technol. 86, 360–371.
Dalumpines, R., Scott, D.M., 2017. Making mode detection transferable: extracting activity and travel episodes from gps data using the multinomial logit model and python. Transp. Plan. Technol. 40 (5), 523–539.
Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. 40 (2), 139–157.
Eftekhari, H.R., Ghatee, M., 2016. An inference engine for smartphones to preprocess data and detect stationary and transportation modes. Transp. Res. Part C: Emerg. Technol. 69, 313–327.
Endo, Y., Toda, H., Nishida, K., Kawanobe, A., 2016. Deep feature extraction from trajectories for transportation mode estimation. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 54–66.
Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., Das, K., 2017. Real-time trip purpose prediction using online location-based search and discovery services. Transp. Res. Part C: Emerg. Technol. 77, 96–112.
Feng, T., Timmermans, H.J., 2013. Transportation mode recognition using gps and accelerometer data. Transp. Res. Part C: Emerg. Technol. 37, 118–130.
Foursquare for Developers: Venue Categories, 2017. https://developer.foursquare.com/docs/venues/categories, [Accessed on June 12, 2017].
Foursquare for developers: Venue response, 2017. https://developer.foursquare.com/docs/responses/venue, [Accessed on June 12, 2017].
Get Elevations – MSDN, 2017. https://msdn.microsoft.com/en-us/library/jj158961.aspx, [Accessed on October 20, 2017].
Ghasri, M., Rashidi, T.H., Waller, S.T., 2017. Developing a disaggregate travel demand system of models using data mining techniques. Transp. Res. Part A: Policy Pract. 105, 138–153.
Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving personal trip data from gps data: a literature review on the existing methodologies. Procedia-Social Behav. Sci. 138, 557–565.
Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J., Labrador, M.A., Winters, P.L., Georggi, N.L., Perez, R., 2010. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. IET Intel. Transport Syst. 4 (1), 37–49.
Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning, vol. 1. MIT Press, Cambridge.
Greene, E., Flake, L., Hathaway, K., Geilich, M., 2016. A seven-day smartphone-based gps household travel survey in indiana 2. In: Transportation Research Board 95th Annual Meeting, no. 16-6274.
Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C: Emerg. Technol. 44, 363–381.
Hasan, S., Ukkusuri, S.V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. PLoS One 10 (5). e0124819.
Jahangiri, A., Rakha, H., 2014. Developing a support vector machine (svm) classifier for transportation mode identification by using mobile phone sensor data. In: Transportation Research Board 93rd Annual Meeting, no. 14-1442.
Kim, Y., Pereira, F.C., Zhao, F., Ghorpade, A., Zegras, P.C., Ben-Akiva, M., 2015. Activity recognition for a smartphone and web based travel survey, arXiv preprint 1502.03634.
Lee, J.H., Davis, A.W., Yoon, S.Y., Goulias, K.G., 2016. Activity space estimation with longitudinal observations of social media data. Transportation 43 (6), 955–977.
Lian, D., Xie, X., 2011. Collaborative activity recognition via check-in history. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM, pp. 45–48.
Liaw, A., Wiener, M., et al, 2002. Classification and regression by randomforest. R News 2 (3), 18–22.
Lin, Z., Yin, M., Feygin, S., Sheehan, M., Paiement, J.-F., Pozdnoukhov, A., 2017. Deep generative models of urban mobility.
Lu, Y., Liu, Y., 2012. Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. Comput. Environ. Urban Syst. 36 (2), 105–108.
MAMROT. Localisation des immeubles 2011. Ministere des Affaires Municipales, des Regions et de l'Occupations du Territoire (MAMROT).
Mandow, L., De la Cruz, J.P., et al., 2005. A new approach to multiobjective a* search. In: IJCAI, vol. 8.
Montini, L., Rieser-Schüssler, N., Horni, A., Axhausen, K., 2014. Trip purpose identification from gps tracks. Transp. Res. Record: J. Transp. Res. Board (2405), 16–23.
Montini, L., Prost, S., Schrammel, J., Rieser-Schssler, N., Axhausen, K.W., 2015. Comparison of travel diaries generated from smartphone data and dedicated gps devices. Transp. Res. Procedia 11 (Supplement C), 227–241.
Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P., 2014. Supporting large-scale travel surveys with smartphones–a practical approach. Transp. Res. Part C: Emerg. Technol. 43, 212–221.
Oliveira, M., Vovsha, P., Wolf, J., Mitchell, M., 2014. Evaluation of two methods for identifying trip purpose in gps-based household travel surveys. Transp. Res. Record: J. Transp. Res. Board (2405), 33–41.

OpenStreetMap contributors, 2017. Planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org.
Opentripplanner bibliography, 2017. http://docs.opentripplanner.org/en/latest/Bibliography/, [Accessed on July 18, 2017].
Patterson, Z., 2017. The itinerum open smartphone travel survey platform. Technical Report. Concordia University TRIP Lab, Montreal, Canada, TRIP Lab Working Paper 2017-1 (July 2017).
Patterson, Z., Fitzsimmons, K., 2016. Datamobile: Smartphone travel survey experiment. Transp. Res. Record: J. Transp. Res. Board (2594), 35–43.
Patterson, Z., Fitzsimmons, K., 2017. MTL Trajet. Working Paper 2017-2. Concordia University. TRIP Lab, Montreal, Canada.
Pearson, D., 2004. A comparison of trip determination methods in gps-enhanced household travel surveys. In: 84th Annual Meeting of the Transportation Research Board, Washington, DC.
Pereira, F., Carrion, C., Zhao, F., Cottrill, C.D., Zegras, C., Ben-Akiva, M., 2013. The future mobility survey: overview and preliminary evaluation. In: Proceedings of the Eastern Asia Society for Transportation Studies, vol. 9.
Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. Transp. Res. Part C: Emerg. Technol. 75, 197–211.
R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Rezaie, M., Patterson, Z., Yu, J.Y., Yazdizadeh, A., 2017. Semi-supervised travel mode detection from smartphone data. In: 2017 International Smart Cities Conference (ISC2). IEEE, pp. 1–8.
Schuessler, N., Axhausen, K., 2105. Processing raw data from global positioning systems without additional information. Transp. Res. Record: J. Transp. Res. Board (2105), 28–36.
Shafique, M.A., Hato, E., 2015. Use of acceleration data for transportation mode prediction. Transportation 42 (1), 163–188.
Shafique, M.A., Hato, E., 2016. Travel mode detection with varying smartphone data collection frequencies. Sensors 16 (5), 716.
Shen, L., Stopher, P.R., 2014. Review of gps travel survey and gps data-processing methods. Transport Rev. 34 (3), 316–334.
Skiena, S., 1990. Dijkstra's Algorithm, Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Addison-Wesley, Reading, MA, pp. 225–227.
Sohn, T., Varshavsky, A., LaMarca, A., Chen, M.Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W.G., De Lara, E., 2006. Mobility detection using everyday gsm traces. In: International Conference 18 on Ubiquitous Computing. Springer, pp. 212–224.
Stenneth, L., Wolfson, O., Yu, P.S., Xu, B., 2011. Transportation mode detection using mobile phones and gis information. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 54–63.
Stewart, B.S., White III, C.C., 1991. Multiobjective a. J. ACM (JACM) 38 (4), 775–814.
Stopher, P.R., 2009. The travel survey toolkit: where to from here? In: Transport Survey Methods: Keeping Up with a Changing World. Emerald Group Publishing Limited, pp. 15–46.
Stopher, P., Zhang, Y., Zhang, J., Halling, B., 2009. Results of an evaluation of travelsmart in south australia.
Sun, Z., Ban, X.J., 2013. Vehicle classification using gps data. Transp. Res. Part C: Emerg. Technol. 37, 102–117.
Ta, N., Kwan, M.-P., Chai, Y., Liu, Z., 2016. Gendered space-time constraints, activity participation and household structure: a case study using a gps-based activity survey in suburban Beijing, China. Tijdschrift voor economische en sociale geografie 107 (5), 505–521.
Ville de Montreal, 2017. MTL Trajet. https://ville.montreal.qc.ca/mtltrajet/en/, [Accessed on June 15, 2017].
Wang, H., Liu, G., Duan, J., Zhang, L., 2017. Detecting transportation modes using deep neural network. IEICE Trans. Inf. Syst. 100 (5), 1132–1135.
Willumsen, L.G. et al, 2011. Modelling Transport. John Wiley & Sons.
Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
Wolf, J.L., 2000. Using gps data loggers to replace travel diaries in the collection of travel data (Ph.D. thesis). School of Civil and Environmental Engineering, Georgia Institute of Technology.
Wolf, J., Guensler, R., Bachman, W., 1768. Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. Transp. Res. Record: J. Transp. Res. Board (1768), 125–134.
Wolf, J., Oliveira, M., Thompson, M., 2003. Impact of underreporting on mileage and travel time estimates: results from global positioning system-enhanced household travel survey. Transp. Res. Record: J. Transp. Res. Board 1854, 189–198.
Wu, J., Jiang, C., Houston, D., Baker, D., Delfino, R., 2011. Automated time activity classification based on global positioning system (gps) tracking data. Environ. Health 10 (1), 101.
Wu, L., Yang, B., Jing, P., 2016. Travel mode detection based on gps raw data collected by smartphones: a systematic review of the existing methodologies. Information 7 (4), 67.
Xiao, G., Juan, Z., Gao, J., 2015. Travel mode detection based on neural networks and particle swarm optimization. Information 6 (3), 522–535.
Xiao, G., Juan, Z., Zhang, C., 2016. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transp. Res. Part C: Emerg. Technol. 71, 447–463.
Xiao, Z., Wang, Y., Fu, K., Wu, F., 2017. Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. ISPRS Int. J. Geo-Inf. 6 (2), 57.
Zahabi, S.A.H., Ajzachi, A., Patterson, Z., 2017. Transit trip itinerary inference with gtfs and smartphone data. Tech. rep.
Zhang, H., Singer, B., 2010. Recursive Partitioning and Applications. Springer Science & Business Media.
Zhang, Z., He, Q., Zhu, S., 2017. Potentials of using social media to infer the longitudinal travel behavior: a sequential model-based clustering method. Transp. Res. Part C: Emerg. Technol. 85, 396–414.
Zhao, F., Ghorpade, A., Pereira, F.C., Zegras, C., Ben-Akiva, M., 2015. Stop detection in smartphone-based travel surveys. Transp. Res. Procedia 11, 218–226.
Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.-Y., 2008. Understanding mobility based on gps data. In: Proceedings of the 10th International Conference on Ubiquitous Computing. ACM, pp. 312–321.
Zheng, Y., Fu, H., Xie, X., Ma, W.-Y., Li, Q., 2011. Geolife GPS trajectory dataset – User Guide. URL https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/.
Zhou, C., Jia, H., Juan, Z., Fu, X., Xiao, G., 2017. A data-driven method for trip ends identification using large-scale smartphone-based gps tracking data. IEEE Trans. Intell. Transp. Syst. 18 (8), 2096–2110.
Zhu, Z., Blanke, U., Tröster, G., 2014. Inferring travel purpose from crowd-augmented human mobility data. In: Proceedings of the First International Conference on IoT in Urban Space. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 44–49.