# Chapter 2. Literature Review

## 2.1. Trip purpose classification

Although a wealth of literature exists regarding the classification of transport mode derived from GPS traces, investigation into the transport purpose from GPS has received far less attention (Yazdizadeh *et al.*, 2019). One reason for this is that the inference of trip purpose in models require that users provide information about why they have actual made a trip to compliment any raw GPS (Gong *et al.*, 2014). Of the classification models which focus on purpose classification, Gong *et al.* (2014) characterise three distinct types:

1. Rule-based (using rules to match GPS signal and ﻿respondents' information),

2. Probabilistic (using the calculated probability of a given purpose);

3. Machine learning.

- Mode classification, on the other hand, can be applied with non-user inputted indicators such acceleration and distance (ref).

The trend in the literature, has been to use these types of models in combination with inputs containing a high number of dimensions (ref). Generally, these input features are then reduced by evaluating the feature importance ref). In the last few years, methods employing ensemble decision trees such as Random Forest classifiers have hence become more popular (Gong *et al.*, 2018). The preference for an approach whereby a lot of variables are considered at first before reduction is likely due to a lack of understanding of specific dynamics which govern why people make trips – a major gap in the research of trip purpose classification (Meng *et al.*, 2019). Moreover, ML methods, as opposed to probabilistic and rule-based methods, are generally better at accounting for hidden relationships in the data (Li *et al.*, 2016). A selection of key classification models from the literature are detailed along with their inputs and accuracy in **Table 2.1**.

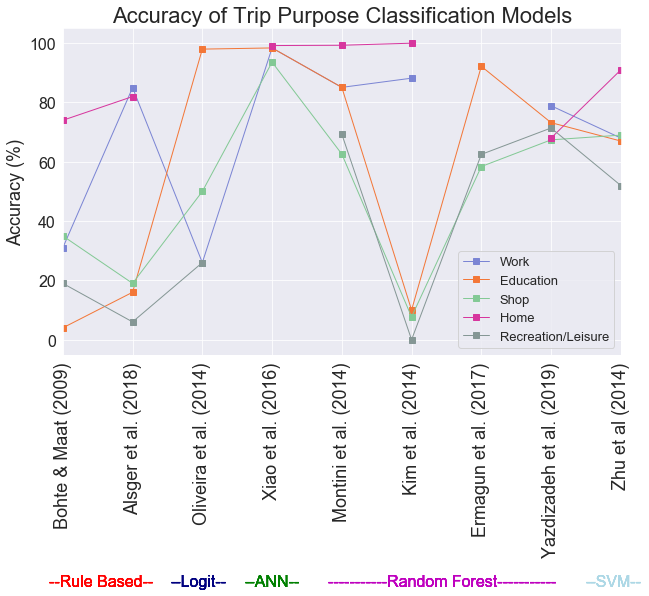
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Author(s)* | *Predictor variables* | *Location and year* | *Number of Trips included in Study* | *Overall classification accuracy* |
| *Rule-Based Methods* | | | | |
| Bohte & Maat (2009) | POI; Personal Locations Proximity | Netherlands, 2007 | ﻿ 33,686 | 43% |
| Alsger *et al.* (2018) | Activity Duration; Land Use; Temporal Features; Trip Frequency | ﻿Queensland, Australia,  2009-2012 | ﻿65,000 | 78% |
| Probabilistic Methods: Multinomial Logit Models | | | | |
| Oliveria *et al.* (2014) | Duration; Mode; Land use; Personal Location proximity | Georgia, USA 2011 | 10,512 | 70% |
| *Machine Learning Methods: Artificial Neural Networks* | | | | |
| Xiao *et al.* (2016) | Land use; POI | Shanghai, China,  ﻿2013-2015. | 7,039 | 96.5% |
| *Machine Learning Methods: Random Forest and Decision Tree Models* | | | | |
| Montini *et al.* (2014) | Land Use; Personal Location Proximity; Socio-demographics; Temporal Features | Zurich, Switzerland, 2012 | 6,938 | 80% |
| Kim *et al.* (2015) | POI; Socio-demographics | Singapore, 2013 | 7,856 | 75.5% |
| Ermagun *et al.* (2017) | POI;  Socio-demographics; Temporal Features;  Travel Mode | Minnesota & Iowa, USA, 2010-2012 | 58,503 | 64% |
| Yadizadeh *et al.* (2019) | Socio-demographics; Personal Location Proximity; Temporal Features; Land Use; Foursquare check-ins | Montreal, Canada, 2016 | ﻿131,777 | 71% |
| *Machine Learning Methods: Support Vector Machines* | | | | |
| Zhu *et al*. (2014) | Foursquare check-ins; Socio-demographics; Temporal Features | Washington State, USA Spring 2014 | ﻿﻿87,600 | 75% |

**Table 2.1** Overview of classification models used in the literature to predict trip purpose (*POI=Points of Interest*).

Inputs used in trip purpose models detailed in **Table 2.1** typically include a combination of user-inputted information and underlying spatial (distance to respondent’s home/work places; POI; Land Usage), temporal (time of day; day of week) and socio-demographic (age; gender; occupation) features. The models are shown to vary in accuracy between 43–96.6% and have been built on a range of different data sizes (7,039–131,777 trips) on different years and area. As a result, significant uncertainties have been raised around the cross-comparability of trip purpose studies, with any findings being tied to specific locations and times (Jahromi *et al*., 2016).

There is also disparity in the accuracy of the classification models based on individual purpose classes. As shown in **Figure 2.1**, the models detailed in **Table 2.1** have broadly struggled in classifying shopping and leisure activities versus activities where of education, work and returning home. Arguably, shopping and leisure activities may tend to be less most temporally and spatially regular as compared to work, education and home locations so this warrants further investigation (ref; ref).

[As opposed to leisure and health, etc.]. individuals’ mobility is found to be highly regular (Lin & Hsu, 2014). Zhu et al -> multi-class one vs all for each



**Figure 2.1** Comparison of trip purpose classification model accuracy within the literature (ANN=Artificial Neural Network; SVM=Support Vector Machine)

Generally, spatial and temporal features have been identified as the key indicators in trip purpose classification, however these have not been applied to any standard throughout the literature (Aslger *et al.*, 2018). A varied amount of importance has been applied to these features and their representation in the models. In some cases, purely distance to POI (e.g. Ermugun *et al.*, 2017). In other cases, more attention has been employed to the importance of these features i.e. clustering techniques have been used in studies by Montini *et al.* (2014) & Kim *et al.* (2015) to better group origin and destination of trips and improve the generalisation ability of the models.

Varied amount of trip purposes -> Jahromi *et al*., 2016

Overall, it has mostly been found that socio-demographic features are less important in the improvement of purpose classification (Montini et al., 2014; ref). Arguably, this finding in the literature may relate is a as when people travel they often pass through and by a range of areas, POI and neighbourhoods (Kwan, 2018). Notably, socio-demographic data has been used as key identifiers in other areas of mobility studies i.e. in mode classification and the predicting of when and how people travel around cities (Xie *et al.* 2016; Bantis & Haworth, 2017).

- Aslger *et al.* (2018) break down influence of individual temporal and spatial indicators of the trips and highlights temporal features to be importance in the classification accuracy across a range of trip purposes.

### 2.1.1 Key issues raised by existing trip purpose research

As evident from a review of the literature, a larger variety of spatial information has been integrated in models than temporal information. The wide range of metrics to account for spatial context such as land use, nearby POIs and Foursquare check-ins have outweighed metrics of temporal importance which are restricted to day of week and time of day. Moreover, there is less attention which has paid to studying the overall temporal profiles of different types of trip purposes (Meng *et al.*, 2019). Further, there is little investigation into the longer term effects and seasonality of the models, which could discount any findings at longer time periods within cities. Gong *et al.* (2018) find seasonality can severely affect accuracy of models which focus on mobility as people tend to change travel patterns and carry out different activities to account for weather.

Xie *et al.* (2016) weather changes how people travel

There is also, a lack of investigation into the spatiality of error terms available in other forms of mobility studies. Semanjski *et al* (2017) use land use to indicate accuracy of classification (more accurate in rural areas)

Finally, there is an inconsistency with the way that the modelling procedure has been approached throughout the literature, with some studies focusing on building individual models for each purpose and others with all-encompassing multi-class classification models [Something about no one method being better].

Further, a range of approaches have been taken with the training and testing of the models – with the majority of the studies ignoring underlying class imbalance present within them (Meng *et al.*, 2019). With work and educational trips generally outweighing shopping and recreational trips across the studies. Consideration of this imbalance has been taken Xiao *et al.* (2016) who use under-sampling technique to train a feed-forward neural network model on an equal number of samples for each purpose. Cross validation

Although, some evidence of spatial clusters being used as explanatory variables (e.g. Montini *et al.,* 2014; Yazdizadeh *et al.*, 2019), less attention paid to temporal clusters. Other forms of mobility research, have also seen less consideration to temporal clusters, although a 2018 study by Liu & Cheng adapt a Latent Dirichlet Allocation model to better account for temporal structures in movements from smart card data.

Research carried out by Zhang & Cheng (2019) discover expected difference in the profiles of people travelling within London based on their employment status. In general, finding regularity in full-time transport patterns compared with those who are un-employed. While, this information is of use to transport authorities, there is still a lack of investigation into more of the local impacts of transport. Insight into which activities occur on which days and times (similar to Zhang & Cheng, 2019). -> (lead onto Batty, 2013)

## 2.2 Volunteered Geographic Information (VGI)

VGI, first described by (Goodchild, 2007),

“﻿the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information” (Goodchild, 2007),

Whether that be geo-tags, georeferences, GPS, (Goodchild, 2007).

“humans as sensors” (Goodchild, 2007, pp. 218)

“﻿The term citizen science is often used to describe communities or networks of citizens who act as observers in some domain of science” (Goodchild, 2007)

New nexus, between data science and VGI (Burini *et al.*, 2017)

Potentially to cross-analyze multisource *big data*” (Burini *et al.*, 2017)

VGI has gained an increasing amount of attention in the literature. Partly due to Crowd-sourcing, mobile phones cost effective (Shi *et al.*, 2018).

OpenStreet Map contributions and crowdsourcing (Goodchild & Li, 2012)

Near-real time (Aubrecht *et al.*, 2011)

“﻿VGI is defined as the subset of user- generated content (UGC) with a geographic reference

(Goodchild 2007).”

Capturing space-time structures (Arribas-Bel & Tranos, 2017)

‘People as sensors’ (ref). “there is an opportunity for smartphones to replace dedicated GPS devices” (Wu *et al.*, 2016)

More User generated content online (Flanagin & Metzger, 2008)

“﻿several instances of VGI involve perceptual information that can only be reliably known and communicated by ‘‘locals’” (Flanagin & Metzger, 2008)

We can capture ﻿“ both mundane and unplanned events” Miller & Goodchild (2014)

“﻿GPS technology allows for comprehensive tracking and sharing of location and route information” (Bricka *et al.*, 2015)

Where users can “opt to share” Geographic information (Elwood *et al.*, 2012)

Populations not samples (Miller & Goodchild (2014))

Detecting “﻿accelerometer sensor, magnetometer sensor in Android-based smartphones.” (Gong *et al.*, 2014)

Li *et al.* (2016) distinguish between two types of VGI participatory (conscious inclusion of their data) and opportunistic (unconscious) forms of VGI. Inferring information (Tu *et al.*, 2017). Through this, VGI can give us insight into processes occurring in space time that more traditionally collected information cannot (Elwood *et al.*, 2012). Indeed, understanding the space and time structures in cities help us understand them better (Chen*g et al.* 2017).

“understanding the interactions within complex system such as a city is a prerequisite for predicting changes within it -> big data gives us an opportunity” Chen*g et al.* 2017

Governments pushing for greater capacity of transport networks rather than efficiency (Attard *et al.*, 2016).

Elwood *et al.* (2012) find that VGI gives us insight that other forms of data do not such as subjectivity which is tied to space (i.e. like purpose of trips).

Tu *et al.* (2017) look at activities relating to mobile phone GPS tracts using social media data, check-ins and a hidden Markov model.

### 2.2.1 The use of VGI for mobility studies

It is only in recent years, with the explosion of VGI collected from smartphones and mobile app surveys that has meant increasingly research using data these sources has become more wide spread within mobility studies (Kim *et al.*, 2015). “﻿The development and use of smartphone travel surveys is opening opportunities to better understanding of travel behavior because of the ability to collect detailed (and previously unavailable) information about people’s travel itineraries.” (Zahabi *et al.*, 2017).

high quality traveller information for public transport is undoubtedly crucial for Government’s transport policies (Lyons & Harman, 2002)

Understanding mobility through mobile phone has kicked off (Zhao *et al.*, 2019)

Less hassle for users, (than traditional travel surverys), thus can reach larger audiences and for longer times in the background (Gong *et al.*, 2018)

Real time transport demand management (Bricka *et al.*, 2015)

Most “﻿human mobility behaviors follow a simple and reproducible pattern [in GPS].” Lin & Hsu (2014) -> go on about predictability and hence mode classification and hence opportunity

Can see people can experience space-time events (i.e. Traffic Jam, Crowds, THINK OF MORE etc.). Both fast and slow dynamics/flows have an impact on changes in a variable across space, accounting for these is essential in modelling [i.e. differencing to remove long-term trend] (Batty, 2013).

Especially as the quality and breadth of travel surveys is not high (Kim *et al.*, 2015)

[VGI has huge potential for trip purpose classification as] Data for trip purpose has typically been collected using travel surveys which tend to emphasise behavioural patterns of people from certain socio-demographics and fail to be representative of the population as a whole (Kim *et al.*, 2015).

VGI, has also allowed study at larger geographic scales

(close to population level) with purpose and sentiment studies (twitter and geo-reffered social media) (Rashidi *et al*., 2017)

* Twitter for geographic understanding of purpose of movement -> “﻿extracting information from social media to track and analyse human movements” (Rashidi *et al*., 2017)

Broader classification eith larger data types of patterns at larger scales (Xie *et al.*, 2016)

MTL Trajet (Patterson & Fitzsimmons, 2017). This has fuelled a shift from tradition methods (i.e. travel surveys, phone surveys) to using more VGI sources (smartphone apps)

Attard *et al.* (2016) advocate the use of VGI in to study transport.

To understand the feasibility of classification of trip purpose we need to understand the current state of modelling of space and time

Ultimately, the value of SBD (Spatial big data analysis) relies on uncertainty handling (Shi *et al.*, 2018)

Data for trip purpose has typically been collected using travel surveys which tend to emphasise behavioural patterns of people from certain socio-demographics and fail to be representative of the population as a whole (Kim *et al.*, 2015).Yazdizadeh *et al* (2019) find that models classifying transport mode techniques tend to be more applicable to at a wider range of spatial and temporal scales and are generally more accurate [predictability, less to identify].

An *et al.* (2015). Problem is that one is often taken in expense of the other with spatial and temporal data An *et al.* (2015).

[On visualising big data maps] New computational and technical paradigms for cartography are accompanying the rise of geospatial big data (Robinson *et al.*, 2017).

Big geospatial data visualisations do not always scale well – because they can become messy (Li *et al.*, 2016)

### 2.2.2 Issues with using VGI

There are a number of drawback of studies using VGI,

problems of representativeness in VGI (Li *et al.*, 2016). Indeed, ﻿“users must opt in to share information on their activities” (Elwood *et al.*, 2012)

credibiliity (Flanagin & Metzger, 2008)

VGI can be biased towards cities (Hecht & Stephens, 2014)

Lack of quality control (Goodchild, 2013)

“Virtually Impossible to create a representive sample in geographic space (Goodchild, 2013)

common VGI, ﻿Twitter, Flickr, and Foursquare (Hecht & Stephens, 2014)

(Attard et al., 2016 referencing Lyons & Harman, 2002). → 1. There are issues with trust over the information provided, 2. Lifestyle changes are opportunities for travel behaviour change, 3a. People have very poor judgement of cost and time when travelling by car with control over their journeys being seen as important. 3b. Public transport in contrast, is seen as difficult as information is sought from unfamiliar and uncertain sources.

“People ﻿in rural areas tend to use technology differently than people who live in cities.” (Hetch & Stephens, 2014). “﻿Foursquare check-ins are not public by default, but can be shared widely if a user connects her/his account to Twitter” (Hetch & Stephens, 2014)

As such, many studies have become useful to specific regions at specific time-period that data were collected (ref).

Miller & Goodchild (2014) becomes very problematic when we make generalisations about populations from inferred data (such as twitter)

Despite this, Li *et al.* (2016) find that use of VGI often “on a fairly ad-hoc basis when compared with traditional data sources, and usually must be repurposed to fulfil research objectives” (Li *et al.*, 2016)

VGI requires rethinking of geographical concepts (Elwood *et al.*, 2012). [With big geospatial data] validity or trust is traded for the velocity of information production (Li *et al.*, 2016)

Difficult to falsify VGI (Elwood *et al*. 2012). Shi *et al.* (2018) greater authenticity issues with VGI than other data sources

Many studies have shown that voluntary contributions by individuals follow a frequency distribution with a long tail, with a few individuals making large numbers of contributions (Goodchild & Li, 2012)

Data collection [for survey data] driven by data availability or convenience of data collection rather than by domain knowledge, theory, or insight into the process(es) of interest. (An *et al.*, 2015).

methodologies originally developed to analyse small data and are not necessarily equipped to tackle some of the distinctive features of newer sources (Gorman, 2013; Arribas-Bel, 2017).

Many previous studies have proposed and validated in limited geographical areas, such as campuses (Jahromi *et al.*, 2016)

## 2.3 Study of mobility within Montreal

Montreal has a relatively high share of transit ridership (for a North American city) Also multimodal public transport network (Eluru *et al.* 2012)

Eluru *et al.* (2012) survey to look at public transport and how people move around Montreal. Trying to encourage it

The city of Montreal itself is the largest city within Quebec and the second largest within Canada ()

See (Zahabi *et al.*, 2017) for datamobile analysis

*Montreal:*

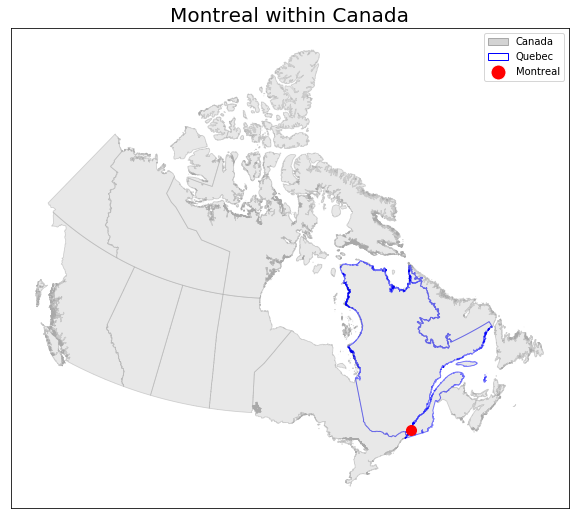
-Chevalier *et al.* (2018) the island of Montreal (Île de Montréal) largest of the Hochelaga Archipelago near the confluence of two rivers.

-WPR (2019) -> 1.75 million people as of 2016. The island in total has an estimated population of 1.95 million. “The city proper has a population density of 4,517 people per square kilometer (11,701 residents per square mile)”. “By 2030, the Greater Montreal Area is expected to grow to 5.275 million”

-The city of Montreal, located within the Greater Montreal region is the largest city in Quebec with a population size of … . It is of particular interest to city transport research due to its unique road and public transport networks. There are a total of

“Montreal with its unique multimodal public transportation system consisting of bus, metro and commuter train offers multiple transit route alternatives to individuals commuting to downtown" Eluru *et al.* (2012)

an region containing around 4 million people (WPR, 2018; **Figure 3.1**).



**Figure 3.1** Location of Montreal within Quebec, Canada

### 2.3.1 MTL Trajet Project

Introduction and purpose

The MTL Trajet project is a large scale mobile phone travel survey that has been run yearly around Oct-Nov since 2016 (Ville de Montreal, 2019). The project relies on participatory volunteered geographic information from its app. User friendly interface

Developed onto of the success of the DataMobile Smartphone Travel Survey example from 2014 carried out by Concordia University in Montreal (Patterson & Fitzsimmons, 2016) 10 November - 5 December 2014. Close to 900 people participated in the survey [Only around the univirsity]. Based in Montreal

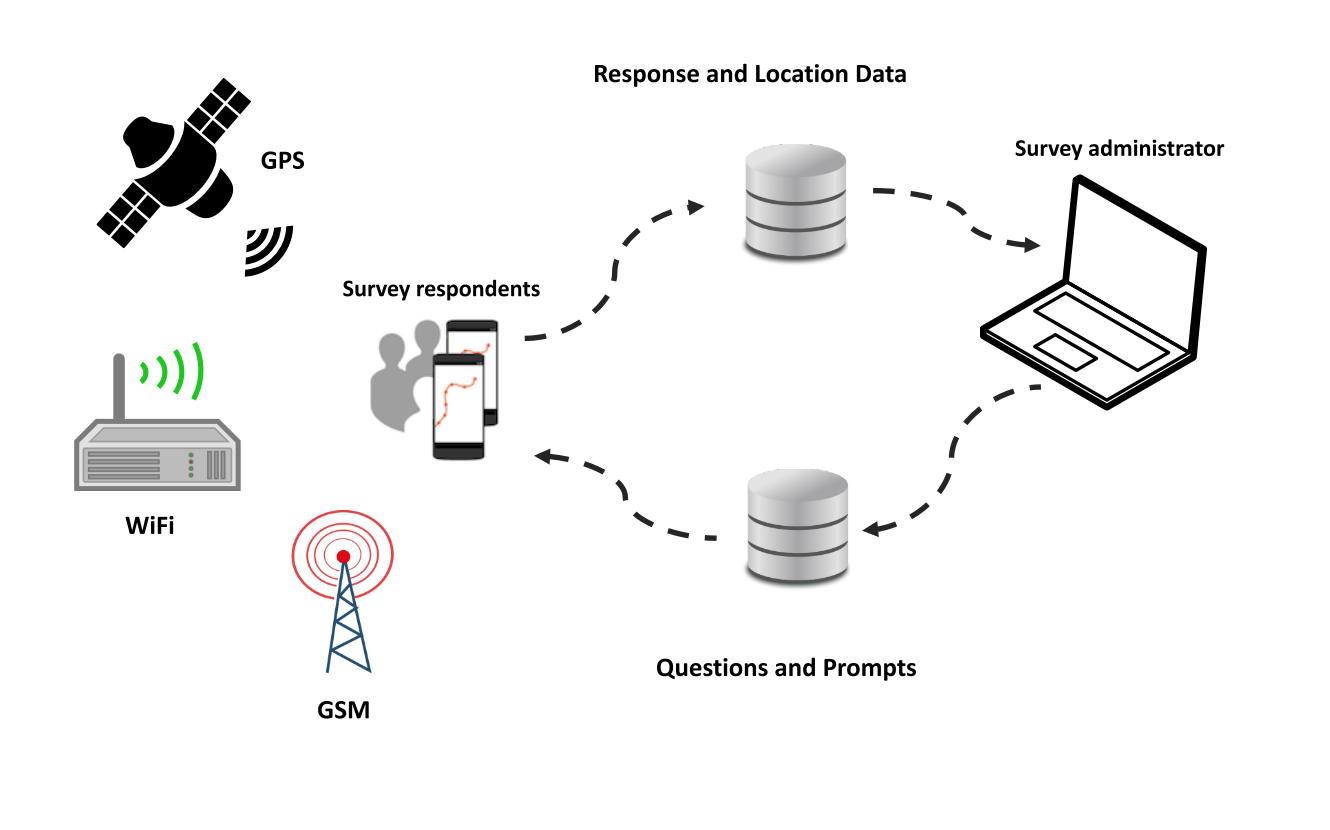
18 and over

Aim of the MTL Trajet is to better understand travel (MTL Trajet, 2017)

The MTL Trajet is a part of the Itinerum platform which is an app providing researchers a platform to develop their own spatial surveys (Yazdizadeh *et al.*, 2019) Now Itinerum Platform **Figure 2.2.**

See (Zahabi *et al.*, 2017) for datamobile analysis

MTL originally had personal locations (See Table 2.1), although have been removed for the data used in this report as this is available from Portail de Ouverte Donnes



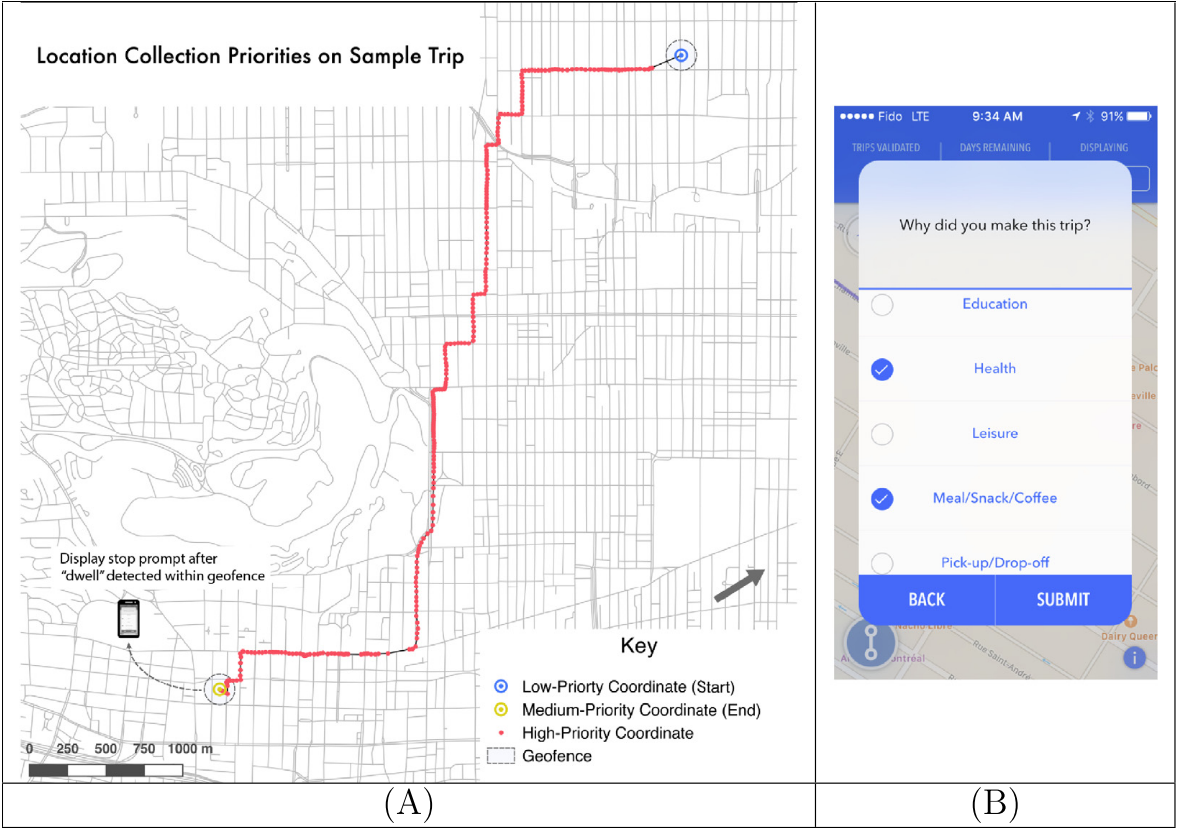
**Figure 2.2** Framework for survey apps developed using the Itinerum platform (Source: Patterson & Fitzsimmons, 2017a).

Bit about the app:

The app employs geofencing

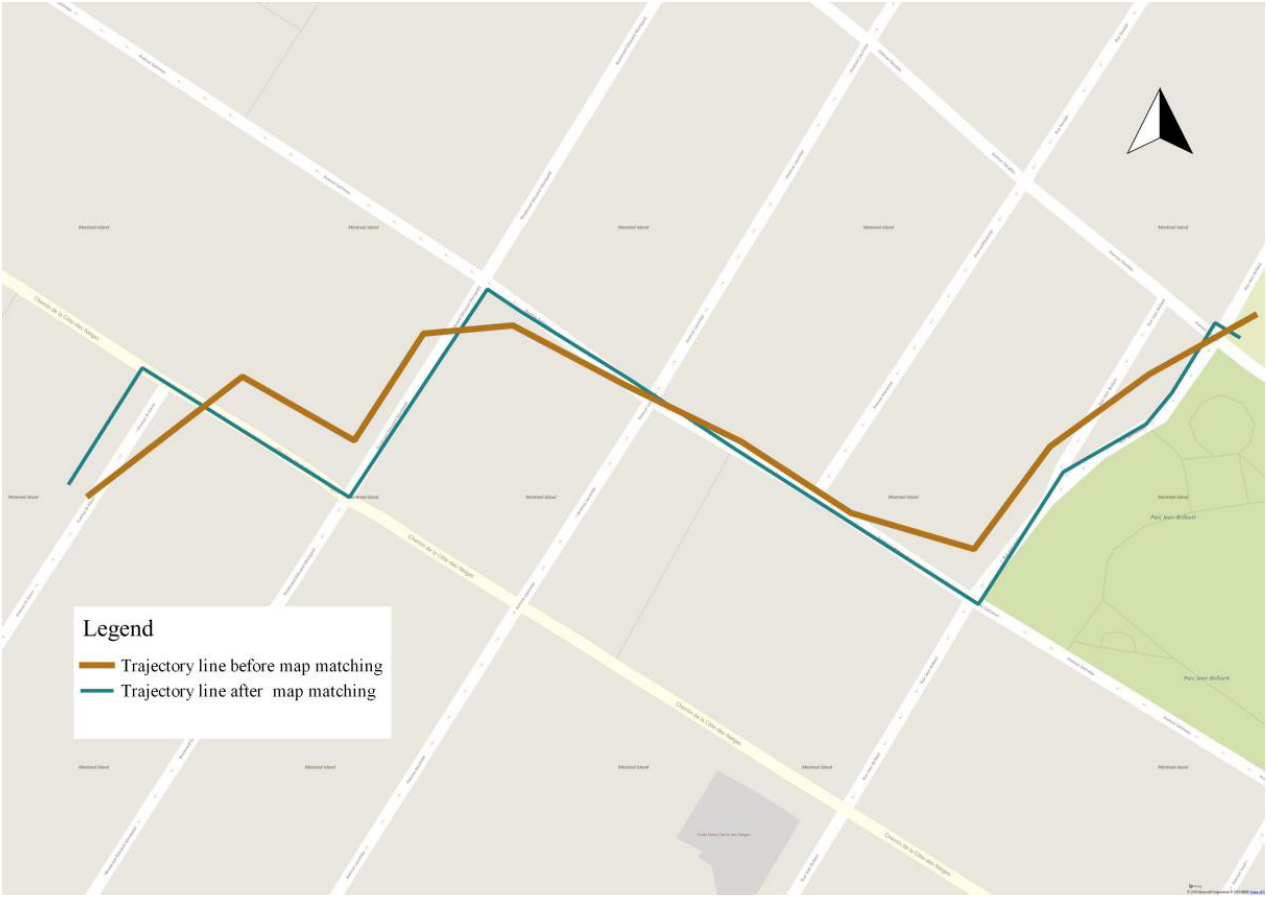
[About the app]:

“﻿Central to the application is a long-running location logging service that requests and monitors changes to location in a battery-efficient way” (Patterson *et al.*, 2019). ‘﻿If the device leaves the geofence at any time, it switches to high-priority mode to request a new point and establish the new ‘‘rolling’’ geofence (see Fig. 4).” (**Figure 2.3**; Patterson *et al.*, 2019). Map-matching (**Figure 2.4**)



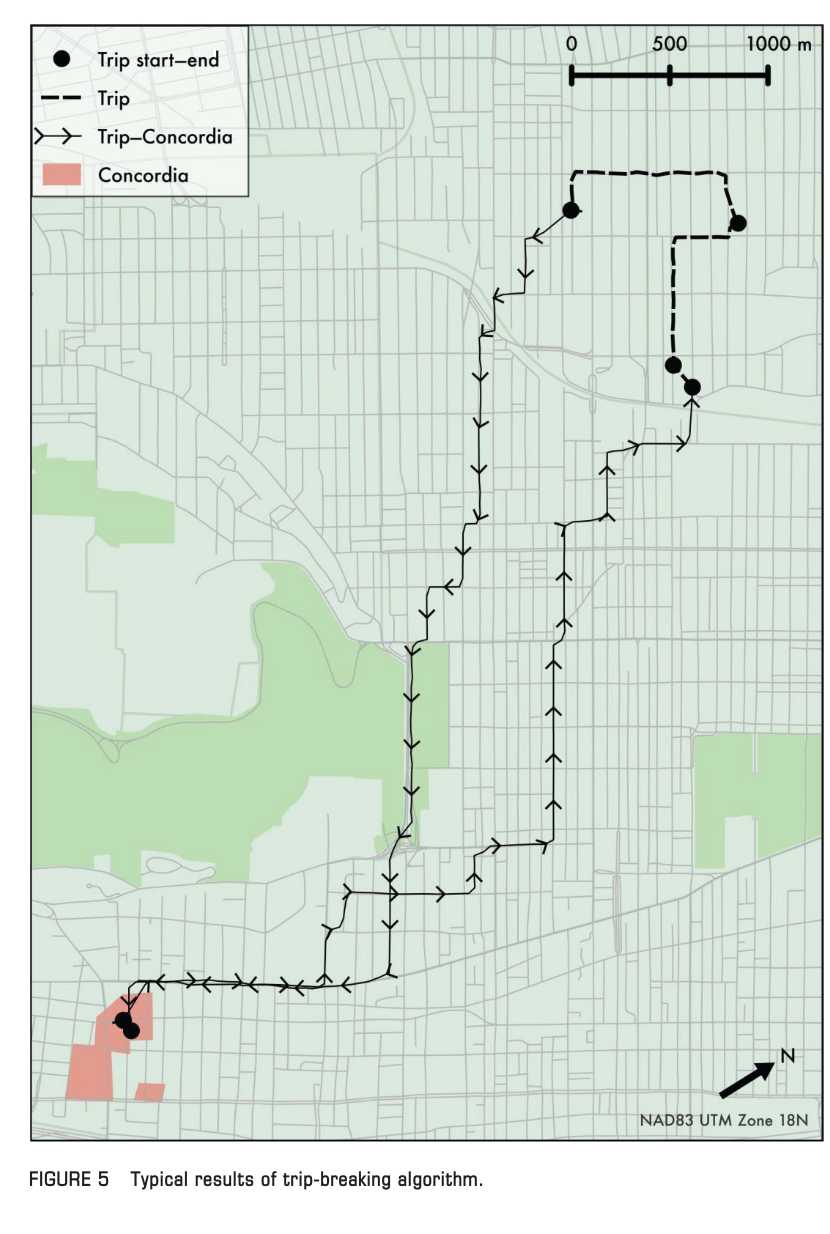
**Figure 2.3** Example of GPS trace with location collection priorities within an Itinerum platform app (A); Example of the on screen prompt after an Itinerum platform app stops recording movement (B) (Source: Patterson *et al.*, 2019).

Each Route fed into the OSRM, some overall and unaccountable inaccuracy (Patterson & Fitzsimmons, 2017b; Ville de Montréal, 2019), unknown unknowns (Shi et al.?) But this is present in all VGI (Elwood *et al*. 2012)



**Figure 2.4** Example of ‘Map Matching’done by the Open Source Routing Machine when processing the raw GPS trace from user devices (Source: Hamouni, 2018).

“﻿Also, the University of Toronto Transportation Research Institute will include a version of Itinerum on iOS as part of their evaluation on the future of travel surveys for their TTS2.0 projec”Patterson & Fitzsimmons (2017a)



**Figure 2.5** Example of trip-breaking algorithm (adapted from Patterson & Fitzsimmons, 2017a)

The MTL trajet has seen limited use in the literature and has instead mainly been restricted to . Yazdizadeh *et al.* (2019) use the 2016 edition of MTL Trajet survey and carry out mode and purpose classification 71%.

The majorit of use has been behind the scences

Still potential for study.

Fallah-Shorshani (2018) uses MTL to look at pollution exposure

Hamouni (2018) Pedestrian Route Choice Model from MTL Trajet 2016.

MTL Trajet is a supplement to the O-D survey in Montreal – also will get younger people involved (MTL Trajet, 2019)

MTL Trajet (2017) results were used to discover how many people went on trips and couldn’t find a parking space (moving round and round in circles downtown). Also, for the redevelopment and planning of the needs of the road network – the study is being used so that the redevelopment meets the needs of people (drivers, cyclists, etc.). Also, in alternative route analysis, where there are road works. Finally, the study helped with identifying common bike routes. Also, survey helped with identifying safety on roads .

The 2018 edition of the MTL Trajet study will take place from September 24 to October 28, 2018 (MTL Trajet, 2019).

References:

Aubrecht, Christoph; Ungar, Joachim; Freire, S. (2011). Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population: a case study for Lisbon Metro using foursquare check-in data. *International Conference Virtual City and Territory (7è: 2011: Lisboa)*, (October), 57–60.

Burini F., Ciriello D. E., Ghisalberti A., Psaila G. (2018) The Urban Nexus Project: When Urban Mobility Analysis, VGI and Data Science Meet Together. *In:* Bordogna G., Carrara P. (*eds*) Mobile Information Systems Leveraging Volunteered Geographic Information for Earth Observation. Earth Systems Data and Models, vol 4. Springer, Cham.

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3–4), 137–148.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Hecht, B., & Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 197–205.