# Chapter 2. Literature Review

## 2.1. Trip purpose classification

Introduction to

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

This creates an issue, as data 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). Oppositely, 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].

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.

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 (hereafter RF; 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). A selection of key classification models from the literature are detailed along with their inputs and accuracy in **Table 2.1**.

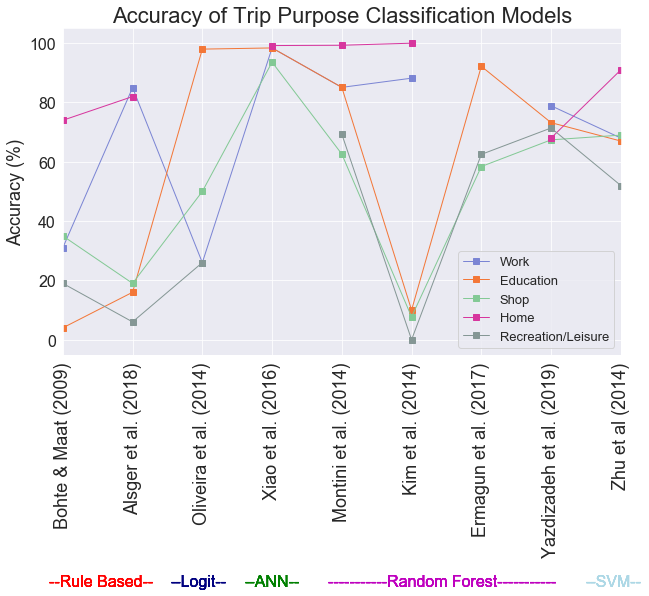
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Author(s)* | *Predictor variables* | *Location and year* | *Number of Trips included in Study* | *Overall classification accuracy* |
| *Rule-Based Methods* | | | | |
| Bohte & Maat (2009) | POI; Personal Locations Proximity | Netherlands, 2007 | ﻿ 33,686 | 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 and inferred 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.

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.

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)

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)

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.

## 2.2 Volunteered Geographic Information in Mobility Studies

About VGI in the literature

The general use of VGI in studies has been

As such, many studies have become useful to specific regions at specific time-period that data were collected (ref). 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 (Kim *et al.*, 2015).

‘People as sensors’. Li *et al.* (2016) distinguish between two types of VGI participatory (conscious inclusion of their data i.e. like MTL Trajet) and opportunistic (unconscious) forms of VGI. Each have their own

Crowd-sourcing, mobile phones cost effective (Shi *et al.*, 2018)

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

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).

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

Use

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

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

On a larger scale (population level) purpose and sentiment:

* Using twitter and sentiment analysis (similar to purpose)
* Twitter for geographic understanding of purpose of movement -> “﻿extracting information from social media to track and analyse human movements” (Rashidi *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

[On big data] However, they are often collected 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)

“there is an opportunity for smartphones to replace dedicated GPS devices” (Wu *et al.*, 2016)

Issues

problems of representativeness in VGI (Li *et al.*, 2016; ref; ref)

The quality of travel surveys is not high (Kim *et al.*, 2015)

(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.

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).

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

Other:

* Using twitter and sentiment analysis (similar to purpose)
* Twitter for geographic understanding of purpose of movement -> “﻿extracting information from social media to track and analyse human movements” (Rashidi *et al*., 2017)
* people travel in distinct patterns broadly based on socio-economic groups, also other classes i.e. returners and explorers (different levels of variance of travel) (Xie *et al.*, 2016)

### 2.2.1 MTL Trajet

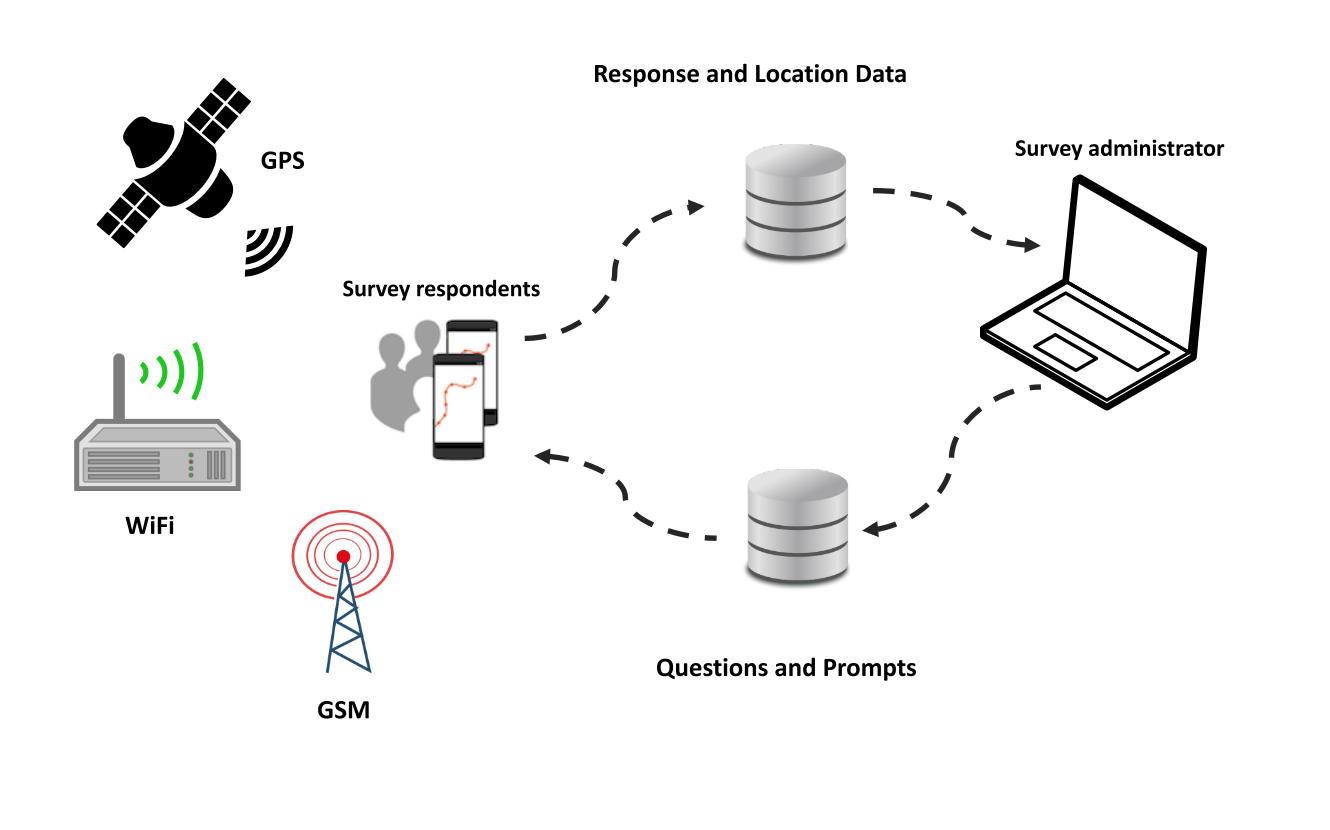
About MTL Trajet and history

The MTL Trajet is a project that has been run yearly around Oct-Nov since 2016, is participatory form of VGI

“Previous studies have mainly concerned themselves on one aspect of trips (e.g. mode) at a time” [but new technology (and the Itinerum platform) has given researchers new opportunity to study purpose] (Yazdizadeh *et al.*, 2019)

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

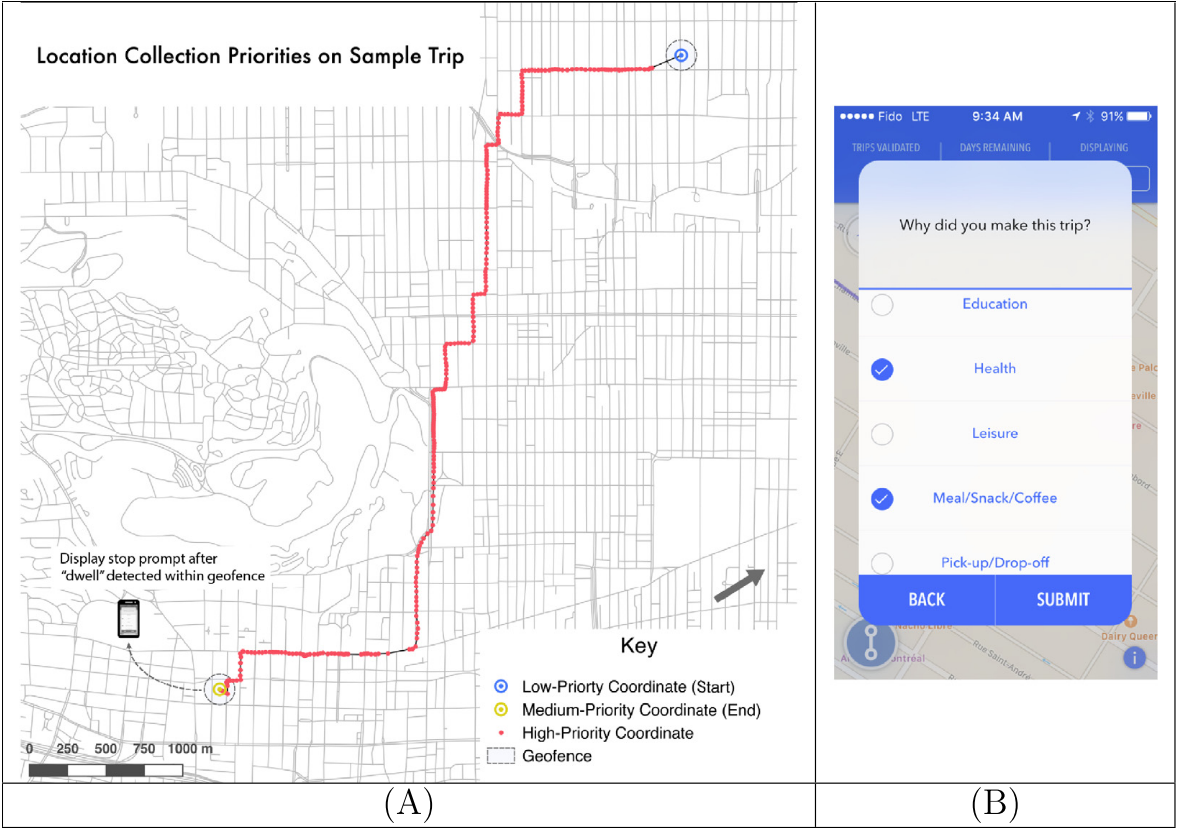
[MTL built on top of the] DataMobile Smartphone Travel Survey example from 2014 carried out by Concordia University in Montreal (Patterson & Fitzsimmons) 10 November - 5 December 2014. Close to 900 people participated in the survey [Only around the univirsity] Many previous studies have proposed and validated in limited geographical areas, such as campuses (Jahromi *et al.*, 2016). Now Itinerum Platform **Figure 2.2**



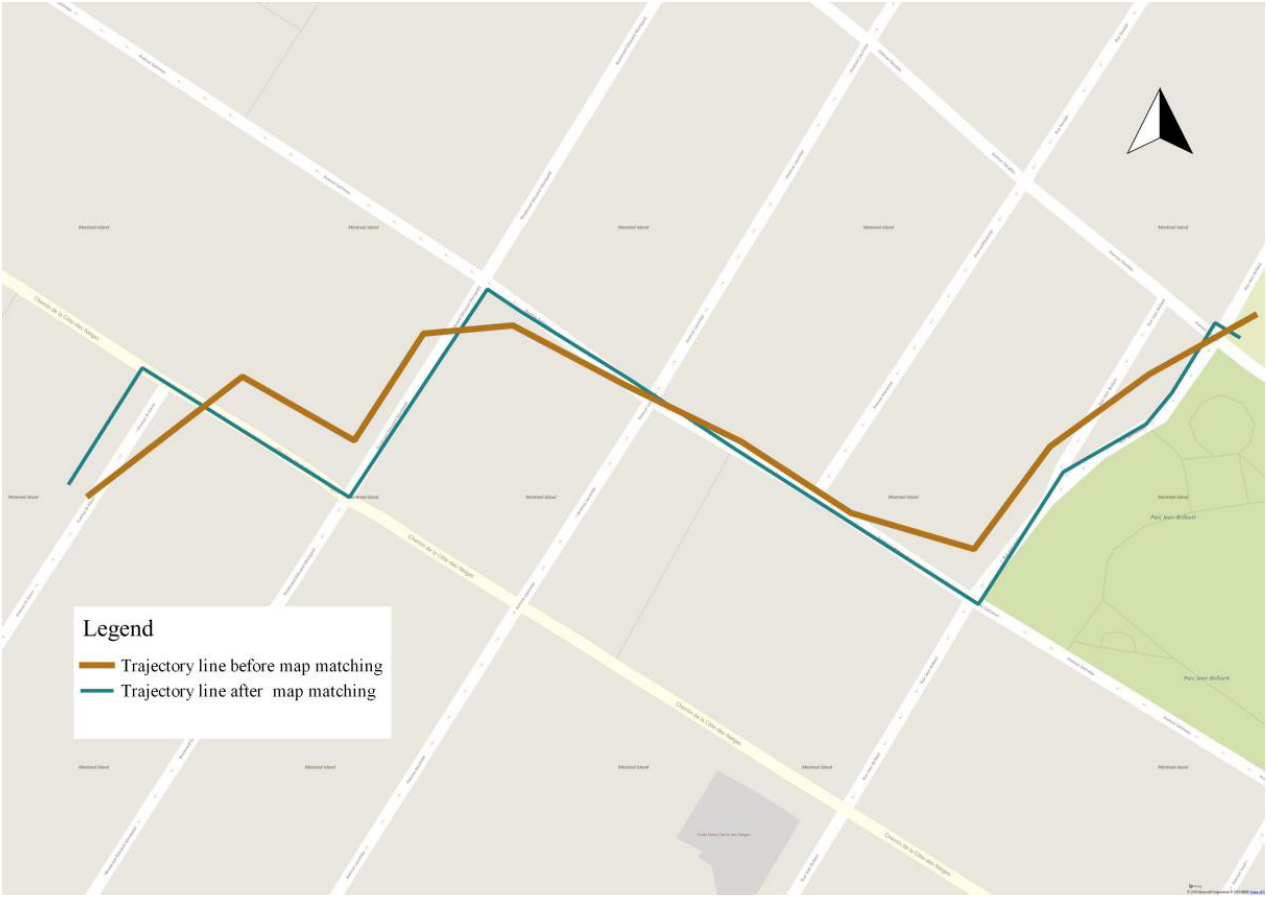
**Figure 2.2** Framework for survey apps developed using the Itinerum platform (Source: Patterson & Fitzsimmons, 2017).

[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).



**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).

Use in literature

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

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

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).

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

“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)

## 2.3. Representation of space and time in modelling

To understand the feasibility of classification of trip purpose we need to understand the current

Difficulty in space, time and space-time metrics in models

1. Models can’t handle

In general, representing space and time and machine learning models has been a difficult notion.

The broader use of space and time metrics in models has been a challenging concept.

machine learning is the unable to directly handle spatio-temporal structure (Cheng *et al.*, 2011)

“Space–time analysis seeks to understand when and where (and sometimes why) things occur.” 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).

Spatio-temporal modeling has always been of interest to researchers in geographical information science (GIS) (Ren *et al.*, 2019).

STARIMA and ANN (Cheng & Wang, 2011)

there is not a close coupling between big data and space-time methods used to analyse them (An *et al.*, 2015)

Significant class-imbalance exists in the MTL Trajet data

Mathematical models being employed without regard of space, often including problems that are inherently tied to spatial considerations (O'Sullivan & Manson, 2015)

Space-Time CNN -> CNN-based models employ a grid map to represent spatial data which is unsuitable for road- network-based data. (Ren *et al.*, 2019)\*

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). However, methods to assess quality assurance of the geospatial data still relevant on big geospatial data as they still describe the same processes (Li *et al.*, 2016)

ML methods are generally effective in tackling nonlinearity in spatial data (Li *et al.,* 2016)

1. Computationally expensive

ST-KDE and space-time decomposition to compute space-time methodologies (Hohl *et al.*, 2016).

Network topology and space-time weight matrix

CNN was not optimal for modelling the spatial patterns of the road because the authors did not consider the road network topology (i.e. direction) (Ren *et al.*, 2019 in their literature review)

network and topological models not well suited for handling geospatial big data (Li *et al.*, 2016)

[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)

1. Things Change

Size of space-time neighbourhoods change (Cheng *et al.*, 2014)

[Problem with time sampling] -> ﻿different patterns can be observed for different temporal resolutions (Zhao *et al.*, 2019).

Determining threshold of time and space (Adepeju & Evans, 2018)

MTUP (Cheng & Adepeju, 2014) and MAUP (Openshaw, 1983)

(Ren *et al.*, 2019)\* Need to account for local trends with models, a lot of papers use models that don’t consider local, but still do space-time analytics (i.e. with locally-weighted CNN layers on a network)

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

Within structures such as cities things change and develop over time in either a state of equilibrium or dis-equilibrium (Batty, 2013)