Can we predict why people travel within a city?: A study analysing the spatial and temporal characteristics of travel intention within Montréal, Canada between September and October 2017.

*This dissertation is submitted in part requirement for the MSc in the Centre for Advanced Spatial Analysis, Bartlett Faculty of the Build Environment, UCL.*

*A colour version of this dissertation is available online.*

Word Count:11,986

Thomas J. Keel – 18110348

CASA0004: MSc Spatial Data Science & Visualisation University College LondonSupervisor: Huanfa Chen

*30th August 2019*

# Abstract

The prediction of why people travel across cities remains an area within the broader mobility studies without extensive investigation. Arguably, this has been hindered by:

(1) an absence of large datasets which details the activities that people’s travel to when they move across a city;

(2) the difficulty in accurately representing space and time within models used to predict why people travel across cities.

Regarding (1), in recent years, smartphones travel surveys have provided researchers a platform to study the attributes characterising travel within a city at increasingly fine temporal and spatial scales. This study makes uses of one such study: the *2017 MTL Trajet* app – a travel survey examining *how* and *why* its participants have moved across Montreal between 18th September 2017 and 18th October 2017. Regarding (2), this project builds upon a small body of research to uncover and categorise spatial and temporal interdependencies of the GPS data from the MTL Trajet project before assessing the performance of three machine-learning classification models: Random Forests, Support Vector Machines and Artificial Neural Networks. Note, these models are built to classify *why* people travel based on spatial and temporal characteristics of individual trip.

**Key Words:** Travel intention classification, Mobility, Spatio-Temporal Investigation, Volunteered Geographic Information.

# 

# Declaration

I, Thomas Keel, hereby declare that this dissertation is all my own original work and that all sources have been acknowledged. It is 12,000 words in length

Signed:

Date: 28th August 2019

Table of Contents

[Abstract 1](#_Toc17552552)

[Declaration 2](#_Toc17552553)

[List of Figures 3](#_Toc17552554)

[List of Tables 4](#_Toc17552555)

[List of Acronyms and Abbreviations 4](#_Toc17552556)

[Acknowledgments 5](#_Toc17552557)

[Chapter 1. Introduction 6](#_Toc17552558)

[1.1 Research Overview and Questions 6](#_Toc17552559)

[1.2 Motivation 7](#_Toc17552560)

[1.3 Approach 8](#_Toc17552561)

[1.4 Outline 10](#_Toc17552562)

# List of Figures

**Figure 1.1** Screenshot from the MTL Trajet App (left; source: Patterson, 2017) and example of prompt within Itinerum apps (right; source: Patterson *et al.*, 2019)………………………..……….1

**Figure 2.1** Comparison of Trip Purpose classification model accuracy within the literature (ANN=Artificial Neural Network; SVM=Support Vector Machine)…………………………………………2

**Figure 2.2** Framework of the Itinerum Mobile and Desktop app (Source: Patterson & Fitzsimmons, 2017)…………………………………………………………………………………………………...…………2

**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)……………………………………………………..4

**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)…………………………..4

**Figure 3.1** Montreal within Quebec, Canada………………………...………………………..……………………5

**Figure** **3.2** GPS routes from the MTL Trajet plotted within the study area…………...………...……5

**Figure 3.3** Location of Montreal within the study area. ……………………………………………………….5

[**Figure 3.4** Bar chart showing the count of each category of land use]

**Figure 3.5** Map showing land use changes within the City of Montreal………..…………...…………5

**Figure 3.6.** Circular contour plot (windrose; left) and circular histogram (right) showing the direction of trips (circle bands indicate count of trips)

**Figure 3.7** Example of the spatial join between a route and the underlying dissemination areas (route in **blue**; overlapping dissemination areas in **red**)

**Figure 4.1** Boxplots (top), Kernel Density Estimation (middle) and Quantile-Quantile (bottom) plots showing the distribution of distance and duration of trips from the 2017 MTL Trajet travel survey.

**Figure 4.2** Bar charts showing the type of trip purpose and travel mode selected by respondents to the 2017 MTL Trajet survey.

**Figure 4.3** Bar chart comparing the proportion of each unique trip purposes accounted for by each unique travel modes.

**Figure 4.4** Map showing the mean direction of trip within each region of Greater Montreal.

**Figure 4.5** Circular contour plot showing the mean direction of trips for each trip purpose.

**Figure 4.6** Bar chart showing the proportion of trips carried out during rush-hour and off-peak as grouped by purpose.

**Figure 4.7** Bar chart showing the proportion of trips carried out within and outside the City of Montreal as grouped by purpose.

**Figure 4.8** Bar chart showing number of trips that have their origins or destinations in each land use category (as defined by Ville de Montreal, 2014).

**Figure 4.9** Bar charts comparing the proportion of each unique trip purposes accounted for by each unique land use category (as defined by Ville de Montreal, 2014) in the trip origins and destinations.

**Figure 4.10** Line graph comparing sum of squared distances and silhouette scores of k-means clustering algorithm for k between 2-20.

**Figure 4.11** Map of origin and destination points from the MTL Trajet trips coloured by cluster label across the study region.

**Figure 4.12** Bar chart showing number of trips per spatial cluster identified by the k-mean clustering algorithm.

**Figure 4.13** Line graph comparing coherence score and log perplexity of LDA models using a topic count of between 1-12.

**Figure 4.14** Calendar plot showing the weighted importance of each ‘temporal word’ in each of the temporal clusters (1-5).

**Figure 4.15** Count of trips associated with each temporal cluster identified for this analysis.

**Figure 4.16** Line plot showing the amount of recorded trips taken from the MTL Trajet app between 18th September 2019– 18th October 2019 (weekends indicated in **purple**).

**Figure 4.17** Average trip distance and duration represented as a percentage of the mean.

**Figure 4.18** Calendar plot showing the temporal profile for each trip purpose class of the count of trips recorded per hour as average per day of the week.

**Figure 4.19** Time series plot showing the average temperature (in Celsius) and precipitation (mm) recorded during the study period

**Figure 4.20** Temporal de-composition of the count of trips recorded by the MTL Trajet travel survey at 24-hour lags.

**Figure 4.21** Local indicator of spatial association (LISA) maps of local Moran’s I of trip origin and destination points for each trip purpose class.

# List of Tables

**Table 3.1** Description of the key variables from data from the MTL Trajet survey before pre-processing.

**Table 3.2** Categories of mode and purpose allowed for trips in the MTL Trajet survey.

**Table 3.3** Description and cover of Land Use categories within the City of Montreal.

**Table 3.4** Definition of Rush hourand Off-peak hours used in this study.

**Table 3.5** Description of the key variables from the MTL Trajet survey after pre-processing.

**Table 4.1** Outline of trips removed from the analysis

**Table 4.2** Summary statistics of overall distance and duration of trips from the 2017 MTL Trajet travel survey

**Table 4.3** Summary statistics of trip distance and duration per trip purpose

**Table 4.4** Results from the application of Rush-hour and City labels to the data.

**Table 4.5** Outline of trip purposes associated with each temporal cluster found by a 5-topic LDA model.

**Table 4.6** Augmented Dickey-Fuller Test (significant below 0.005 shown in **bold**)

**Table 4.7** Global Moran’s I tests by trip purpose.

# List of Acronyms and Abbreviations

**ANN** – Artificial Neural Networks

**DA** – Dissemination Areas

**GPS** – Global Positioning System

**LDA** – Latent Dirichlet Allocation

**MAUP** – Modifiable Areal Unit Problem

**MLP** – Multi-Layer Perceptron

**MTUP** – Modifiable Temporal Unit Problem

**RF** – Random Forest

**SVM** – Support Vector Machine

**VGI** – Volunteered Geographic Information

# Acknowledgments

I would firstly like to thank my family and friends for their thorough support during the planning and writing of this dissertation. A special thanks goes out to my flat mates James, James and George for putting up with me this year.

Secondly, I would like to thank my supervisor Huanfa Chen for his useful comments and advice during the development of this project.

Finally, I am extremely grateful to all the people I have met at CASA and I thank them for their enthusiasm and all the support they have given me throughout this year.

# Chapter 1. Introduction

## 1.1 Research Overview and Questions

The purposes by which people use transport networks on a large scale remains an area with a distinct lack of investigation within the broader mobility studies (Yazdizadeh *et al.,* 2019). In the past, this has primarily been due to an absence of data, at such a large scale, available to study the activities that people are travelling to, when they move around a city.

In recent years, however, improvements to GPS and processing power of smartphones has provided researchers a new opportunity to study and record the large scale geospatial movement of people (Zhao *et al.*, 2019). Travel survey apps created for smartphones require much less effort from their participants than traditional travel surveys (i.e. where a separate GPS device is required to record movement) (Li et al., 2016). Therefore, it has become increasingly easy to collect qualitative information about movement within a city – including information about *how* and *why* people travel.

The ability of smartphone users to create a large amount of geographically-referenced in these travel survey apps can help us generate unique insight into transport behaviour at much greater scales than ever before. This form of participatory data creation is known as Volunteered Geographic Information (hereafter, VGI) (after Goodchild, 2007).

Despite the potential to produce more VGI that can be used to generate insight into mobility within a city, there are many cities globally that have no form of formal research initiated within them (Attard *et al.*, 2016). One exception to this, is Montreal, Canada, where a number of mobile travel survey applications have been created to study *how* and *why* participants have moved within the city. This report makes use of the most recent available dataset from one of these studies: The *2017 MTL Trajet* travel survey project (Ville de Montréal, 2019). The *MTL Trajet* project was carried out between 18th September 2017 and 18th October 2017 and is used in this dissertation to following assess the following research questions:

**Main Research Question:**

Can we effectively classify the purpose of trips using spatial and temporal indicators?

**Sub-Questions:**

1. Which spatial and temporal indicators are most important for the classification of trip purpose?
2. Which type of classification model is most effective in the classification of trip purpose?

## 1.2 Motivation

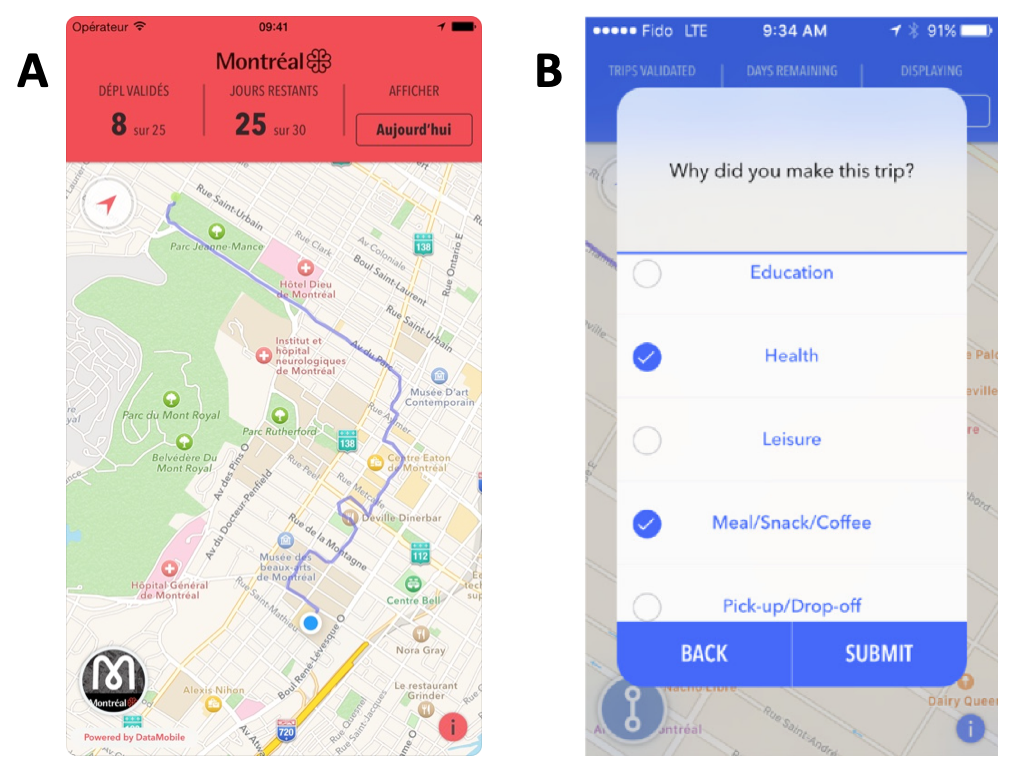
Movement can be thought of as an interaction between an origin and destination (Murray *et al.*, 2012). People move across space and through time to go from where they are to where they want to be (Murray *et al.*, 2012). Transport, is the by-product of the interaction between an origin and destination, and can thus is best considered a ‘derived demand’ for a given destination (Golledge & Gärling, 2001). Studying the patterns in the types of destinations that people demand to travel to, thus, underpins our comprehension of behavioural patterns within a city (Kwan & Neutens, 2012).

If we are able to discern the activities for which individual’s make movements (hereafter, their ‘*trip purpose’*), we may be able to use this information to inform policy and account for demand in essential (e.g. health & educational services) and non-essential (e.g. leisure & commercial) services throughout a city (Attard *et al.*, 2016).

To better understanding and classify trip purpose (or intention of movement), we first need to understand the temporal and spatial scales at which people are travelling at for certain activities. The motivation of this study is thus to evaluate whether we can use spatial and temporal dependencies discovered within different categories of trip purposes to model and classify them.

## 1.3 Approach

This study makes use of data from the *2017 MTL Trajet* survey originally collected by researchers at the Transportation Research for Integrated Planning (TRIP) lab, Concordia University (Patterson & Fitzsimmons, 2017a). This survey was part of the 2015-2017 Montréal Smart and Digital City Action Plan and was created to study travel behaviour across the city (MTL Trajet, 2017). Data collection for this survey was carried out through a mobile app (available on both iOS and Android platforms) which automatically recorded a location trace using GPS provided from a user’s phone (**Figure 1.1A**; Patterson & Fitzsimmons, 2017a). When users were stopped in a given location for more than intervals of 120 seconds the app would prompt the user to end the trip and input a travel mode and travel purpose for the trip (see **Figure 1.1B**).

****

***Figure 1.1*** *(A) Screenshot from the MTL Trajet app showing recorded GPS trace (source: Patterson, 2017a). (B) Example of prompt similar to one used in the MTL Trajet app (source: Patterson et al., 2019).*

Data from the *MTL Trajet* forms the backbone of the information used to test three types of classification models that look to characterise the purpose of movement:

1. Random Forests

2. Support Vector Machines

3. Artificial Neural Networks

Temporal and spatial clustering techniques will be used to generalise about the space and time trends seen within the data before being input into the models.

## 1.4 Outline

The following chapters of the report are organised as follows:

*Chapter 2* reviews literature relating to trip purpose classification, the use of VGI in mobility studies and existing travel surveys based in Montreal.

*Chapter 3* details the steps carried out in the data pre-processing and collection, the development of space and time metrics from the MTL Trajet data, and the set-up for each trip-purpose classification model.

*Chapter 4*, presents the results from the analysis procedure and compares the performance of the classification models.

*Chapter 5* discusses the extent to which the research objectives (set out in 1.1) have been achieved in the results and highlights uncertainty within them the analysis procedure.

Finally, *Chapter 6,* draws conclusion from the research carried out in this project and suggests areas of further research.

# Chapter 2. Literature Review

## 2.1. Trip purpose classification

Although a wealth of literature exists regarding the classification of transport mode from GPS traces, investigation into the classification of transport purpose has received far less attention (Yazdizadeh *et al.*, 2019). One reason for this imbalance is that that users are required to 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**.

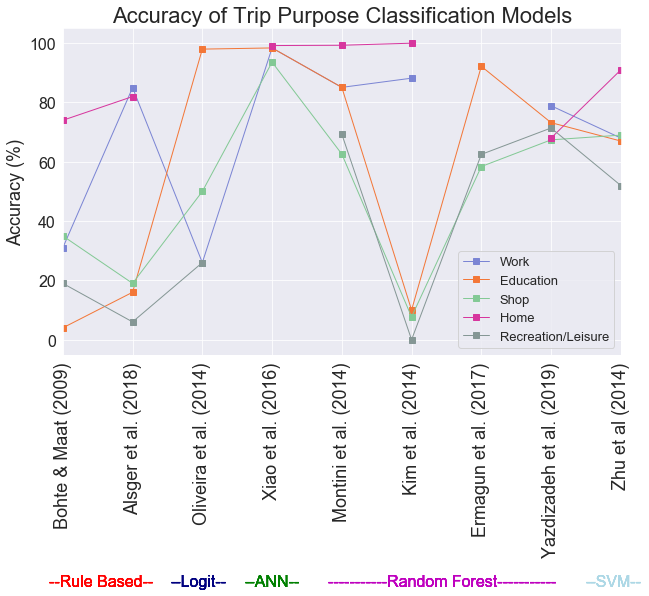
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Author(s)* | *Predictor variables* | *Location and year* | *Number of Trips included in Study* | *Overall classification accuracy* |
| *Rule-Based Methods* | | | | |
| Bohte & Maat (2009) | POI; Personal Locations Proximity | Netherlands, 2007 | ﻿ 33,686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

Employment status and other socio-demographic (Zhang *et al.*, 2019).

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)

With transport mode, we have defined features which give-away which class something is i.e. acceleroation, velocity, space -> less so with purpose classification (Dabiri & Heaslip, 2018)

## 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)

Spatial and temporal information provided from this VGI can be integrated into city-level decision-making to help inform planning a variety of essential and non-essential services (Attard *et al.*, 2016). For example, if we knew that people tended to cycle to cafés during lunch breaks, policy could be implemented to introduce bike racks near the cafés.

Smartphones being able to better record similar mobility behaviour as their carriers (Jahromi *et al.*, 2016).

VGI has gained an increasing amount of attention in the literature and allowed researchers to begin to study

* “Getting a deeper understanding of human mobility is a prerequisite for a broad range of possible studies on smart cities and related research areas”. (Xie *et al.*, 2016)
* Understanding the interactions within complex system such as a city is a prerequisite for predicting changes within it… big data offers us an opportunity to study this (Cheng *et al.*, 2017)
* :”In practice, people’s trip purposes are very important in understanding travel behaviors and estimating travel demands” (Meng *et al.*, 2019)

“﻿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)

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)

Further, improving our understanding of the context surrounding human mobility in a city can even be used in the estimation of travel demand in the longer term (Meng *et al.*, 2019). This is as, the modes of travel that people use around a city are often tied to socio-demographic characteristics of underlying populations such as employment status and affluence (Zhang & Cheng, 2019). Through shifts in these characteristics, such as through underlying process within a city such as gentrification, this has an effect on the travel patterns that people display and the types of activities that they partake in (Bricka *et al.*, 2015).

Big geographic data allows us to not only study the spatial and temporal interactions but also interactions of socio-economic factors [this is what this research aims to do] (Cheng *et al.*, 2017).

“Combining such information [detailed GPS speed, acceleration, etc] with socio-demographic characteristics of travellers has the potential of offering a richer modelling framework that could facilitate better transportation mode detection using variables such as age and disability” [mention it has success in mode transport classification but not purpose] (Bantis & Haworth, 2017)

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

Despite the potential to produce more VGI that can be used to generate insight into mobility within a city, there are many cities globally that have no form of formal research initiated within them (Attard *et al.*, 2016).

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)

"The public release of such precise information, particularly location data such as place of residence, opens the risk of privacy violation" Badu-Marfo *et al.* (2019)

Car sharing in Montreal – some trips always car (Sioui*, 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 ()

Grimsrud & El‐Geneidy (2013) looking at public transport usage in youth

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

*Montreal:*

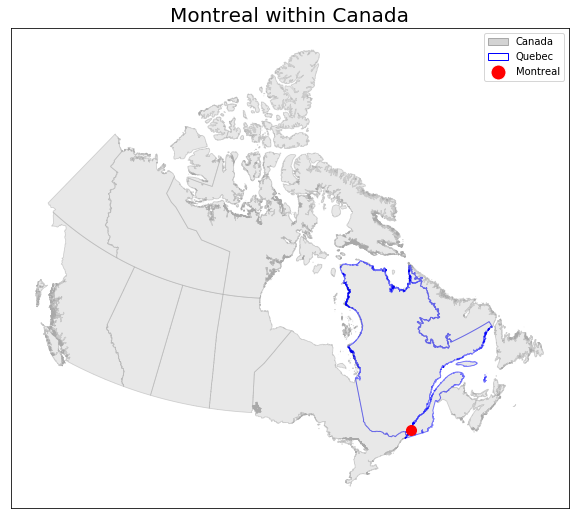
-Chevalier *et al.* (2018) the island of Montreal (Île de Montréal) largest of the Hochelaga Archipelago near the confluence of two rivers.

-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

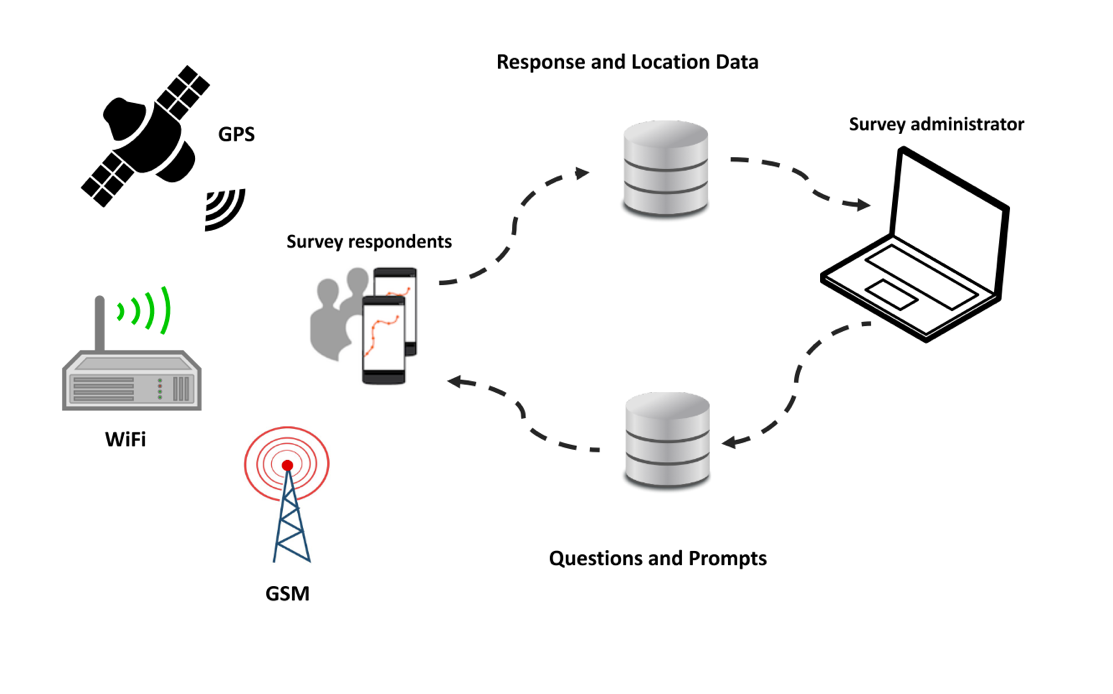
18 and over

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



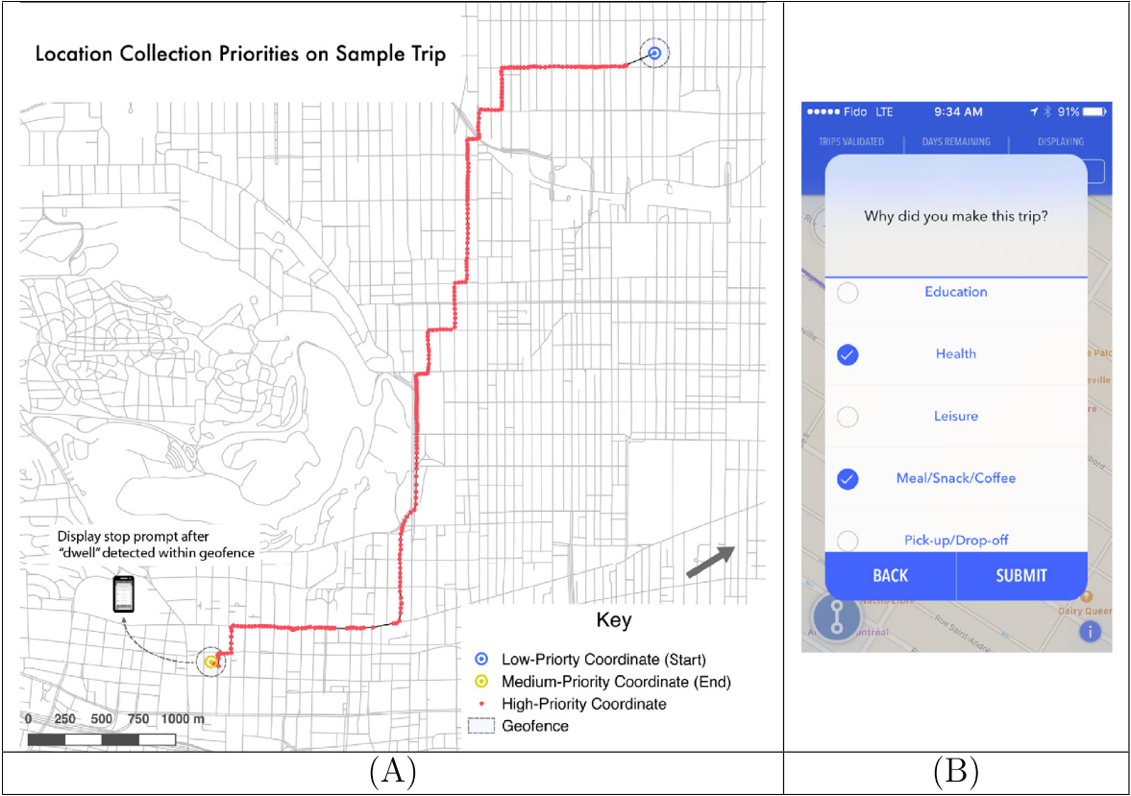
***Figure 2.2*** *Framework for survey apps developed using the Itinerum platform (Source: Patterson & Fitzsimmons, 2017a).*

Bit about the app:

The app employs geofencing

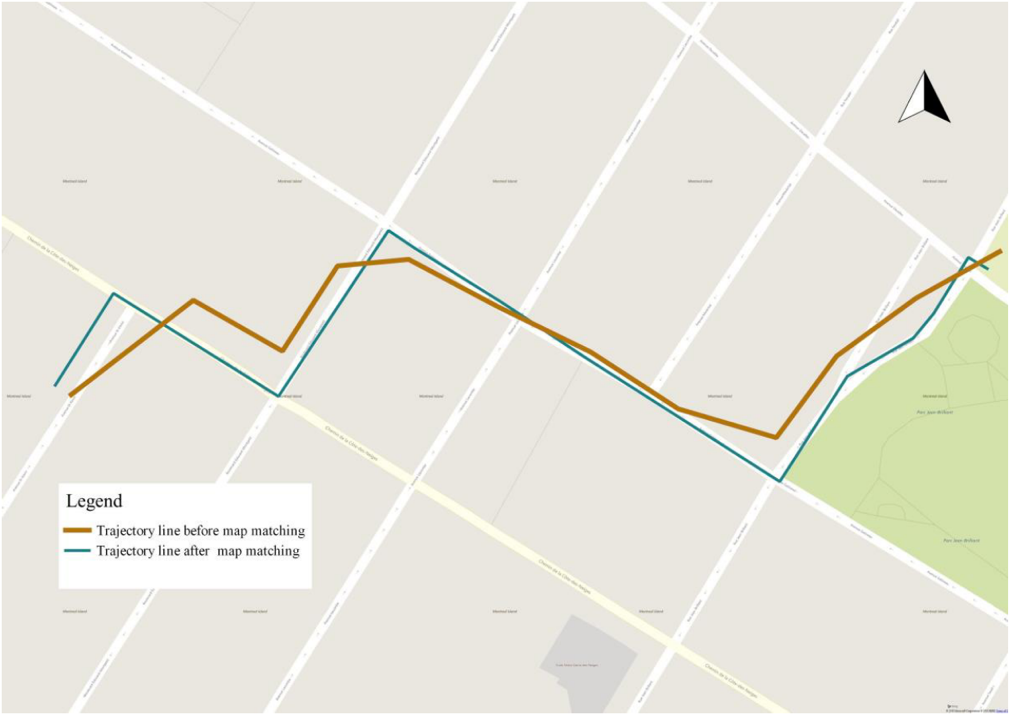
[About the app]:

“﻿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**)



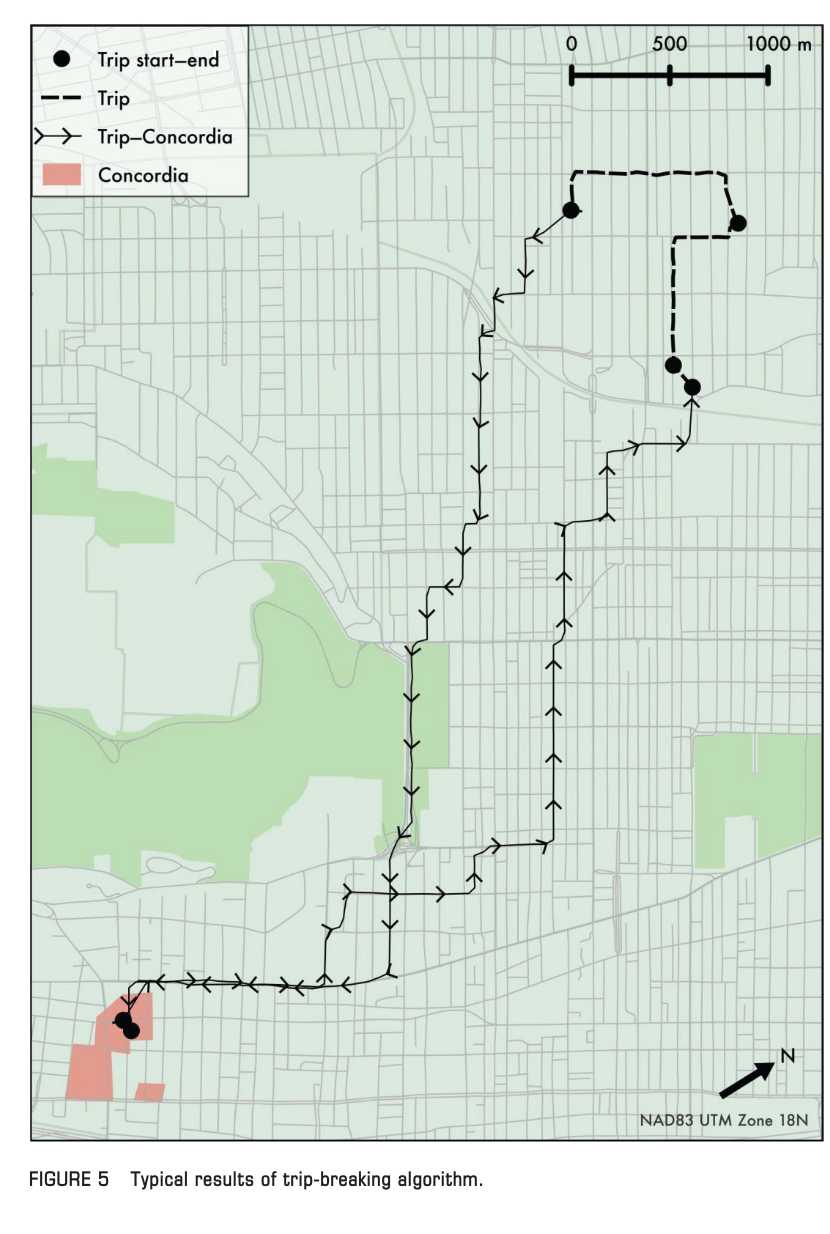
***Figure 2.3*** *Example of GPS trace with location collection priorities within an Itinerum platform app (A); Example of the on screen prompt after an Itinerum platform app stops recording movement (B) (Source: Patterson et al., 2019).*

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

Aluja-Banet, T., Prat, A. & Schwab, I. (2009). Enhancing Socioeconomic Surveys by Data about Internet Usage. [Online]. Available at:

https://www.researchgate.net/publication/244110920\_Enhancing\_Socioeconomic\_Surveys\_by\_Data\_about\_Internet\_Usage [Accessed 23rd August 2019]

An, L., Tsou, M. H., Crook, S. E. S., Chun, Y., Spitzberg, B., Gawron, J. M., & Gupta, D. K. (2015). Space–Time Analysis: Concepts, Quantitative Methods, and Future Directions. *Annals of the Association of American Geographers*, 105(5), 891–914.

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), 93–115.

Arribas-Bel, D., & Tranos, E. (2017). Characterizing the Spatial Structure(s) of Cities “on the fly”: The Space-Time Calendar. *Geographical Analysis*, 1–20.

Attard, M., Haklay, M., & Capineri, C. (2016). The Potential of Volunteered Geographic Information (VGI) in Future Transport Systems. *Urban Planning*, 1(4), 6.

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.

Badu-Marfo, G., Farooq, B., & Patterson, Z. (2019). Perturbation Privacy for Sensitive Locations in Transit Data Publication: A Case Study of Montreal Trajet Surveys. *Transportation Research Records of Transportation Research Board*, 1–19.

Bantis, T., & Haworth, J. (2017). Who you are is how you travel : A framework for transportation mode detection using individual and environmental characteristics. *Transportation Research Part C*, 80, 286–309.

Batty, M., Fosca, G., Pozdnoukhov, A., Bazzani, A., Ouzounis, G., Portugali, Y., & Wachowicz, M. (2012). Smart Cities of the Future. *UCL Working Papers Series*, (188), 0–40.

Batty, M. (2013). Visually-driven urban simulation: Exploring fast and slow change in residential location. *Environment and Planning* A, 45(3), 532–552.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning* Research, 3, 993–1022.

Bricka, S., Moran, M., Miller, K., & Hudson, J. (2015). The Future of TDM: Technology and Demographic Shifts and Their Implications for Transportation Demand Management. ﻿*Texas A&M Transp. Inst.,* Houston, TX, USA, Tech, Rep. PRC 15-25F, 2015.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks, 249–259.

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.

Chandradevan, R. (2017) Random Forest Learning-Essential Understanding. [Online]. Available at: https://towardsdatascience.com/random-forest-learning-essential-understanding-1ca856a963cb [Accessed 26th March 2018].

Cheng, T., & Adepeju, M. (2014). Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PLoS ONE*, 9(6), 1–10.

Cheng, J., Gould, N., Han, L., & Jin, C. (2017). Big Data for Urban Studies: Opportunities and Challenges: A Comparative Perspective. Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE Internationa, 2000(Fig 1), 1229–1234.

Chevalier, W. J., Felteau, C. & McGillivray, B. (2018) Montreal. *In:* Encyclopaedia Britainnica [Online]. Available at: https://www.britannica.com/place/Montreal [Accessed 13th August 2019].

Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS) [Data]. Avaialble at: https://cds.climate.copernicus.eu/cdsapp#!/home [Accessed 10th August 2019].

Dabiri, S., & Heaslip, K. (2018). Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C*, 86(November 2017), 360–371.

De Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences,* 324(2015), 126–145.

Doll, T. (2018) ‘LDA Topic Modeling: An Explanation’, *Medium*, 24 June [Online]. Available at: https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd [Accessed 12th August 2019].

Dubos-golain, A., Trépanier, M., & Morency, C. (2017). Understanding Transit use Patterns in Montreal. Centre interuniversitaire de recherche sur les réseaux d’entreprise, la logistique et le transport (*CIRRELT*).

Eluru, N., Chakour, V., & El-geneidy, A. M. (2012). Travel mode choice and transit route choice behavior in Montreal : insights from McGill University members commute patterns. *Public Tranport*, 129–149.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, 102(3), 571–590.

Fallah-Shorshani, M., Hatzopoulou, M., Ross, N. A., Patterson, Z., & Weichenthal, S. (2018). Evaluating the Impact of Neighborhood Characteristics on Differences between Residential and Mobility-Based Exposures to Outdoor Air Pollution. *Environmental Science & Technology*, 52, 10777–10786.

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

Glenn, S. (2016) ADF — Augmented Dickey Fuller Test, *Statistics, How to* [Online]. Available at: https://www.statisticshowto.datasciencecentral.com/adf-augmented-dickey-fuller-test/ [Accessed 25th August 2019].

Golledge, R. G., & Gärling, T. (2001). Spatial Behavior in Transportation Modeling and Planning. *In:* K. Goulias (Ed.), *Transportation and Engineering Handbook* (1st ed., Vol. 46, pp. 0–37).

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 and Behavioral Sciences*, 138(0), 557–565.

Gong, L., Kanamori, R., & Yamamoto, T. (2018). Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. *Travel Behaviour and Society*, *11*, 131–140.

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

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120.

Goodchild, M. F. (2013) The quality of big (geo)data, Dialogues in Human Geography, 3(3), 280-284.

Grimsrud, M. & El‐Geneidy, A. (2013). Driving transit retention to renaissance: Trends in Montreal commute public transport mode share and factors by age group and birth cohort. *Public transport: Planning and Operations*, 5(3), 119‐241.

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.

Heuvelink, G. B. M., & Griffith, D. A. (2010). Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. *Geographical Analysis*, 42(2), 161–179.

Jahromi, K. K., Zignani, M., Gaito, S., & Rossi, G. P. (2016). Simulating human mobility patterns in urban areas. *Simulation Modelling Practice and Theory*, 62, 137–156

Koehrsen, W. (2017) Random Forest Simple Explanation. *Medium* [Online]. Available at: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d [Accessed 23rd August 2019].

Kumar, K. (2018) Evaluation of Topic Modeling: Topic Coherence, *DataSciencePlus* [Online]. Available at: https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/ [Accessed 23rd August 2019].

Kwan, M. P., & Neutens, T. (2014). Space-time research in GIScience. International Journal of *Geographical Information Science*, 28(5), 851–854.

Kwan, M. P. (2018). The Limits of the Neighborhood Effect: Contextual Uncertainties in Geographic, Environmental Health, and Social Science Research. *Annals of the American Association of Geographers*, 108(6), 1482–1490.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An introduction to statistical learning. London: *Springer*, 102, pp.303-368.

Japkowicz, N. (2000). The Class Imbalance Problem : Signi cance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI).

Li, S., Dragicevic, S., Antón, F., Sester, M., Winter, S., Coltekin, A., … Cheng, T. (2016). ISPRS Journal of Photogrammetry and Remote Sensing Geospatial big data handling theory and methods : A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119–133.

Lin, M., & Hsu, W. (2014). Mining GPS data for mobility patterns : A survey. *Pervasive and* *Mobile Computing*, 12, 1–16.

Liu, Y., & Cheng, T. (2018). Transportmetrica A : Transport Science Understanding public transit patterns with open geodemographics to facilitate public transport planning. Transportmetrica A: Transport Science, 0(0), 1–28.

Lyons, G., & Harman, R. (2002). The UK public transport industry and provision of multi-modal traveller information. *International Journal of Transport Management*, 1(1), 1–13.

Miller, H. J. (2013). Beyond sharing: Cultivating cooperative transportation systems through geographic information science. *Journal of Transport Geography*, 31, 296–308.

Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. (2014). Trip purpose identification from GPS tracks. *Transportation Research Record*, (2405), 16–23.

MTL Trajet (2017) Results of the 2017 MTL Trajet study [Online]. Available at: https://ville.montreal.qc.ca/mtltrajet/en/etude/ [Accessed 13th August 2019].

Murray, A. T., Liu, Y., Rey, S. J., & Anselin, L. (2012). Exploring movement object patterns. *Annals of Regional Science*, 49(2), 471–484.

Openshaw, S. (1984) Ecological fallacies and the analysis of areal census data. *Environment and Planning*, 16, 17–31.

Patterson, Z., & Fitzsimmons, K. (2016). DataMobile Smartphone Travel Survey Experiment. *Transportation Research Record: Journal of the Transportation Research Board*, 2593, 35–43

Patterson, Z. and Fitzsimmons, K. (2017a) MTL Trajet. *Working Paper* *2017-2*, Concordia University, TRIP Lab, Montreal, Canada, 2017.

Patterson, Z., & Fitzsimmons, K. (2017b). The Itinerum Open Smartphone Travel Survey Platform The Itinerum Open Smartphone Travel Survey Platform, (July).

Patterson, Z. (2017). MTL Trajet. [*Mobile* *app*]. Version 2.0.6. Available from: https://apps.apple.com/us/app/mtl-trajet/id1131355971 [Accessed 2nd August 2019].

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. *Transportation Research Part C: Emerging Technologies*, 75, 197–211.

Rey, S.J. & Anselin, L. (2007) PySAL: A Python Library of Spatial Analytical Methods, *Review of Regional Studies* 37, 5-27.

Roubeyrie, L. & Celles, S. (2018) *windrose v1.6.7* [Software]. Available at: https://github.com/python-windrose/windrose [Accessed 1st August 2019]

Semanjski, I., Gautama, S., Ahas, R., & Witlox, F. (2017). Spatial context mining approach for transport mode recognition from mobile sensed big data. Computers, Environment and Urban Systems, 66, 38–52.

Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics.

Shi, W., Zhang, A., Zhou, X., & Zhang, M. (2018). Challenges and Prospects of Uncertainties in Spatial Big Data Analytics. *Annals of the American Association of Geographers*, 108(6), 1513–1520.

Sioui, L., Morency, C., & Trépanier, M. (2012). How Carsharing Affects the Travel Behavior of Households: A Case Study of Montréal, Canada. *International Journal of Sustainable Transportation*, 7(1), 52–69

Statistics Canada (2016) Dissemination Areas, Digital Boundary File - 2016 Census [Data]. Available at: https://open.canada.ca/data/en/dataset/5a6e8f76-cfd2-4a69-acee-9ed205dd9556 [Accessed 1st August 2019].

Statistics Canada (2019) Census subdivisions – 2011 Census - Boundary files [Online]. Available at: https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2011-eng.cfm [Accessed 1st August 2019].

Tayyab, M., Dauwels, J., Goh, C. Y., Oran, A., Fathi, E., Xu, M., … Jaillet, P. (2014). Spatial and Temporal Patterns in Large-Scale Traffic Speed Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2), 794–804.

The Data Science Blog (TDSB) (2016) A Quick Introduction to Neural Networks [Online]. Available at: https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/ [Accessed 23rd April 2019].

Tu, W., Cao, J., Yue, Y., Shaw, S. L., Zhou, M., Wang, Z., … Li, Q. (2017). Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*, 31(12), 2331–2358.

Ville de Montréal – Portail Données Ouverts (2017) Déplacements MTL Trajet [*Data*]. Available at: http://donnees.ville.montreal.qc.ca/dataset/mtl-trajet [Accessed 10th July 2019].

World Population Review (WPR) (2019) Montreal Population 2019 [Online]. Available at: http://worldpopulationreview.com/world-cities/montreal-population/ [Accessed 13th August 2019].

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(67), 1–19.

Xie, K., Xiong, H., & Preparation, A. D. (2016). The Correlation between Human Mobility and Socio-demographic in Megacity. 2016 *IEEE International Smart Cities Conference* (ISC2), 1–6.

Xue, A. Y., Zhang, R., Zheng, Y., Xie, X., Huang, J., & Xu, Z. (2013). Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In: *Proceedings of the 29th IEEE International Conference on Data Engineering*. *IEEE*, 254–265

Yamada, I., & Thill, J. C. (2010). Local Indicators of Network-Constrained Clusters in Spatial Patterns Represented by a Link Attribute. *Annals of the Association of American Geographers*, 100(2), 269–285.

Yazdizadeh, A., Patterson, Z., & Farooq, B. (2019). An automated approach from GPS traces to complete trip information. International Journal of Transportation Science and Technology, 8(1), 82–100.

Zahabi, S. A. H., Ajzachi, A., & Patterson, Z. (2017). Transit Trip Itinerary Inference with GTFS and Smartphone Data. *Transportation Research Record: Journal of the Transportation Research* Board, 2652(1), 59–69.

Zhang, Y., & Cheng, T. (2019). A Deep Learning Approach to Infer Employment Status of Passengers by Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, PP, 1–13.

Zhang, Y., Cheng, T., & Aslam, N. S. (2019). Exploring the Relationship Between Travel Pattern and Social-Demographics Using Smart Card Data and Household Survey. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W13(June), 1375–1382.

Zhao, Z., Shaw, S. L., Yin, L., Fang, Z., Yang, X., Zhang, F., & Wu, S. (2019). The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data. International Journal of Geographical Information Science, 33(7), 1471–1495.

# Appendices

## Appendix 1 Notification not to apply for Ethical Approval

As this study makes use of data that is public available from the City of Montreal’s Open Data Portal, this study was considered to be minimal risk due to the subsection X of UCL’s ethics guide.

# Appendix 2 Mean Direction and Distance Calculations

The python script used to carry out these calculations is available in *direction\_functions.py* available athttps://github.com/Thomasjkeel/MSc\_Dissertation under the ‘*utils’*

## Appendix 3 Python Scripts used for the analysis carried out in this report.

Python script is available online at https://github.com/Thomasjkeel/MSc\_Dissertation under ‘. (This program consists of 500 lines written in Python computer language).  
*Description:* contains the code to create the figures and carry out the statistical analysis of this IGS.