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

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 London

Supervisor: Huanfa Chen

30th August 2019

Abstract

The prediction of *why* people travel when they move 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 detail the purposes of individual's travel 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, Volunteered Geographic Information (VGI) provided by smartphones travel surveys have provided researchers an opportunity to study the attributes characterising urban mobility patterns within a city at increasingly fine temporal and spatial scales. This study makes uses of one such source of VGI: the *2017 MTL Trajet* travel survey app – a project with the aim to study *how* and *why* people move within the City of Montreal, Canada. Regarding (2), this project builds upon a small body of research to uncover and categorise spatial and temporal interdependencies of GPS data provided from the MTL Trajet project, before assessing the performance of three machine-learning classification models used to classify this GPS data: Random Forests, Support Vector Machines and Artificial Neural Networks. Specifically, 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

List of Figures

List of Tables

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

TC – Temporal Clusters

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

The purposes by which populations 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 large datasets which combine both the geographically coordinates of people's movement (i.e. a GPS trace) and the activities for *why* people make these movements (i.e. for Work, Leisure, etc).

In recent years, improvements to GPS within 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 data in these travel survey apps can help researchers generate unique insight into transport behaviour at much finer 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 urban mobility patterns 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* people move along the city's transport network. 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. 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 (referred to as 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, we first need to understand the temporal and spatial scales at which people travel for certain activities. The motivation of this study is thus to evaluate whether we can use spatial and temporal dependencies as key indicators for trip purpose classification models.

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 'complete' that trip and would be asked:

- 'Which travel modes did you use for this trip?'
- 'Why did you make this trip?' (see similar example in **Figure 1.1B**).

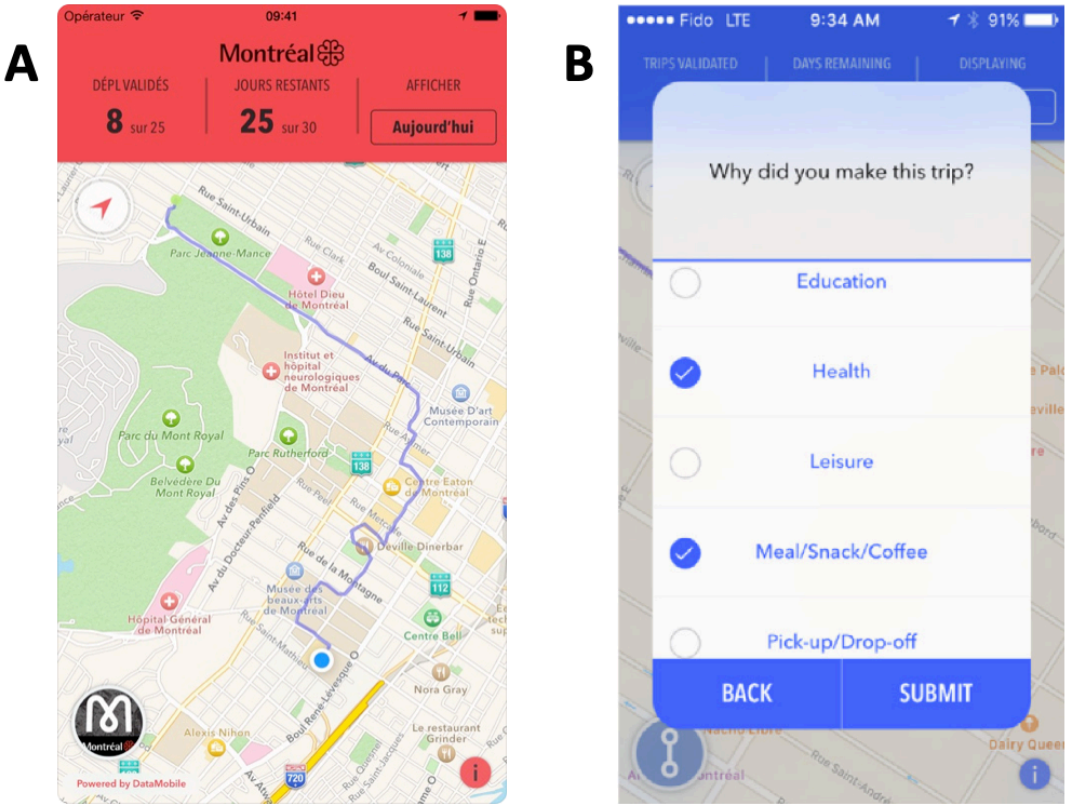

Figure 1.1 *(A) Screenshot from the MTL Trajet app showing a recorded GPS trace (source: Patterson, 2017a). (B) Example of prompt similar to one used in the 2017 MTL Trajet app (source: Patterson et al., 2019).*

Both the responses to these questions and GPS trace available in the data from the MTL Trajet are used to test three types of classification models that look to characterise the *purpose* of movement. These models are specifically:

1. Random Forests

2. Support Vector Machines

3. Artificial Neural Networks

Various temporal and spatial generalisation techniques will be used to represent the space and time trends seen within the data before being used in the models as predictors for trip purpose.

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 the MTL Trajet survey.

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

2.1.1 Overview

Although a wealth of literature exists regarding the classification of transport modes from GPS traces, investigation into the classification of transport purpose has received far less attention (Yazdizadeh *et al.*, 2019). One reason for this is that users are required to manually provide information about *why* they have made a trip (as a GPS trace and timestamp is not sufficient alone) (Gong *et al.*, 2014). Notably, mode-classification algorithms often only need few key-identifiers such as speed, acceleration and distance (which are recorded automatically without user-input) to have high accuracy (Dabiri & Heaslip, 2018). This differs from purpose-classification algorithms where some degree of qualitative information about the individual users is needed. Correspondingly, Yazdizadeh *et al.* (2019) find that mode-classification models are often shown to be more accurate on average than purpose-classification.

Of studies that set out to build purpose-classification models, Gong *et al.* (2014) characterise three distinct types:

1. Rule-based (using rules to match GPS signal and qualitative identifiers),
2. Probabilistic (using the calculated probability of a given purpose);
3. Machine learning.

And a selection of key classification models from the literature from each one these types are detailed in **Table 2.1** along with their inputs and accuracy.

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

Author(s)	*Predictor variables*	*Location and date of data*	*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)	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; Spatial Clustering	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%

As highlighted in **Table 2.1**, methods employing the use of Random Forest classifiers (RF) are currently the most popular used (Gong *et al.*, 2018). The trend in the literature has been to train RFs with a high number of inputs and then reduce these using the *feature importance* as indicator of which inputs are pertinent to the model's performance. It is likely this trend owes to a lack of understanding around the specific combination of dynamics which govern why people make trips – a major gap in the research of trip purpose classification (Meng *et al.*, 2019).

The inputs used in trip purpose models detailed in **Table 2.1** typically include a combination of user-inputted information and underlying spatial (e.g. distance to respondent's home/work places; POI; Land Usage), temporal (e.g. time of day; day of week) and socio-demographic (e.g. 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 around education, work and returning home. Arguably, shopping and leisure activities may tend to be less temporally and spatially *structured* compared to work, education and returning home activities (Lin & Hsu, 2014).

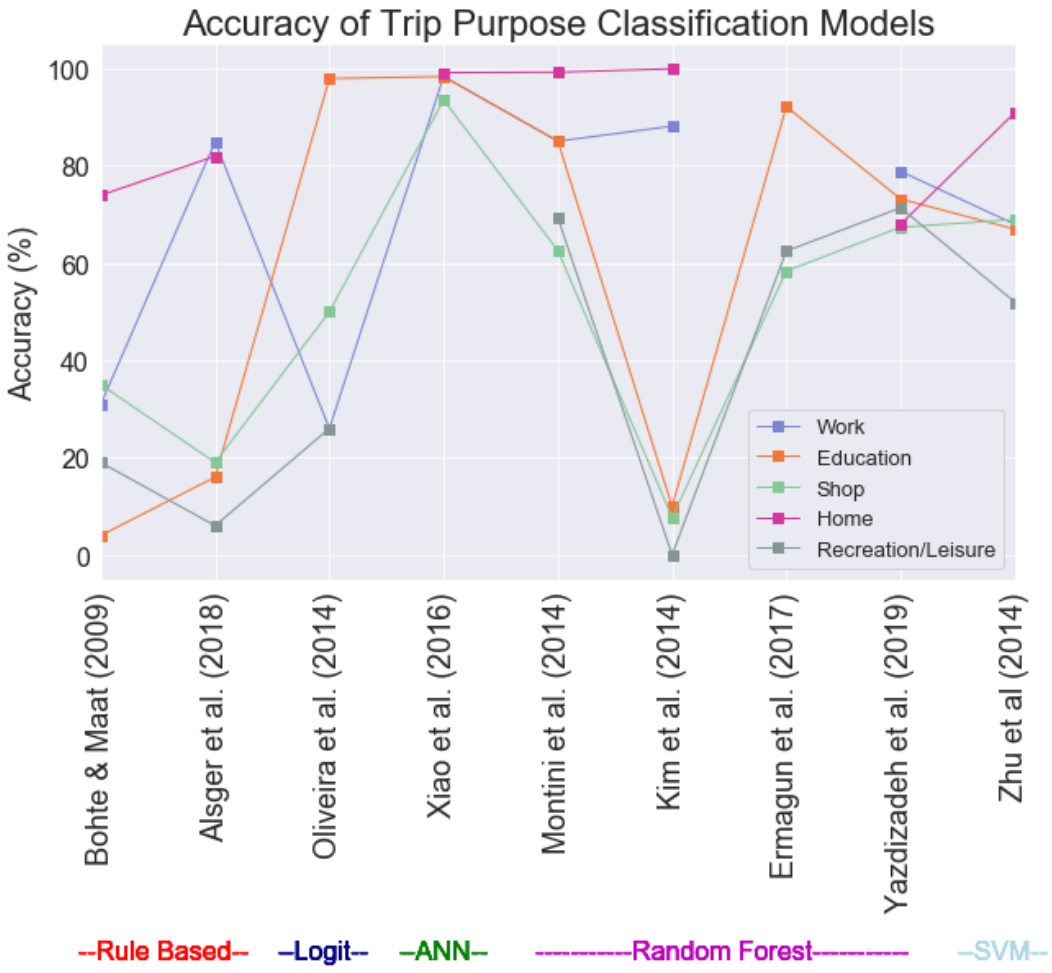

Figure 2.1 *Comparison of trip purpose classification model accuracy within the literature (ANN=Artificial Neural Network; SVM=Support Vector Machine)*

2.1.2 Spatial and temporal representation in trip purpose classification models

Generally, spatial and temporal features have been identified as the key indicators in trip purpose classification (e.g. Zhu *et al.*, 2014; Yadizadeh *et al.*, 2019) as opposed to socio-demographic features. Despite this, spatial and temporal features have not been applied with any uniform standard throughout the literature (Aslger *et al.*, 2018).

In some cases, only the proximity of the start and end points of the trips to local POIs (Points of Interest) and Personal Locations (i.e. Home and Work) are used to infer about the

purpose of a respondents trips (e.g. Kim *et al.*, 2015 & Ermugun *et al.*, 2017; **Table 3.1**). In other cases, closer attention has been paid to reducing the spatial and temporal complexity of the trips, such as generalising these features through clustering. An example of this spatial generalisation is seen in Montini *et al.* (2014) who build a high performance trip purpose classification model that makes use of clustering algorithms to group origin and destination points of user trips.

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. There is also less attention on studying the changes in different types of trip purposes based on daily and weekly trends (Meng *et al.*, 2019).

2.1.3 Key issues raised by existing trip purpose research

One major issue in the literature is that there is little investigation into the longer term effects and seasonality of changes to the trip purpose. Xie *et al.* (2016) find that weather can fundamentally change how people travel, so including weather in any model that seeks to predict travel is vital. Further, it has even been found that seasonality can severely alter which activities (or purposes) people carry out (Gong *et al.*, 2018). Correspondingly, many of Montreal's Festivals take place during the months of July–September, which has an effect on the activities people carry out within the city during these months (Grimsrud, M. & El-Geneidy, 2013).

Also evident in the literature is the fact that the modelling procedure has been approached in a range of different ways. Some studies focus on building individual models for each unique trip purpose and others build all-encompassing, multi-class classification models which account for all unique trip purposes at once. Generally, multi-class has been more effective in the literature (Alsger *et al.*, 2018).

Finally, the majority of the studies ignore the underlying class imbalance of the answers selected by respondents relating to *why* they have made a particular trip. In most studies, the majority of trips are where the respondent has travelled to *work* or is *returning* home, as opposed to a minority trips where the respondent has visited *shops* or *hospitals* (Meng *et al.*, 2019). One case where class imbalance is considered is in Xiao *et al.* (2016) who use under-sampling technique to account for the disproportion of these trip purpose categories.

2.2 Volunteered Geographic Information in mobility research

Volunteered Geographic information is essentially crowd sourced data which is defined formally as the "widespread *engagement of large numbers of private citizens […] in the creation of geographic information*" (Goodchild, 2007, p.212). As VGI is, by definition, volunteered, in some cases it gives us the opportunity to study more personal and subjective forms of information than traditional forms of data collection (such as telephone surveys) (Elwood *et al.*, 2012).

Within mobility studies, VGI has enabled us to study movement at increasingly fine spatial and temporal scales, allowing us to better understand travel in urban mobility patterns than

ever before (Arribas-Bel & Tranos, 2017; Zahabi *et al.*, 2017). This is because research initiated through smartphones has the potential to reach a larger number of people (as more people have smartphones than GPS devices) (Wu *et al.*, 2016). Notedly, smartphones offer a more cost effective solution than GPS devices, which have been traditionally used in mobility research (Gong *et al.*, 2014; Shi *et al.*, 2018).

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 because the modes of travel 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 e.g. with gentrification, this can an effect on the travel patterns and activities carried out within some parts of a city (Bricka *et al.*, 2015). VGI gives us the opportunity to witness these patterns within crowd sourced data.

Finally, Li *et al.* (2016) distinguish between two types of VGI:

1. Participatory – which is the conscious inclusion of data by private citizens (in the context of this study this may be in-app responses to trip purpose of the MTL Trajet),
2. Opportunistic – which is unconscious inclusion of data (in the context of this study this may be a GPS trace).

The success of mobility research is often dependent on combination of both types being present within a study.

2.2.2 Issues with in VGI in mobility studies

As many forms of volunteered information require that the users share their data themselves, this creates problems of representativeness in VGI (Li *et al.*, 2016). The general trend in most travel-based surveys is that minority a respondents make up the majority of data, as only some people are willing to share their information (Goodchild & Li, 2012).

There are also problem with geographical representativeness as, VGI tend to be biased towards cities and in richer nations (Hecht & Stephens, 2014). As such, Miller & Goodchild (2014) argue that we must be careful when making generalisations about larger populations (than sample size) from any form of VGI. In terms of travel surveys where trip-purpose has been collected, these often tend to emphasise behavioural patterns of people from certain socio-demographics such as people that want to make their trip intentions known (Kim *et al.*, 2015).

Moreover, there are credibility issues with VGI due to a lack of quality control in its creation (Flanagin & Metzger, 2008; Goodchild, 2013). VGI collected from travel surveys are particular hindered by this, as we do not ultimately know who each one of the individual participants are and whether they have inputted data correctly (Shi *et al.*, 2018). As such, researchers collecting VGI often have to trust that respondents do not purposefully create mis-information (Attard et al., 2016).

2.3 The *MTL Trajet* mobile travel survey

Despite the potential the potential of VGI, there are many cities globally that have no form of formal research initiated within them (Attard *et al.*, 2016). One exception to this, is in Montreal, Canada where a number of mobility-based travel surveys have been conducted in the last few years. One of these is the MTL Trajet travel survey project.

The MTL Trajet is a large scale mobile phone travel survey app that has been run yearly between Oct-Nov and has been conducted since 2016 (Ville de Montreal, 2019). The app itself is built from the *Itinerum platform* which is a framework providing researchers a platform to develop their own travel surveys (Yazdizadeh *et al.*, 2019)

The original aim of the MTL Trajet was to better understand transport behaviours within Montreal (MTL Trajet, 2017), although it is repurposed in this report to investigate trip purpose classification. The MTL Trajet survey originally contained information from its users about their personal locations (i.e. work and residency), although these have been removed from the data before being published in Montreal's Open Data Portal (Hamouni, 2018).

There are two available sources of MTL Trajet dataset available on Montreal's Open Data Portal, one that contains raw GPS points and one that contains geo-routed GPS traces from the respondents, the latter of which is used in this report (Ville de Montréal, 2019). Specifically, the geo-routed traces were created after feeding the raw GPS points into the Open-source Routing Machine, which maps the users' trajectory (see **Figure 2.2**; Patterson & Fitzsimmons, 2017b). Because of this, there is a degree of unaccountable inaccuracy in the data used is in this report.

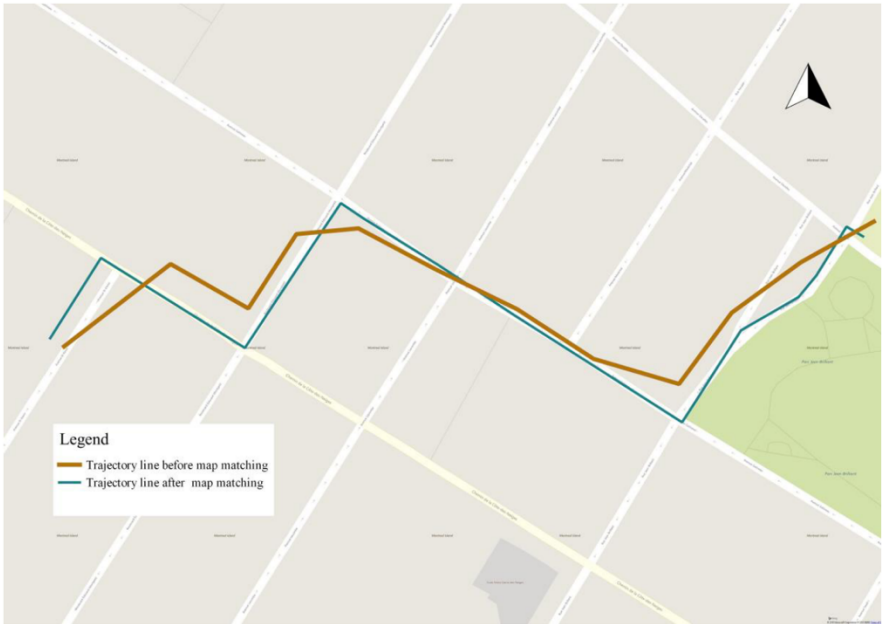

Figure 2.2 *Example of routing done by the Open Source Routing Machine when processing the raw GPS trace from user devices (Source: Hamouni, 2018).*

An *Itinerum Platform* survey app works by employing a geofencing technique, which updates the sampling frequency of the GPS recordings from the smartphone whilst the user is moving (Patterson *et al.*, 2019). When the user stops for more than 120 seconds, the update rate of GPS recordings drops and the user is prompted to end the trip and answer questions relating to *how and why* they have made the trip (**Figure 2.3**; Patterson *et al.*, 2019). The app begins recording movement again when the user leaves the geofence once again.

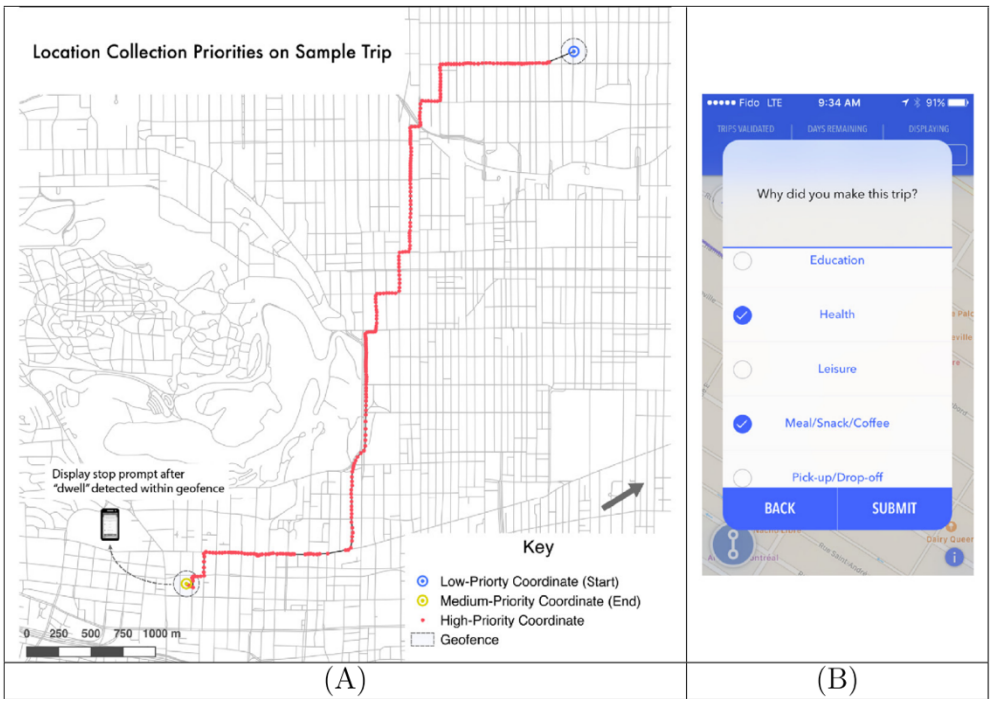

Figure 2.3 *Example of GPS trace with location collection priorities within an Itinerum platform app (A); Example of the on screen prompt after an Itinerum platform app stops recording movement (B) (Source: Patterson et al., 2019).*

The MTL Trajet has seen limited use in the literature and has instead mainly been restricted to use within the City of Montreal's transport department (MTL Trajet, 2017). One example in the literature is in Yazdizadeh *et al.* (2019) who use the 2016 edition of MTL Trajet survey data for transport purpose classification models (which is described in **Table 2.1**).

Chapter 3. Methodology

3.1 Study Area

The study area chosen for this project spans the Greater Montreal region in Eastern Canada. To create this study area, a shapefile containing all of Canada's 54,000 dissemination areas (DAs) – which are the smallest standard geographic area available on the 2016 Canadian census – was retrieved from Statistics Canada (2016). Using QGIS, a spatial intersect was then calculated between all of the DAs and the GPS traces of respondents to the 2017 MTL Trajet survey to select only areas where data there was an overlap. An illustration of **Figure 3.1**, the result of this selection is a study area of 7,046 DAs which are used in the analysis of this report.

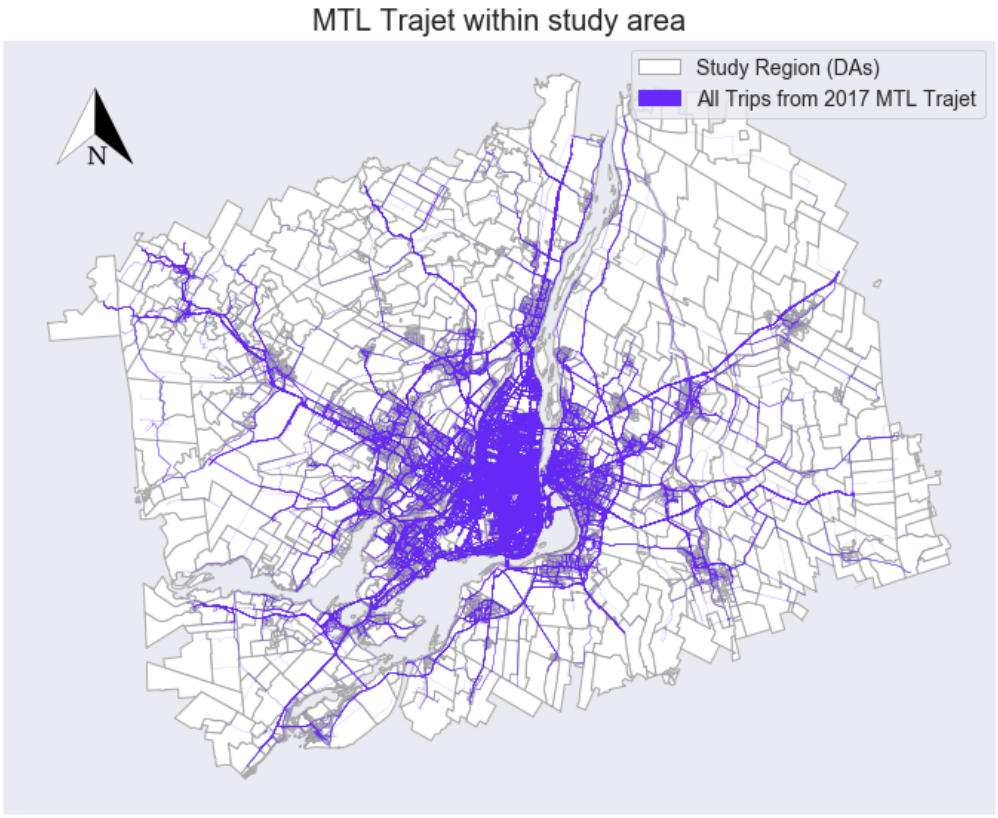

Figure 3.1 *GPS routes from the 2017 MTL Trajet travel survey plotted within the study area.*

Two further shapefiles outlining the geographical boundaries of the city of Montreal, Greater Montreal region were retrieved from Canada's *Open Government Portal* (Statistics Canada, 2019). To allow the analysis of this project, all geographically-referenced data were re-projected into the Statistics Canada Lambert (or NAD83). This is a Canadian-centric projection with a 1 metre unit (EPSG, 2019). The re-projection of the data was carried out using Python's *Geopandas* library. The total of area of the region chosen for this study is 4,279 km^2 and the extent of the City and Greater regions of Montreal are shown in **Figure 3.2**.

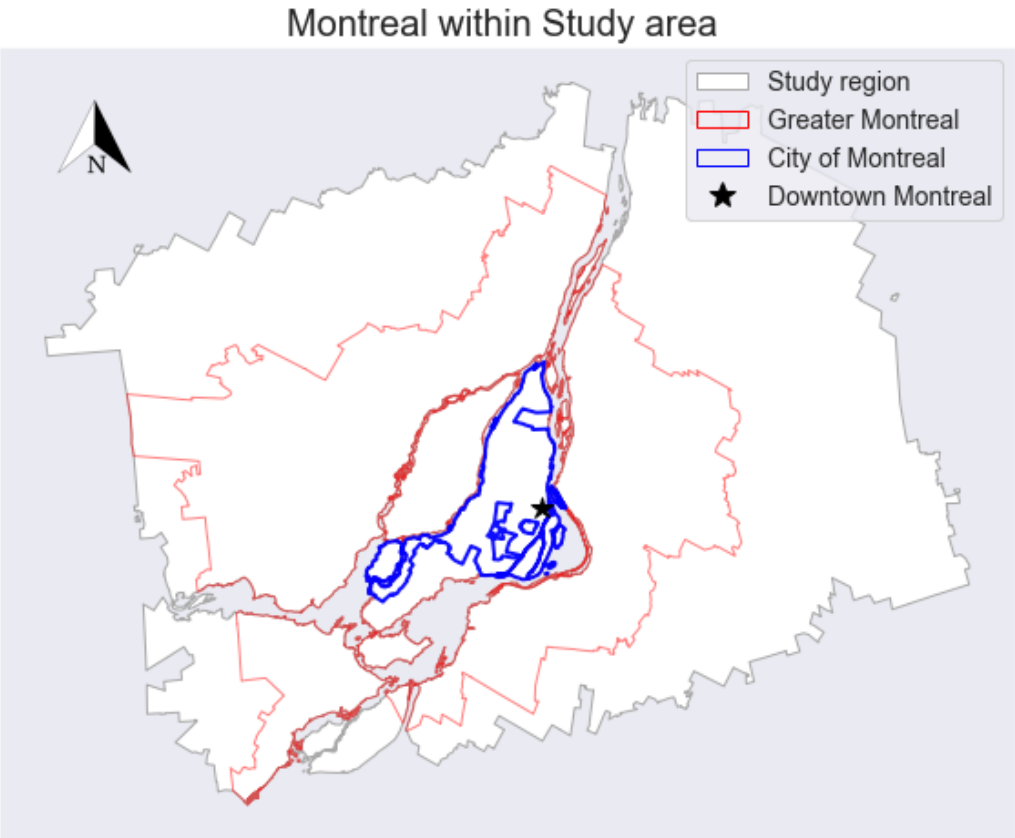

Figure 3.2 *Location of Montreal within the study area.*

3.2 Data collection and pre-processing

3.2.1 2017 MTL Trajet Survey

Data detailing the results of the *2017 MTL Trajet* smartphone travel survey carried out within Montreal, Canada between 18th September 2017 and 18th October 2017, was retrieved from the Montreal Open Database (Ville de Montréal, 2017). This data, which is in a GeoJSON format, has already been pre-processed and cleaned and details 185,285 unique trips from 4,425 unique respondents (Ville de Montréal, 2017). Each unique trip in the dataset contains a unique identification number, a user-defined label for the *mode* and *purpose* of the trip; a start and end timestamp, and a spatial reference or geometry. An outline and description of these variables are given in **Table 3.1**.

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

Column	*Description*	*Format*	*N*
id_trip	Unique identification number of the trip	Integer	185,285
mode	The means of transport used for a trip	String	74,218
purpose	The class of activity for which that trip is for	String	74,218
starttime	Date and time when the trip begun	Datetime	185,285
endtime	Date and time when the trip finished	Datetime	185,285
geometry	Coordinates detailing the route of a trip	LineString	185,285

The geometry of each trip, specifically, contains a collection of line segments (LineString format) derived from the original GPS trace from the user's smartphone (see *2.3*). For this analysis, the geometry has been re-projected from WGS84 into NAD83 using *GeoPandas.*

All aspects of the data has been translated from French to English and the unique categories of the mode and purpose of the trips are shown in **Table 3.2**. Note that although the MTL

Trajet app allowed respondents to choose any combination of travel mode categories per trip, it only allowed *one* category of travel purpose per trip.

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

Category (variable name)	*Number of unique categories*	*Unique categories*
Mode of Trip	70*	Car, Cycling, Not available, Other, Public transport, Taxi, Walking
Purpose of Trip	11	Café, Education; Health, Leisure, Not available, Other, Pick up a person, Returning home, Shops, Work

* combination of any number of unique categories

The time signature for the start and end of each trip has been converted from Coordinated Universal Time (UTC) to Eastern Daylight Time (EDT). This being the time zone that Montreal falls within and for the purpose of analysis using Python's *datetime* library. The duration of each trip was calculated in seconds by taking the difference between these two time signatures. The total distance in metres of each trip was calculated using Python by taking the sum of Euclidean distances between pairs of points within a given trip.

3.2.2 Supplementary data

This study makes use of two supplementary data sources detailing the land use categories in the city of Montreal and weather in Montreal between 18th September 2017 and 18th October 2017.

Land Use Data from the *City of Montreal's 2014 Plan d'urbanisme* was collected from the Montreal Open Database (Ville de Montréal, 2014). The data, which is in a GeoJSON format,

contains ten unique categories of land use within the City of Montreal, these are detailed in **Table 3.3** and mapped in **Figure 3.3**. The purpose of adding this data is primarily to add contextual spatial information to where trips begin and end within Montreal. A *spatial join* is carried out using *Geopandas* to find the category of land use for each trip's origin and destination.

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

Land Use Category	*Description*	*Total Area (%)*
Agricultural	Farmland	6.38
Conservational	Wildlife reserves	8.85
Diversified activities	No one category can be applied	6.81
Employment	Company offices and places of work	16.67
Infrastructure	stations, railway lines, airports, etc	8.12
Institution	Major facilities including governmental and private	5.20
Mixed	Residential and employment	8.19
Park	Including green spaces	9.09
Religious	Churches, Mosques, Synagogues, etc.	2.84
Residential	Homes	27.86

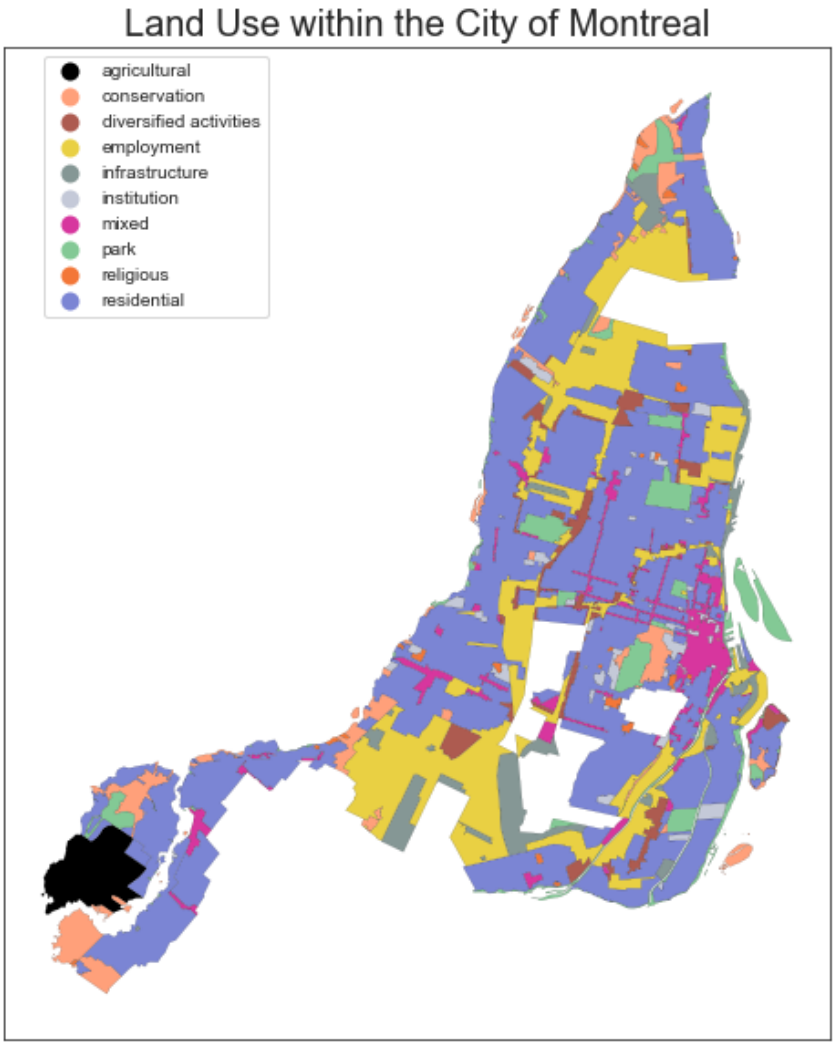

Figure 3.3 *Map showing land use categories within the City of Montreal (data from: Ville de Montréal, 2014).*

For the weather data, 2-m surface temperature (°C) and precipitation level (mm) data at 1464 1-hour intervals for the dates 18th September 2017 – 18th October 2017 were retrieved from the ERA-5 climate reanalysis dataset produced by the Copernicus Climate Change Service (C3S, 2017). This data covers a 1° × 1° degree area over Greater Montreal (45° N, -73° W) and was retrieved in a *netcdf4* format through Python using the Climate Data Store API client (see Appendix 2). This data was loaded into Python using the *iris 2.0* library (Met Office, 2018) before being re-formatted and output into csv. The purpose is to supplement

the information from the trips, as it has been found in the literature, that weather can influence people's activities and affect how they choose to travel (Dubos-Golain *et al.*, 2017; Gong *et al.*, 2018). Python's *datetime* and *Pandas* libraries are used to join the relevant temperature and precipitation level to timestamp of each trip within the MTL Trajet data.

3.3 Development of space and time model inputs

3.3.1 Rush hour and City Labels

A number of both spatial and temporal metrics have been created from the data to aid the ability of the trip purpose classification models used in this project. Binary labels were created from the data to indicate whether a trip occurred inside or outside of the City of Montreal (**Figure 3.2**) after a similar method for transport mode inference in Zahabi *et al.*, 2017)**.** This was calculated using the *intersects* method from Python's *Shapely* between each trip and a shapefile of the City of Montreal.

Binary labels were created to distinguish whether a trip begun in, passed through or ended in a 'Rush-Hour' or 'Off-Peak' period (**Table 3.4**) and whether the trip had begun or ended on a weekday or weekend (after Liu & Cheng, 2018). These binary labels were created to give more context to the study area by:

- Differentiating between the city proper and its suburbs
- Differentiating between times of day and days of week.

As we expect the governing spatial and temporal dynamics to change throughout the day and across the city (Cheng *et al.*, 2014), these labels give the classification models in this project more context between broad units of space and time.

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

Section	*Times*	*Days*	*Hours (each day)*
Rush hour	6:00–10:00 & 15:00–19:00*	Monday – Friday	8
Off-peak	Times outside *Rush hour*	Saturday – Sunday	16

* after Howell (2018)

3.3.2 Trip direction

The mean cardinal direction of each trip (e.g. North, North-East, South-West, etc.) has been calculated to investigate the directional dependence of each given category of trip purpose within the MTL Trajet data. See an example of an Eastbound trip within Montreal in **Figure 3.4**.

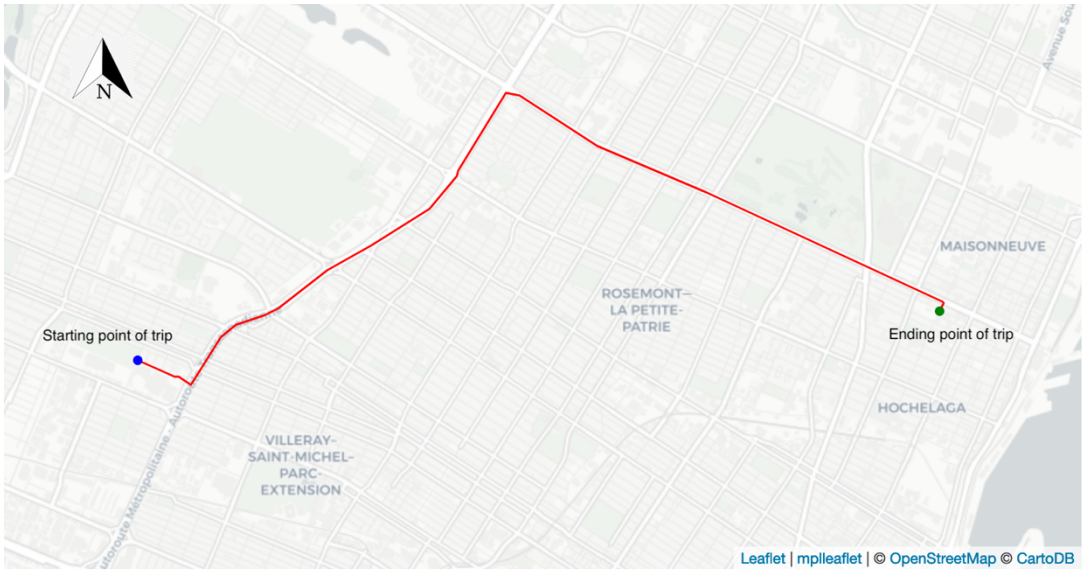

Figure 3.4 *Example of an eastbound trip across Montreal (trip-id= 150744)*

To achieve this, each trip (in a LineString format) was first re-projected from Canada Lambert into World Geodetic System 1984 projection (*EPSG-4326*) using *Shapely*. Individual

trips were then broken down into an array of latitude-longitude points-pairs. The bearing (θ) in decimal degrees between each point pair was then calculated using the following:

$$\theta(°) = arctan\left(\begin{matrix}\sin(\Delta lon)\cdot\cos(lat_2), \\ \cos(lat_1)\cdot\sin(lat_2) - \sin(lat_1)\cdot\cos(lat_2)\cdot\cos(\Delta lon)\end{matrix}\right) \quad (1)$$

where $lat_1, lon_1, lat_2, lon_2$ refer to the coordinates in decimal degrees of the first and second points of a pair respectively. Euclidean distance in degrees (D (°)) was also calculated for each point pair using:

$$D\ (°) = \sqrt{(lat_2 - lat_1)^2 + (lon_2 - lon_1)^2} \quad (2)$$

Using the collection of distances and bearings, calculated in (1), (2), for each trip, we can calculate the mean cardinal direction and distance magnitude for each trip (calculations of which are detailed in Appendix 2). Python's *windrose* library have been used to create circular histograms of trip direction for each unique purposes (Roubeyrie & Celles, 2018).

3.3.3 Spatial and temporal clustering

To reduce the complexity of the space and time signatures in the MTL Trajet data (see **Table 3.1**), each trip has been assigned a label relating to one spatial and temporal cluster. The use of clusters as model inputs, as opposed to raw latitude/longitude coordinates, is chosen in the hope that they will improve ability of the classifiers to generalise about spatial and temporal structures across the trips and to speed up model training times (Montini *et al.*, 2014).

For use in a k-means clustering algorithm, the starting and ending coordinates of each individual trip are extracted. Using only these coordinates was decided this on the assumption that each trip is an interaction between an origin and destination (Murray *et al.*,

2012), and thus the choice of travel route (i.e. which roads are taken) between them is likely less important for a classification model.

The k-means clustering algorithm is an unsupervised technique to iteratively partition a given (k) amount of data classes within data space and was chosen over density-based clustering techniques such as DBSCAN in the interest of computational time (De Amorim & Hennig, 2016). This algorithm is carried out using *Scikit-Learn* for a range of values of k between 2-20, each of which are compared and evaluated for their effectiveness using their silhouette score – a metric evaluating how well each data point fits into its assigned cluster (De Amorim & Hennig, 2016).

Although, there are some mobility studies that use spatial clusters as explanatory variables in trip purpose classification (e.g. Montini *et al.,* 2014; Yazdizadeh *et al.*, 2019), less attention has been paid to temporal clustering. One example of temporal clustering being used is in a Liu & Cheng (2018) who adapt a Latent Dirichlet Allocation model to better account for temporal structures in movement data from smart cards. Temporal clusters (TCs) therefore have been extracted from the data in this report similarly using a Latent Dirichlet Allocation (LDA) model adapted from Liu & Cheng (2018).

LDAs are probabilistic topic identification models commonly used in the classification of semantics in large bodies of unstructured text (Blei *et al.*, 2003). As LDA models are natural language processing algorithms, information from the MTL Trajet regarding the day, time and purpose were converted into ‘temporal words’ (after Liu & Cheng, 2018). For example,

a trip to work beginning at 7.15am on Monday becomes a sentence containing one temporal word: 'Monday_7' and one trip-purpose word: 'work'.

As shown visually in **Figure 3.5**, a total of 168 temporal words (and further 10 unique trip-purpose words) are used to build the LDA in this report. Specifically, the model will be able to discover patterns and discern words that have a high probability of clustering (i.e. 'work' and 'Monday_7'). This probability is called the *weighted importance* which each temporal and trip-purpose word will be assigned in the LDA (Doll, 2018).

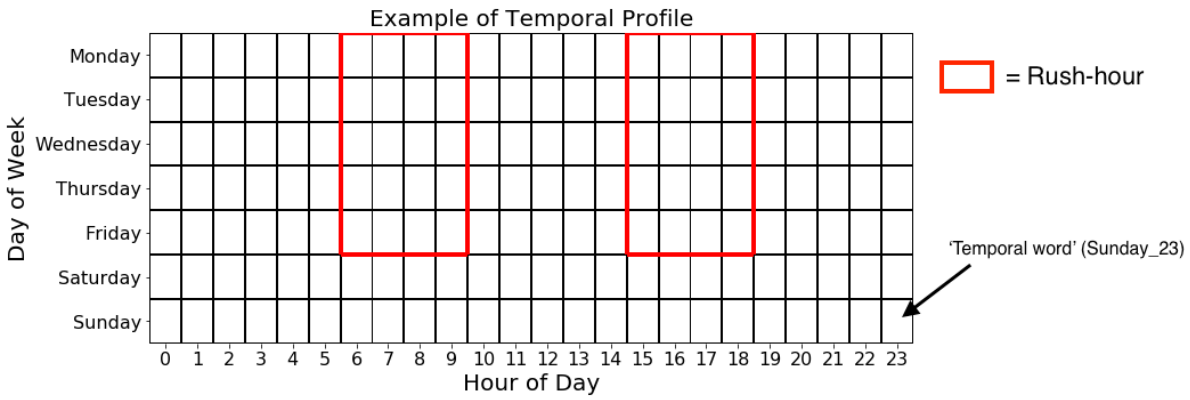

***Figure 3.5** Example of temporal profile*

To carry out this methodology, we use Python's *Genism* and *Natural Language Toolkit* libraries. Metrics used to analyse the accuracy of the LDA (perplexity and coherence), are then used to select the optimum number of topics (TCs) that best fits the data (after Liu & Cheng, 2018).

Finally, we join the results from the clustering back to the data. To do this we look at the characteristics of each of the topics identified by the LDA and assign these to each trip in the

MTL Trajet. For example, if temporal cluster 1 has a high probability to include the temporal words: “Monday_7”, “Tuesday_7”, “Work”, “Education”, trips in the data with these characteristic will be assigned to cluster 1.

3.4 Evaluation of model inputs

3.4.1 Discovery of spatial and temporal dependency in model inputs

To assess the feasibility of modelling each class of trip purpose, this section sets out the methodology used to investigate underlying spatial and temporal inter-dependencies within the individual trip purpose classes (see **Table 3.2**), and within the model inputs used to predict them.

To examine spatial dependency, the start and end coordinates from each trip have first been aggregated into the 7,046 underlying DAs of the study area (see 3.1) using a *Spatial Join* method within *Geopandas*. After this, a Queen’s case contiguity spatial weight matrix has then been computed using Python’s *Pysal* library (Rey & Anselin, 2007). This matrix is used in the calculation of Global and Local Moran’s I statistics, which assess the level of global (across the study area) and local (at the neighbourhood-level) spatial autocorrelation within each class of trip purpose.

Then, Local Indicator of Spatial Association (LISA) maps will be built from the values of Local Moran’s I to visually indicate areas of high spatial association (or ‘hotspots’) and areas of low association (or ‘coldspots’) for each trip-purpose class (Anselin, 1995).

For temporal dependency, the data has been grouped by hour and day of week for each purpose class. These are then plotted in a temporal profile (see **Figure 3.5**) which allow us to capture daily and hourly temporal trends of each purpose (after Arribas-Bel & Tranos, 2017). In examining these calendars we can determine the time-variant and time-invariant properties of given modes of transport and purposes in the trip.

Moreover, temporal stationarity of the frequency of each purpose class is examined across the study period (18th September 2017–18th October 2017) using Augmented Dickey-Fuller (ADF) test statistics. Specifically, these unit-root tests look for statistically significant trends ($p>0.005$; Benjamin *et al.*, 2018) within the temporal structure of the data, thus can be used to determine whether any of the trip purpose become more/less frequent during the study period (Glenn, 2016).

3.4.3 Outlier detection

We first remove any trips where the user had not revealed trip-purpose from the analyses. After this, trip distance and duration (3.2.1) will be used in conjunction with each other to inform the outlier removal process. Trips less than 1 minute and more than 3 hours and less than 50 meters and more 100 kilometres being removed from the analysis. Note, the MTL Trajet has already been cleaned for trips with errors in speed and acceleration (Patterson & Fitzsimmons, 2017b). And, these thresholds in distance and duration have been decided after initial testing and what is likely to skew the data and affect the ability of the classifiers.

3.5 Classification models

3.5.1 Overview

In this study, we evaluate the performance of three distinct machine learning models used to identify hidden relationships in the input features and classify trip purpose:

1. Multi-class Random Forest Classifier
2. Support Vector Machine Classifier
3. Multi-Layer Perceptron Neural Network

A description of each type of model as well as the set-up used are detailed in this section.

3.5.1 Random Forest Classifier

A Random Forest (RF) model is a type of machine learning structure containing a collection (or ensemble) of individual decision trees (Breiman, 2001). Individually, decision trees represent the probability of all possible outcomes of given a set of inputs. In the context of this study, one tree may represent the probability that a trip is for work versus another type of trip based on whether that trip is in rush-hour or in the city of Montreal or not (see Tree-1 in **Figure 3.6**). In a RF classifier structure, shown in **Figure 3.6**, a multitude of similar decision trees are used in combination with each other with each tree having one vote as to which class (i.e. work, leisure, etc.) they expect an input (i.e. user trip) to be part of – the model's prediction is the class with the most votes (Montini *et al.*, 2014). The ability for RF to handle class imbalance have meant they have become the primary tool in trip purpose classification (Gong *et al.*, 2018).

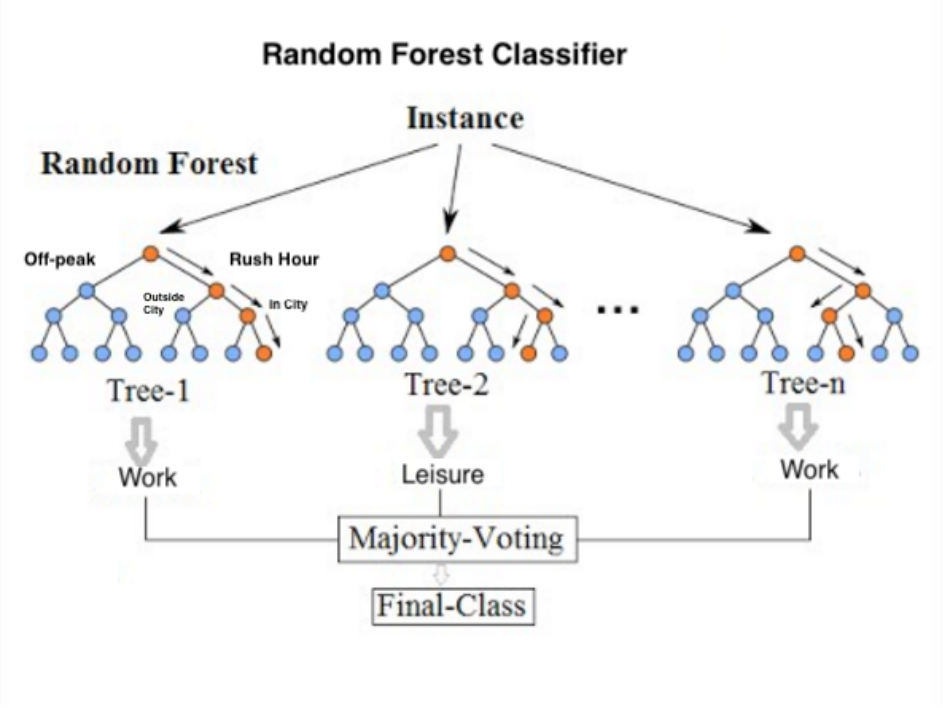

Figure 3.6 *Structure of Random Forest Classifiers (adapted from Koehersen, 2017).*

3.5.2 Support Vector Machine Classifier

Support Vector Machines (SVMs) are kernel-based methods which primarily act to maximise distance between different data classes by drawing a boundary known as a 'separating hyperplane' between them in feature space (James *et al.*, 2013). Features that exist at the minimum distance from the hyperplane are referred to as 'support vectors' and the hyperplane itself is linear in feature space (James *et al.,* 2013). Where non-linear boundaries exist between input data classes, a given kernel function is applied to the data to transform it into higher dimensions where a linear hyperplane can be drawn that better classifies the data: this is known as the 'kernel trick'. The cost function determines the threshold for how accurate the separation of data classes by a hyperplane needs to be. Where cost functions are set too high, a model can suffer from overfitting (Semanjski *et al.*, 2017).

For use in this study, the SVM classifier will undergo hyperparameter tuning to assess various input parameters including cost function, kernel and gamma. Also, we assess two distinct training strategies for the SVM model: a one vs one (i.e. between pairs of trip purposes) and one vs all (each trip purpose vs all other trip purpose classes) approach. Zhu *et al.*, (2014), note SVM as useful methods in trip purpose classification as they can handle high numbers of feature and map non-linear patterns within them.

3.5.3 Multi-Layer Perceptron Classifier

The Multi-Layer Perceptron (MLP) is the final type of classifier used to assess trip purposes in this report. MLPs are a type of feed-forward artificial neural network (ANN) which comprises a framework of connections between an input node layer, a number of hidden node layers and output layer (TDSB, 2016).

In basic ANNs (outlined in **Figure 3.7**), data is sent from the input layer to the hidden layer via links called weighted synapses. These synapses either *activate* or *deactivate* the information passed through them (i.e. determining whether or not a particular feature is passed on to eventually make a prediction of class). Through a number of iterations, these weighted synapses are error adjusted and this changes which information is fed through the model to make predictions in process known as *error back propagation* (TSDB, 2016).

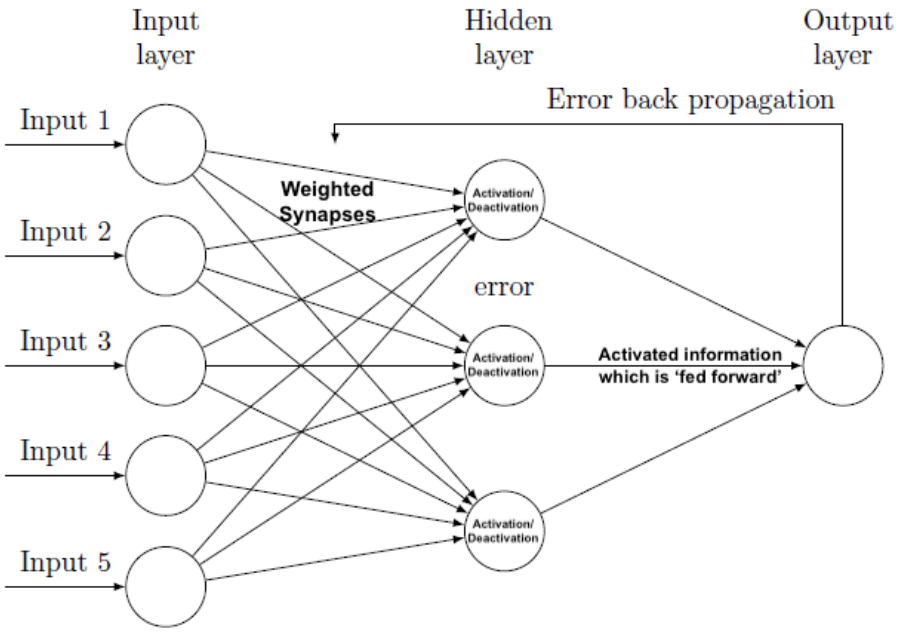

Figure 3.7 *Structure of a generic feed-forward Artificial Neural Network.*

In MLPs there are essentially multiple hidden layers and hence multiple weighted synapses, making them better at predicting non-linear structures than simple ANNs. MLPs can handle binary or multi-class classification, meaning they are useful when tackling trip purpose classification (**Table 2.1**; Xiao *et al.*, 2016). We tune an MLP on a three of different hidden layer structures between 1-3 hidden layers. (1: 100; 2: 50,100 and 3: 50,100,50).

3.5.4 Model training

Each model will be trained using 5-fold cross-validation and a training/test split of 67/33. After sections 3.2-3.3, a total of 17 input variables are used as inputs in theses classification models and these are detailed in **Table 3.5**.

Table 3.5 *Description of the key all model inputs used within the trip purpose classifiers.*

Column	*Description*	*Data Type*
	Dependent variable	
Trip Purpose	The class of activity for which that trip is for	*Categorical*
	General Explanatory variables	
Trip Distance	Total distance of trip (m)	Ratio
Trip Duration	Seconds between start and end of a trip	Ratio
Trip Mode	The means of transport used for a trip	Categorical
Precipitation	Mean precipitation (mm) at the time of the trip	Ratio
Temperature	Mean temperature (°C) at the time of the trip	Ratio
	Spatial explanatory variables	
Start in City	Trip starts in the City of Montreal?	Binary
End in City	Trip ends in the City of Montreal?	Binary
Cardinal direction	Mean cardinal direction of trip	Categorical
Magnitude	Magnitude of direction of trip	Ratio
Land Use Start	Underlying land use of where the trip started	Categorical
Land Use End	Underlying land use of where the trip ended	Categorical
Spatial Cluster	k-means cluster label*	Categorical
	Temporal explanatory variables	
Weekday	Trip starts on a weekday?	Binary
Start in Rush Hour	Trip starts in Rush-hour?	Binary
Through Rush Hour	Trip passes through Rush-hour	Binary
End in Rush Hour	Trip ends in rush-hour?	Binary
Temporal Cluster	LDA temporal cluster label*	Categorical

* (see 3.3.3)

To assess class imbalance present within the MTL trajet data, we evaluate the classification on a random minority over-sampling and random majority under-sampling technique using *imbalanced-learn* library in Python (after Xiao *et al.*, 2016). We use this technique to improve the representative of minority trip purpose classes in the training of the MLP and SVM models to improve classification accuracy (Japkowicz, 2000; Buda *et al.*, 2018).

The specific combination of inputs used in each model will be determined by running an initial Random Forest Classification model on the data and examining feature importance.

Arbitrarily, feature selection in the models detailed in *3.4.1–3.4.3* will be based on those input variables with a feature importance score of 0.05 and above.

For the MLP and SVM model, *Scikit-Learn* is used to apply a One-Hot Encoder to all categorical model inputs (listed in **Table 3.5**). Also, to speed up the training process and effectiveness of the models, all ratio values included in the model inputs are standardised between 0–1 (after Xiao *et al.*, 2016). Both encoding and standardisation methods are adopted after non-normality was discovered in initial examination of the model inputs.

3.6 Limitations:

3.6.1 Methodological

A number of methodological limitations are noted from the methods used in this study. Firstly, the spatial and temporal units used throughout this analysis are chosen relatively arbitrarily and are as such subject to both the Modifiable Areal Unit Problem (Openshaw, 1984) and Modifiable Temporal Unit Problem (MTUP; Cheng & Adepeju, 2014). Specifically, we separate defined sections of both space (i.e. City/Non-City) and time (i.e. Rush-Hour), so any conclusion drawn are only relevant at these sections/scales (an example of the '*Ecological Fallacy*'; Openshaw, 1984).

Further, we discount any spatial or temporal '*edge effect*' that units just outside of these boundaries may have on the dynamics of the units within them. For example, in the temporal clustering (3.3.3), choosing temporal words at a temporal unit of hours per week is problematic as it means we ignore the dynamics that exist within each unit (i.e. the

dynamics that occur at less than hour) and across the study period (i.e. the dynamics that occur across the month). Indeed, we may observe different patterns with the temporal clusters if we chose different temporal resolutions (Zhao *et al.*, 2019).

For the modelling procedure we make no attempt to train and test the classifiers in any sequence and are instead split using a random sampling technique. Arguably, there may be temporal structure within the data which would mean training and testing the models on trips from different times of the day and week become ineffective (Gong *et al.*, 2018). There are significant computational restrictions with the processing of these analyses of this project, therefore we only fine tune the hyperparameters SVM and MLP models to a small degree.

Finally, no representation of space-time interdependencies are considered for the analysis. We cannot, therefore, be confident in assuming that the any temporal trends happen uniformly across the study area or uniform spatial trends through time (Ren *et al.*, 2019).

3.6.2 Data

A few limitations exist within the quality of the various datasets used in project. Firstly, the MTL Trajet data used in the classification is limited as it only applies to study period and only to Montreal (Patterson & Fitzsimmons, 2017a). Notably, it represents 185,285 trips from 4,425 so it is not particularly representative of the region of Montreal – an area containing around 3.8 million people (Chevalier, 2018).

Chapter 4. Results

This chapter is divided into three sections, the first (4.1) examines general trends in the model inputs and identifies key areas of analysis, the second (4.2) reviews the space, time and structures and interdependencies within the model inputs before the third analyses the performance of the different classification models and their outputs (4.3).

4.1 Overview of model inputs

4.1.1 Trip distance and duration

After calculating the distances and duration of the individual trips of the MTL Trajet survey, the analysis finds a total of 7,594 trips which are removed from the analysis based on the outlier strategy adopted in 3.4.3. As shown in **Table 4.1**, the majority (6,709) of these were from trips that were less than 50 m in length. These trips are potentially from cases where the app had started recording a GPS trace after witnessing slight movement or the user had mistakenly ended a trip while being stopped for more than 2 minutes (Patterson & Fitzsimmons, 2017b).

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

Outlier Type	*Number removed*
Distance below 50 m	6,709
Distance above 100 km	62
Duration below 60 seconds	412
Duration above 3 hours	411
Total	7,594

The resulting trips are shown to have a mean distance and duration of around 6.6 km and 26 mins, respectively (**Table 4.2**). Here, we see that both these variables are positively skewed, although distance is more so. The disparity between the mean and median in both trip distance and duration is indicative of both these variables exhibiting a long-tailed distributions, and this can be visually identified by examining the univariate kernel density estimations and Quantile-Quantile plots shown in **Figure 4.1**.

Table 4.2 *Summary statistics of distance and duration of trips from the 2017 MTL Trajet travel survey (converted to km and minutes; N=177,938).*

	mean	*STD*	*min*	*25%*	*Median*	*75%*	*95%*	*max*	*kurtosis*	*Skewness*
Distance (km)	6.63	9.92	0.05	0.84	3.14	8.09	25.25	99.81	15.216	3.355
Duration (min)	25.62	21.42	1.00	10.27	20.07	34.68	65.81	179.98	6.097	1.967

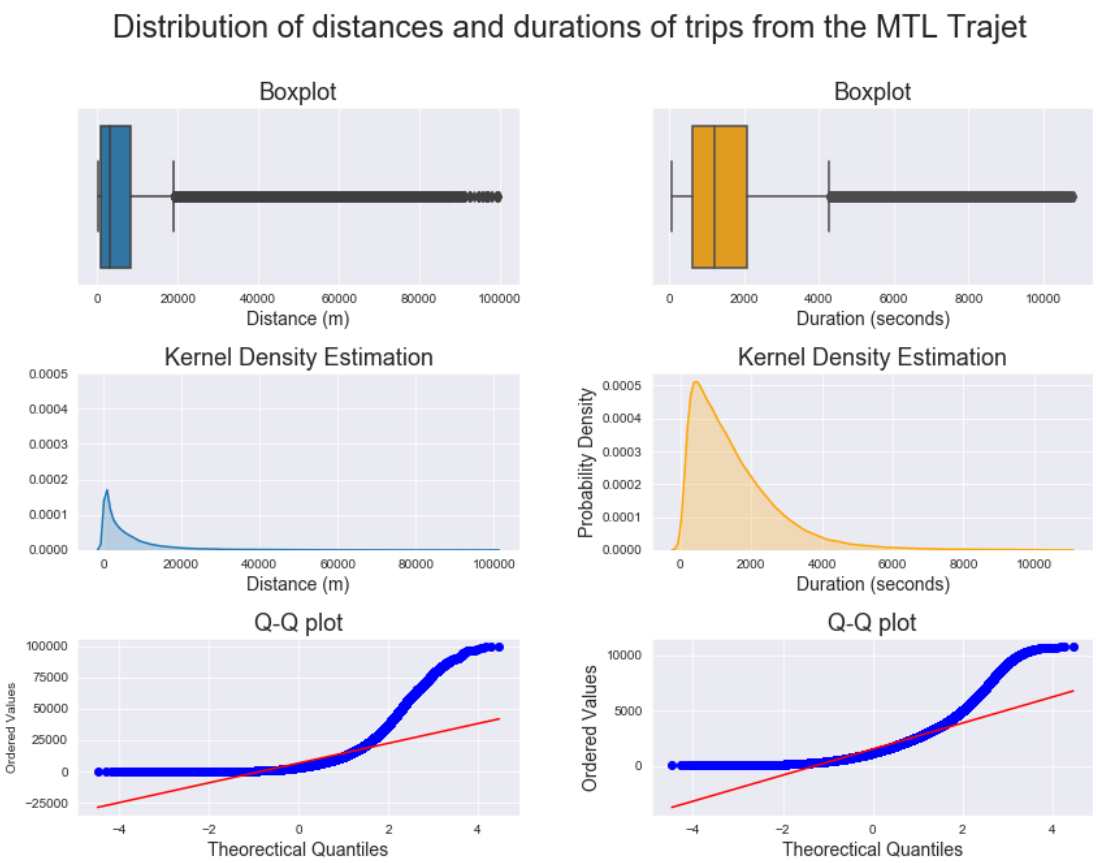

Figure 4.1 *Boxplots (top), Kernel Density Estimation (middle) and Quantile-Quantile (bottom) plots showing the distribution of distance and duration of trips from the 2017 MTL Trajet travel survey.*

4.1.2 Travel purpose and mode

There are total of 73,029 trips from the MTL Trajet survey containing both a travel mode and purpose label. As shown in **Figure 4.2**, the categories of these variables have not been selected by the respondents in equal proportions. It is shown that severe class imbalance exists within both of these categories, with around 63.7% (45,769) of the trips labelled as either trips to work and back to home, and 33.6% of the trips being taken by car. This finding is not unexpected for a survey that has taken place in North American city with a

high level of employment, however, so we can argue that this study is relatively representative of trips occurring across Montréal (Meng *et al.*, 2019).

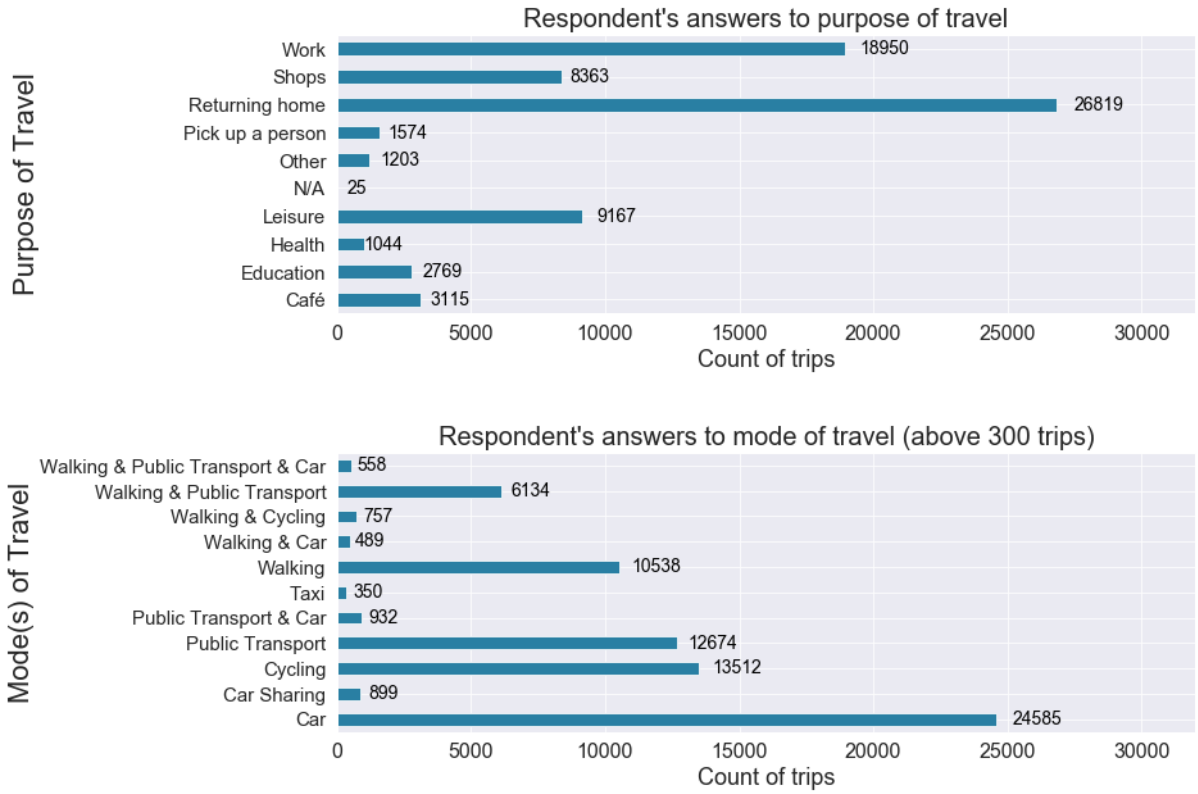

Figure 4.2 *Bar charts showing the type of trip purpose and travel mode selected by respondents to the 2017 MTL Trajet survey.*

When the travel mode is broken down by purpose, in **Figure 4.3**, we see that there is higher usage of cars in trips for shopping and picking people up and lower usage of cars in trips for education. Notably, a higher proportion of respondent have walked or cycled when taking trips to work, cafés and places of education.

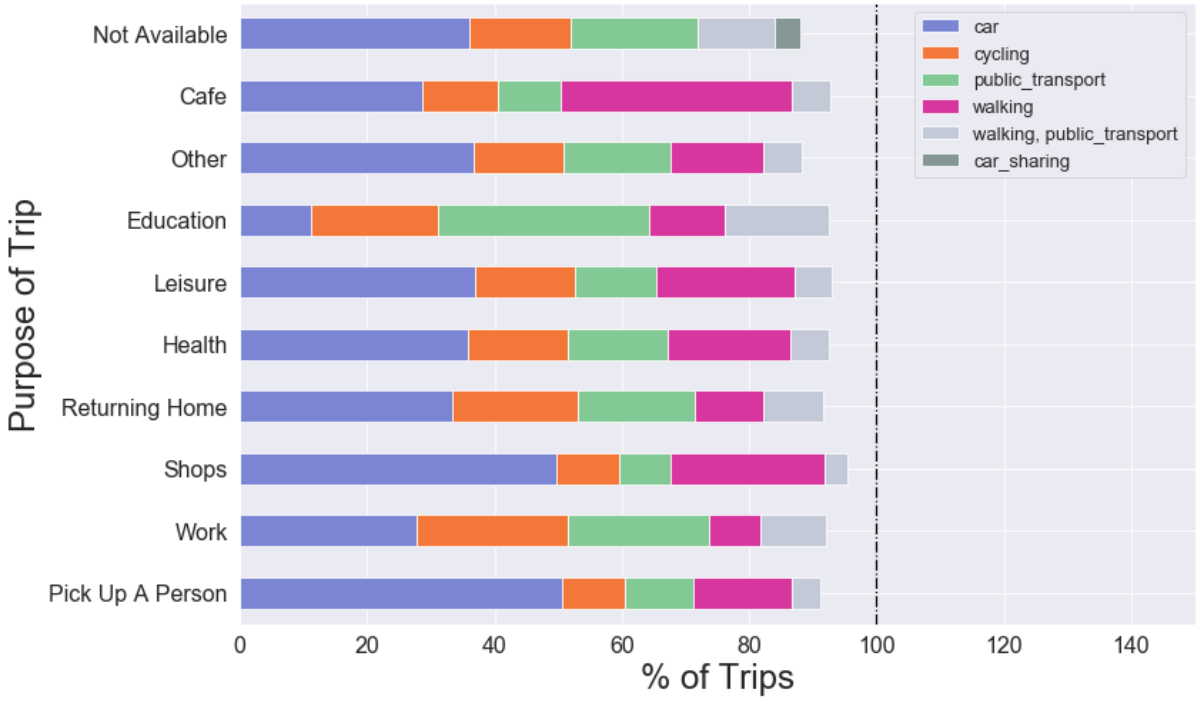

Figure 4.3 *Bar chart comparing the proportion of each unique trip purposes accounted for by each unique travel modes.*

Comparing the distances and duration of the trips as grouped by trip purpose, in **Table 4.3**, we see that trips to cafés and shops are shorter in both mean distance (4.5 & 4.8 km) and duration (23 & 20 mins) compared to the other forms of trips such as work and returning home. When cross-referencing with travel mode (from **Figure 4.3**), it could be proposed that these values are a product of the fact that a higher proportion of these trips are walked.

Table 4.3 *Summary statistics of trip distance and duration per trip purpose (Note: trips that are classed as 'Not Available' have been omitted from this table).*

Trip Purpose	*Trip distance (km)*			*Trip duration (mins)*		
	μ	σ	Skew	μ	σ	Skew
Café	4.5	7.6	+4.5	22.9	19.5	+2.4
Education	5.8	6.8	+3.5	28.9	19.4	+1.3
Leisure	6.8	10.3	+3.1	35.2	20.8	+2.0
Health	6.2	8.0	+2.8	25.2	19.7	+2.0
Other	8.7	12.9	+3.1	31.3	25.3	+1.9
Returning home	7.5	9.8	+3.2	29.0	22.2	+1.7
Pick up a person	7.8	10.9	+2.9	25.1	20.9	+2.1
Shops	4.8	7.2	+3.5	20.4	17.6	+2.4
Work	7.6	8.3	+2.6	28.8	19.8	+1.5

4.1.3 Trip direction:

The mean direction of all trips taken across all 91 regions of Greater Montreal is shown in **Figure 4.4**. Here, we see that the direction is general towards the city of Montreal (see **Figure 3.2**) indicating we can be somewhat confident in assuming we have accounted for a some degree of the MAUP – with the directional dynamics of trips facing 'inward' towards downtown and the study area chosen versus out of the study area (Openshaw, 1984).

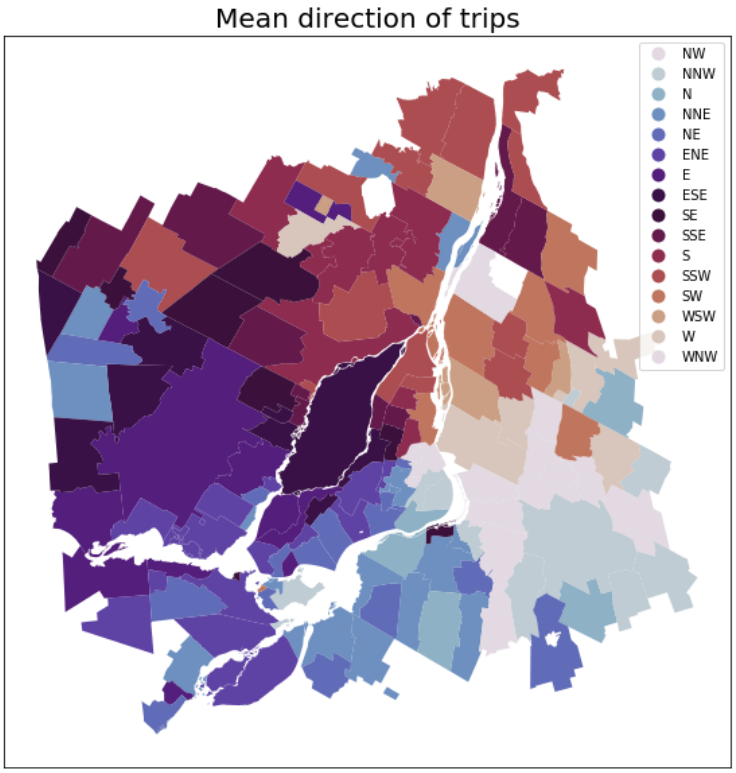

Figure 4.4 *Map showing the mean direction of trip within each region of Greater Montreal.*

Across the individual purpose class, in **Figure 4.5,** the mean direction of the trips are generally shown to be in the NNE and SSW directions, something which is similar to the morphology of the island of Montreal. Notably, work and returning home trips shown to be more directionally dependent, in the SSW & NNE respectively, than the other purpose classes. Directional independence is shown in the trips categorized by other, café and purposes.

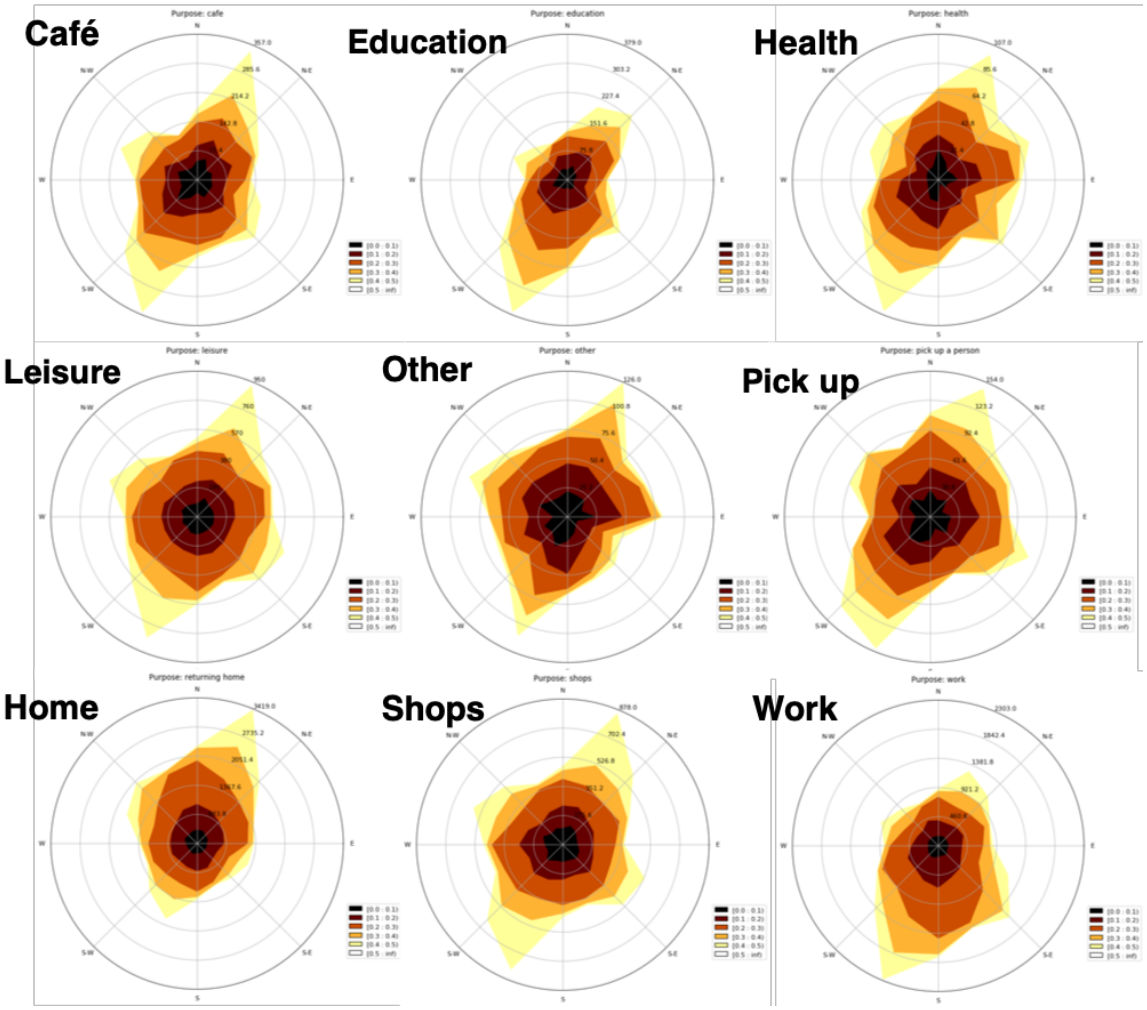

Figure 4.5 *Circular contour plot showing the mean direction of trips for each trip purpose.*

4.1.4 Rush-hour & City Labels

After applying city and rush hour labels to the origin and destination points (see 3.3.1), the majority of trips are found to have occurred within the City of Montreal (93.5%) and are evenly split between rush hour and off-peak (**Table 4.4**).

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

	Rush hour?		City?	
	Yes	No	Yes	No
Origin of trip	36785	36244	63811	9218
Destination of trip*	38650	34490	64136	8893

* including trips that have passed through rush hour or city

When separated by purpose class, a higher proportion of trips are discovered to be carried out for work and education at rush-hour times versus trips to shops which are proportionally carried out at off-peak times (**Figure 4.6**). **Figure 4.7**, highlights that work and home-bound trips are disproportionately represented in trips occurring in the city as opposed to outside the city, where shopping and leisure trips are more proportionally represented.

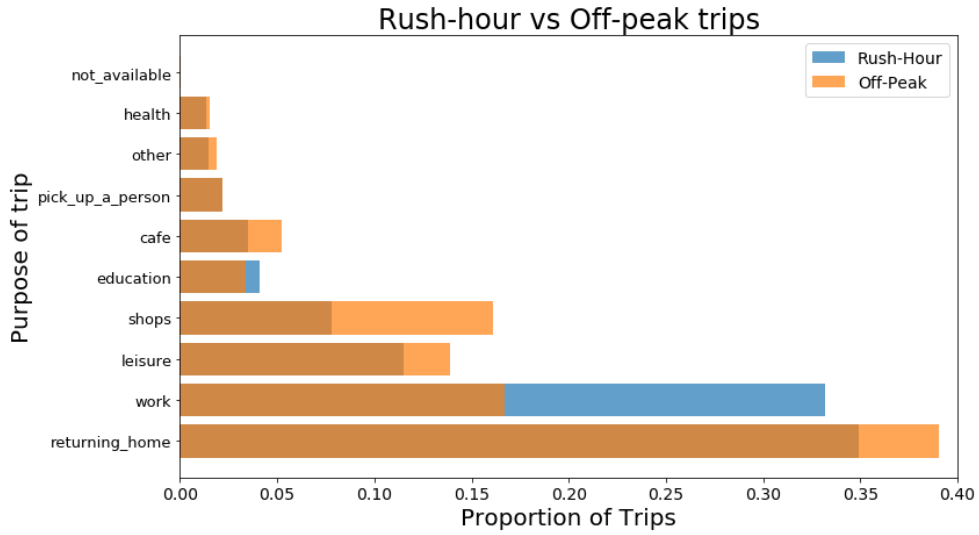

***Figure 4.6** Bar chart showing the proportion of trips carried out during rush-hour and off-peak as grouped by purpose.*

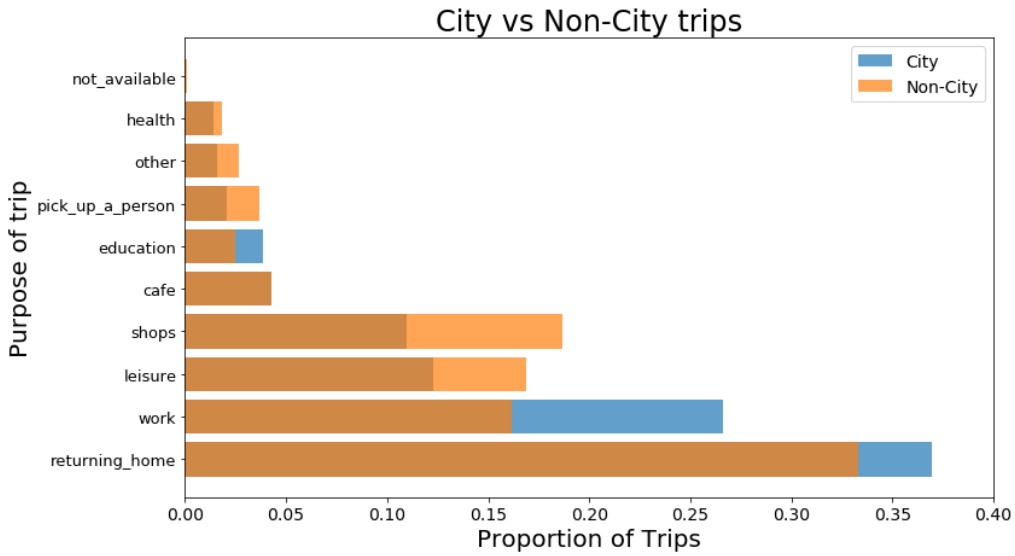

Figure 4.7 *Bar chart showing the proportion of trips carried out within and outside the City of Montreal as grouped by purpose.*

4.1.5 Land Use

The majority of trips are found to have their origins and destinations in areas of the City of Montreal classified as residential and mixed use categories (Ville de Montreal, 2014; **Figure 4.8**). Note that, most trips are found to begin in areas of mixed and residential land use (around 65-75% of trips), whereas most trip end in a variety of land use categories dependent on trip purpose (see **Figure 4.9**).

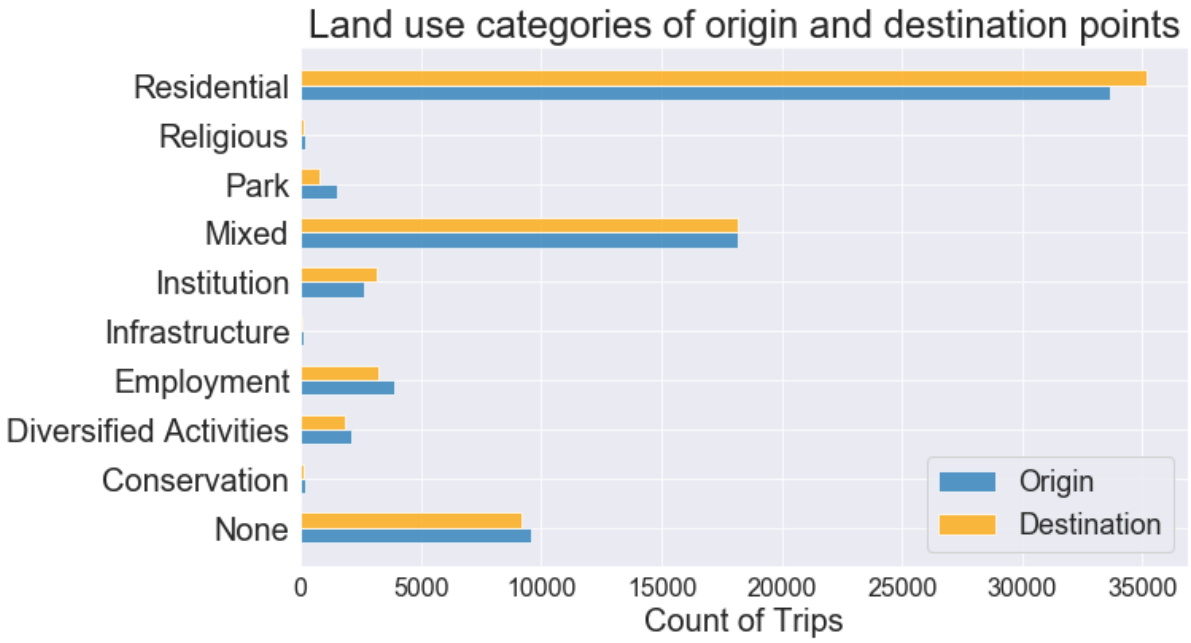

Figure 4.8 *Bar chart showing number of trips that have their origins or destinations in each land use category (as defined by Ville de Montreal, 2014).*

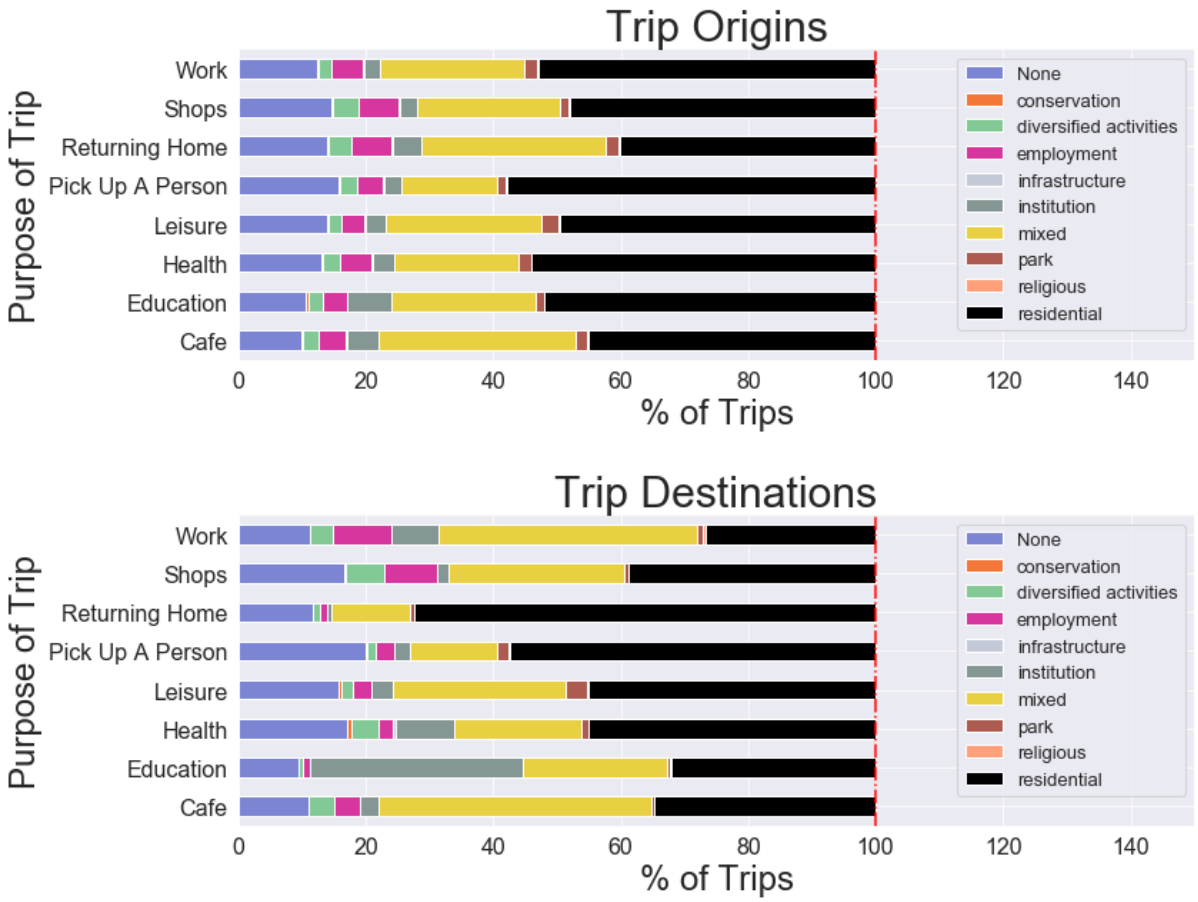

Figure 4.9 *Bar charts comparing the proportion of each unique trip purposes accounted for by each unique land use category (as defined by Ville de Montreal, 2014) in the trip origins and destinations.*

4.1.6 Clustering

4.1.6.1 Spatial

After fine-tuning values of *k* between 2-20 and evaluating the sum of squared distances and silhouette score within each k-number of clusters (**Figure 4.10**), we select a total of 12 for the k-means clustering algorithm to be built upon. These clusters mapped across the study region in **Figure 4.11** and a summary of how many trips have been assigned to each cluster

is shown in **Figure 4.12**. Note that, the algorithm separates a region containing Downtown Montreal (*cluster-id=0*) and the clusters of 0,5 and 10 are most prominent in the data.

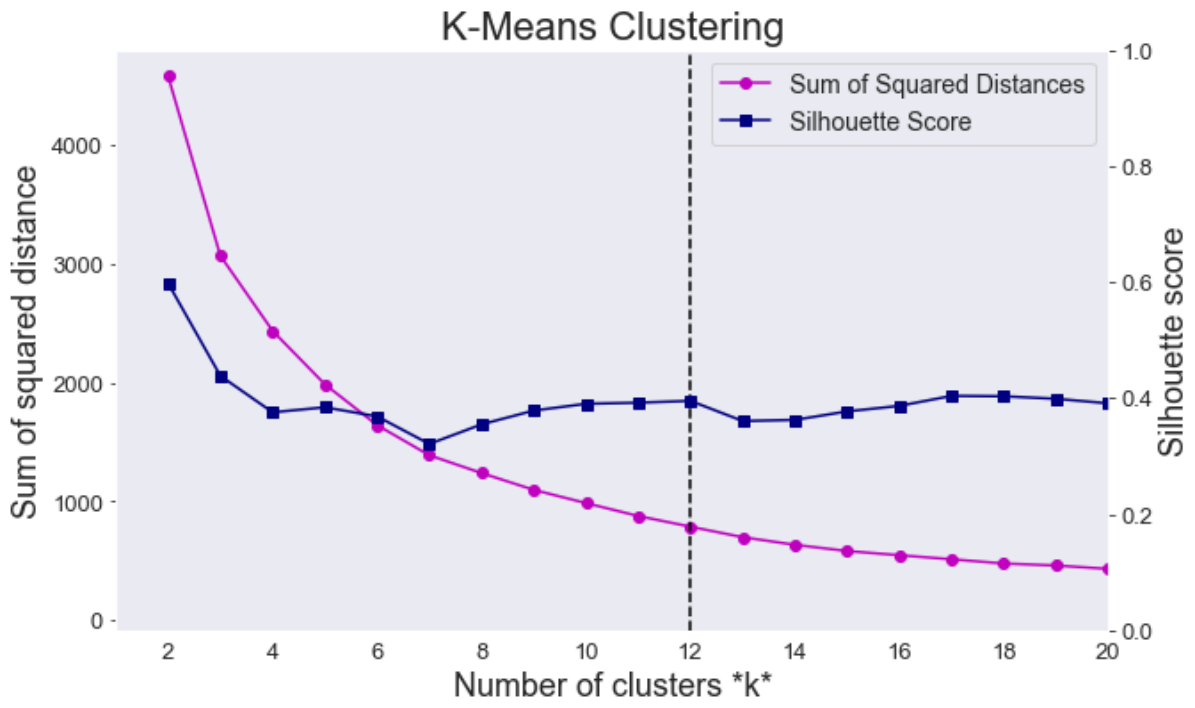

Figure 4.10 *Line graph comparing sum of squared distances and silhouette scores of k-means clustering algorithm for k between 2-20.*

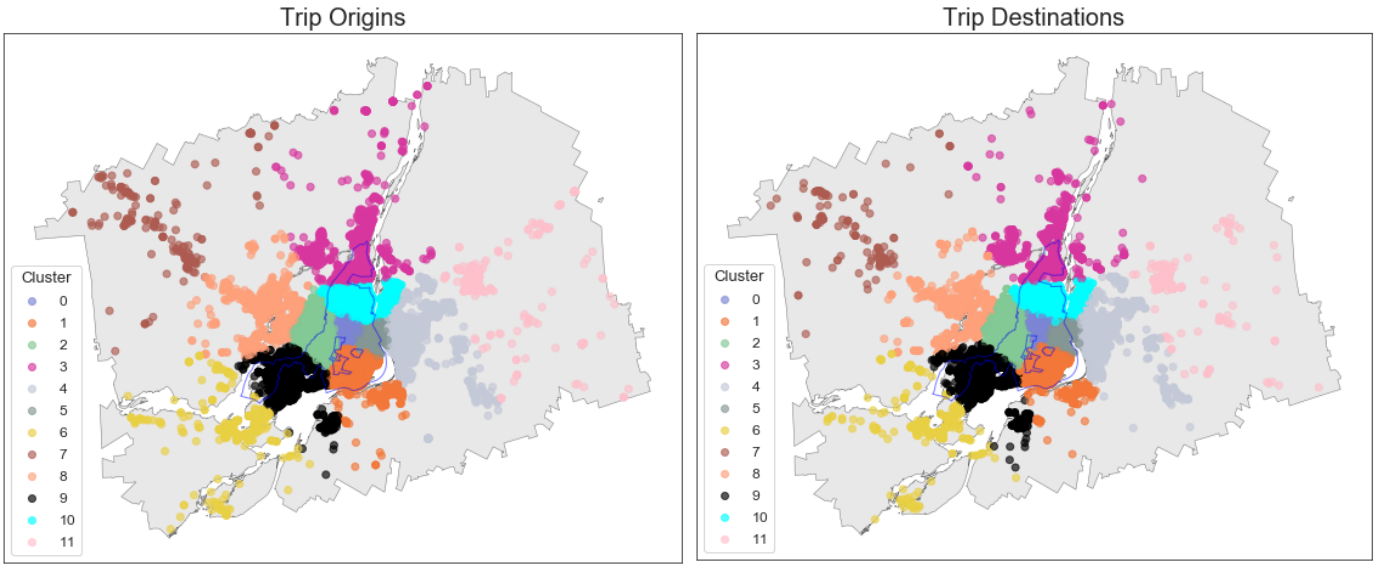

Figure 4.11 *Map of origin and destination points from the MTL Trajet trips coloured by cluster label across the study region.*

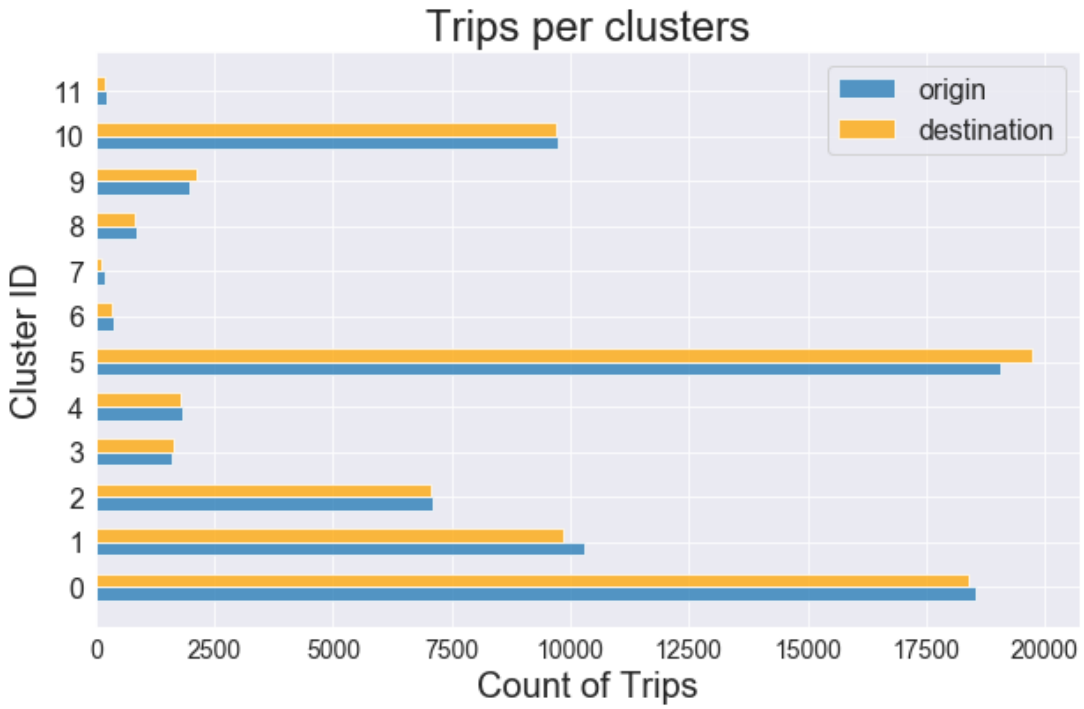

Figure 4.12 *Bar chart showing number of trips per spatial cluster identified by the k-mean clustering algorithm.*

4.1.6.1 Temporal

After examining coherence and log perplexity of LDA models modelled between 1-12 topics (here, temporal clusters or TCs), we select 5 topics to build from the data. As shown in **Figure 4.13**, at 5 topics we maximise the coherence of the LDA model, suggesting that the words within the topics are most similar at this point (Kumar, 2018). Notably, perplexity of model continues to drop pass this point, indicating that the model is better at predicting the topics, however in practice it has been found that coherence is a more stable metric for an LDA (Kumar, 2018).

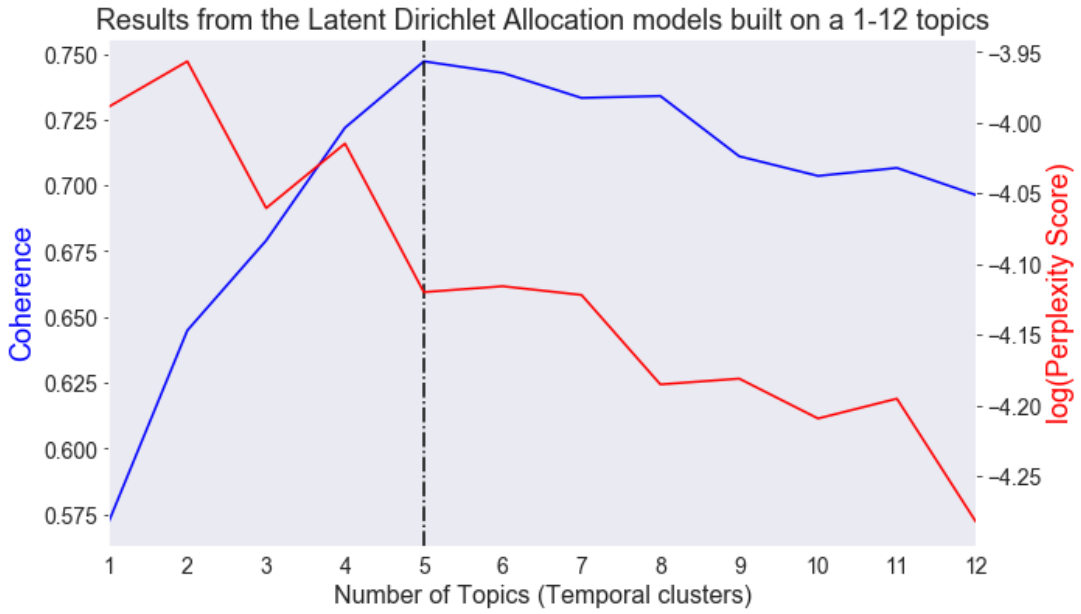

***Figure 4.13** Line graph comparing coherence score and log perplexity of LDA models using a topic count of between 1-12.*

Results showing the temporal profile of the 5-topic LDA model are shown in the form of calendar plots in **Figure 4.14**. Note that, each 'temporal profile' is a 7-day by 24-hour matrix, which indicates the probability that a given 'temporal word' (i.e. "Sunday_7") is associated

with that temporal cluster or TC (which known as its *weighted importance*). The trip purposes classes associated with each TC along with their weighted importance are outlined in **Table 4.5**.

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

Temporal cluster (TC)	*Associated trip purpose*	*Weighted importance*
1	Returning Home	0.549
2	Leisure	0.349
	Education	0.111
	Other	0.041
3	Work	0.488
4	Work	0.276
	Not Available	0.002
5	Shop	0.306
	Café	0.113
	Pick Up a Person	0.060
	Health	0.004

The strongest association found between trip purpose classes and the TCs is found with the returning home class (with a probability of 0.549 to be found in cluster 1). On examining the temporal profile of TC1, 14 out of the 20 evening rush hour segments have been associated. **Table 4.5**, also indicates the relatively weak temporal clustering of Leisure & Education in TC2 and Shop, Café & Pick up a Person trips in TC5 suggesting that these purposes broadly share similar temporal characteristics within the data.

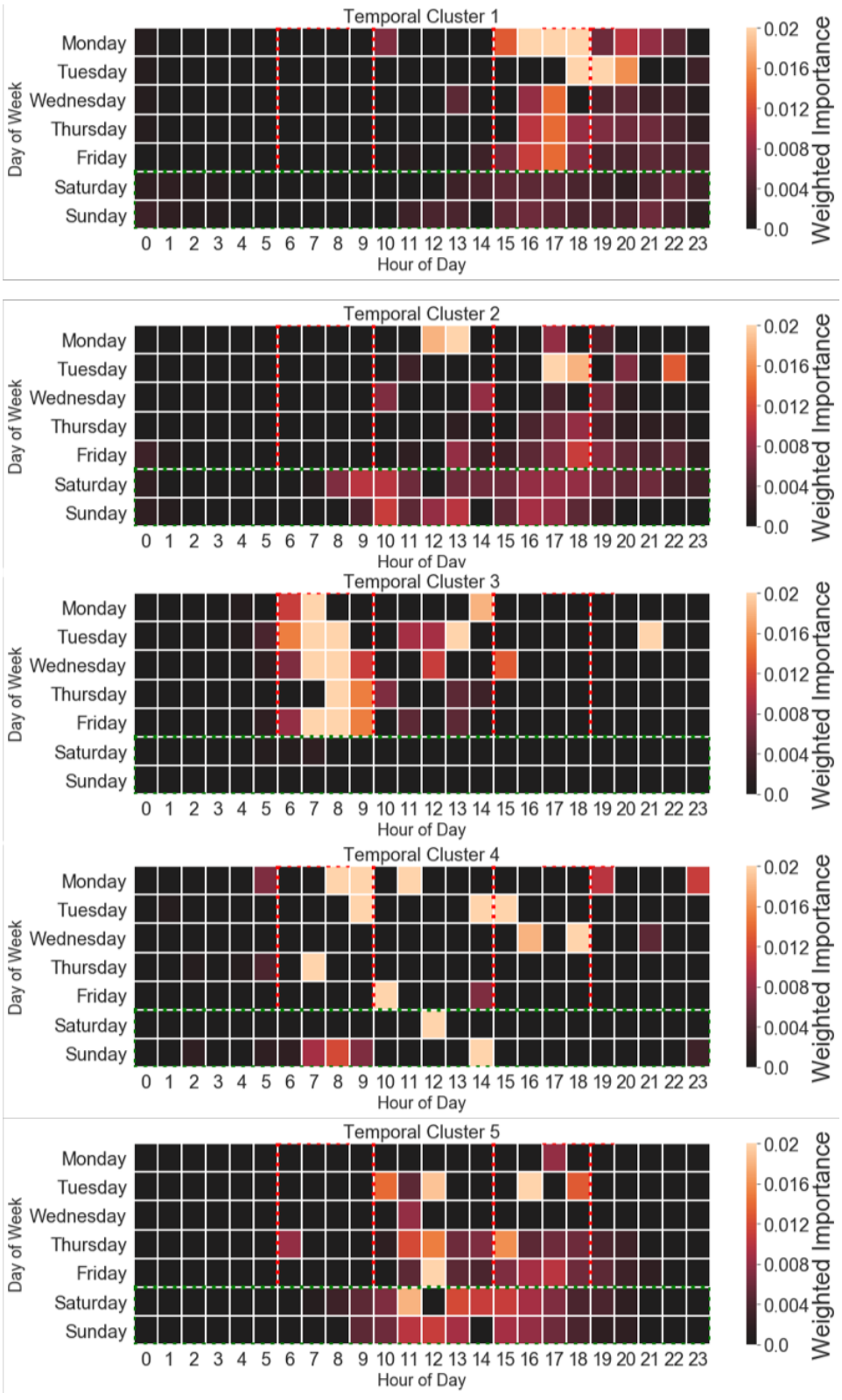

Figure 4.14 *Calendar plot showing the weighted importance of each 'temporal word' in each of the 5 temporal clusters (rush hour periods as defined by this study are outlined in* ***red*** *and weekends are outlined in* ***green****).*

Work is found to be the only trip purpose existing in more than one TC as it is found to be the most dominant topic in both TC3+TC4. On examining the inter-topic distance map (**Figure 4.15**), we see the LDA model has been fairly successful in separating the first two principle components (PC1+PC2) of the clusters for all topics expect TC3+TC4 where there is a slight overlap. After joining the TCs back to the data (using step set out in 3.3.3), we find that TC2, TC4 and TC5 are most frequent in the data (**Figure 4.16**).

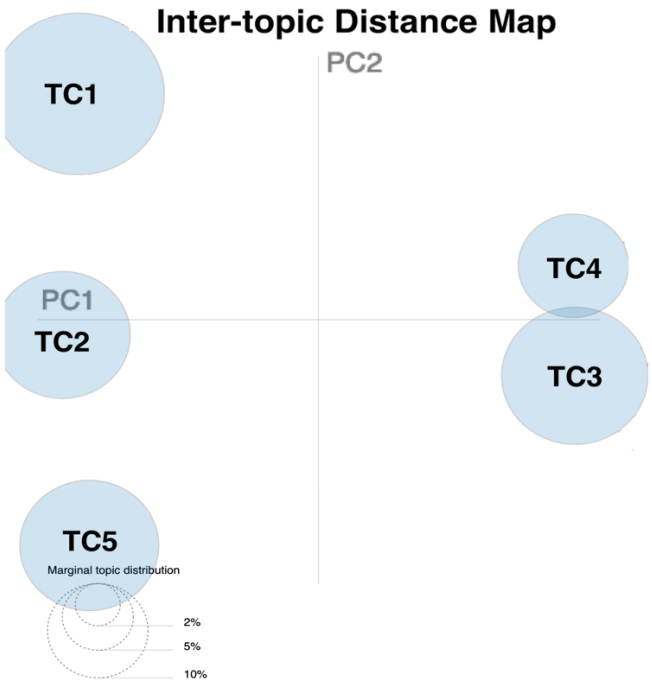

Figure 4.15 *Inter-topic Distance map between each of the Five Temporal Clusters identified by the LDA model*

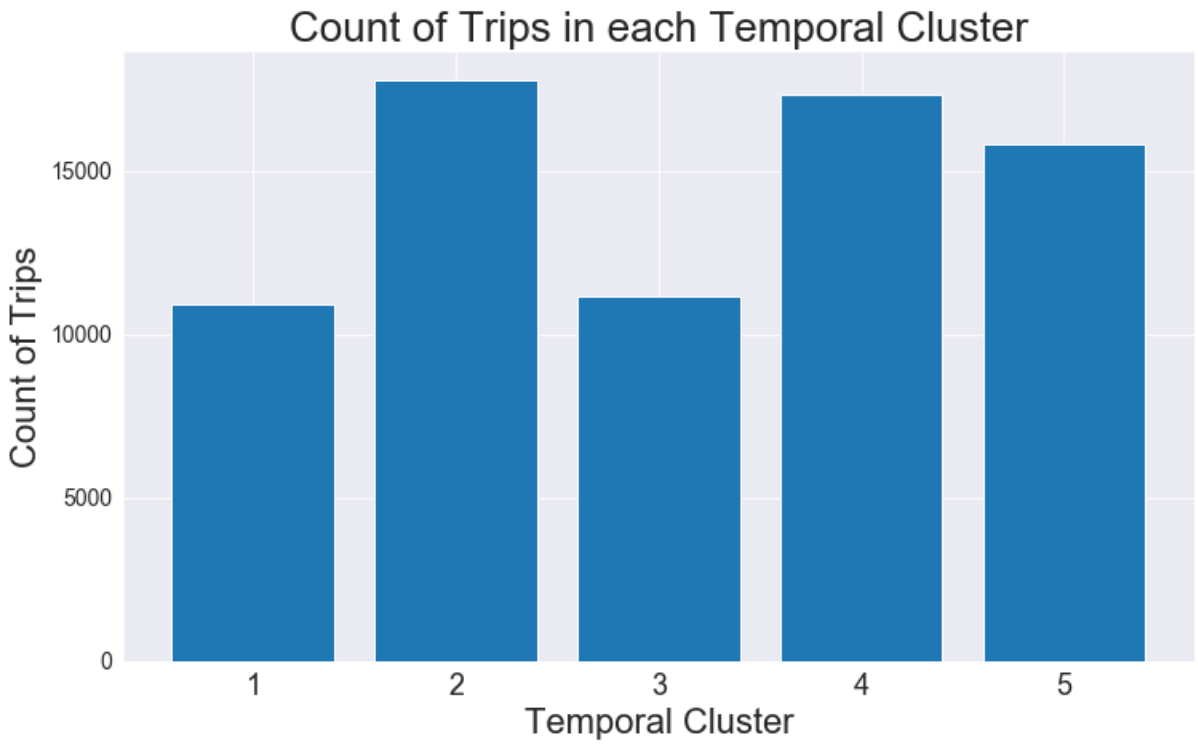

Figure 4.16 *Count of trips associated with each temporal cluster identified for this analysis.*

4.2 Spatial and temporal dependency in model inputs

This section highlights the methods carried out to investigate time, space and space-time signatures in the data. It is hoped that the identification of these forms of trends will assess the ability for the purposes to be modelled inform the modelling process (detailed in 4.3).

4.2.1 Temporal trends

A total of 73,029 trips were recorded across the study period (18th September 2017– 17th October 2017), but there is significant variation in the amount of recorded trips per day. As shown in **Figure 4.17**, during the first 7 days of the study less than around 1000 trips were recorded per day compared with more than 1500 trips in the remaining days (with the most amount of trips being recorded on Thursdays/Fridays). Here, less trips are recorded on

weekends versus weekdays, other than Monday 9th October, which was the day Thanksgiving was celebrated that year in Canada.

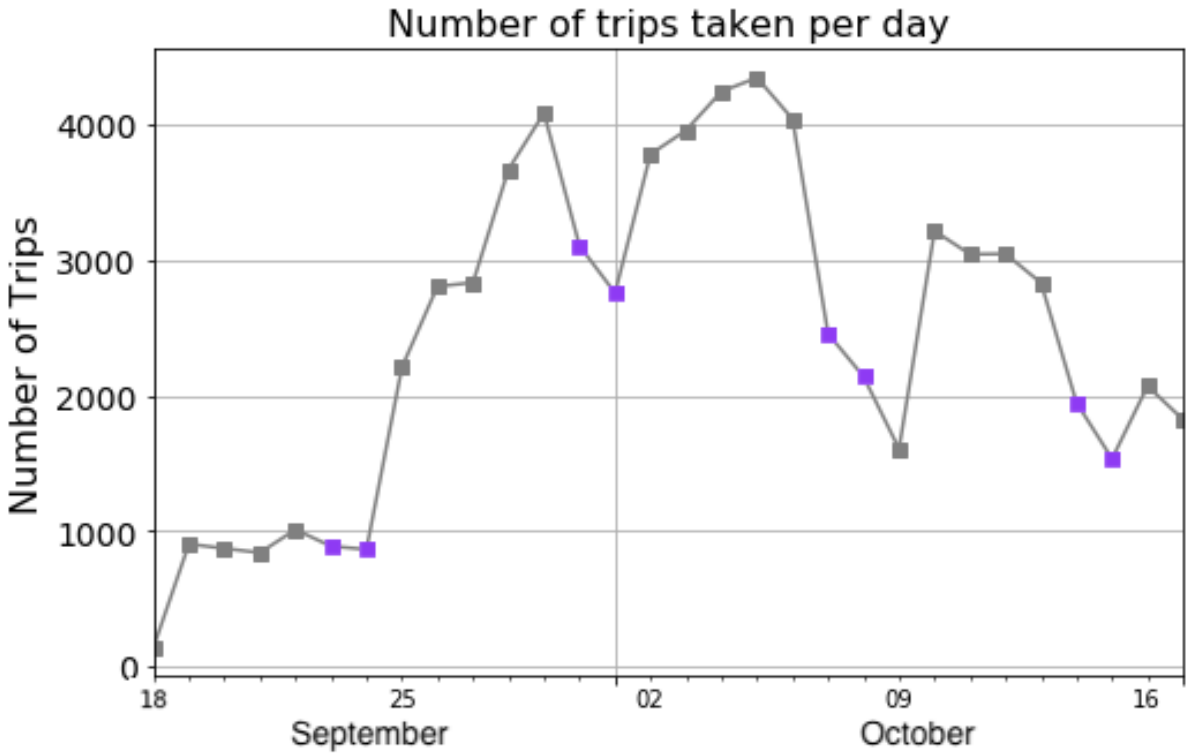

Figure 4.17 *Line plot showing the amount of recorded trips taken from the MTL Trajet app between 18th September 2019– 18th October 2017 (weekends indicated in* ***purple****).*

As broken down by week, on average, trips of longer distances are taken on the weekends versus weekday (**Figure 4.18**). Arguably this could result from the influence of work, with people travelling further into rural areas during weekends. Notably, there is no deviation from the mean travel duration across the week.

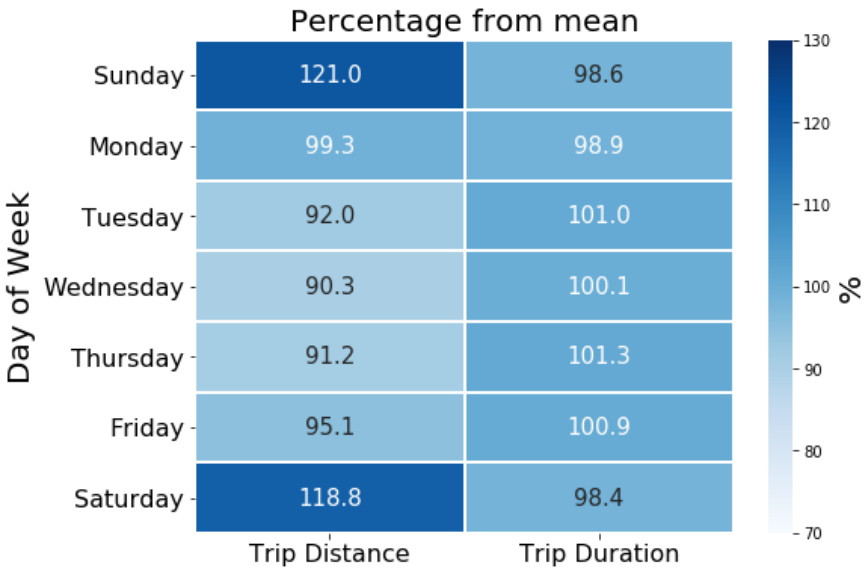

Figure 4.18 *Average trip distance and duration represented as a percentage of the mean.*

The temporal profile of each trip purpose class as an average per hour per day across the study period is shown in **Figure 4.19**. Here we see a clear temporal dependency in work and education and returning home trips. They are shown to be restricted to rush hour periods and the week days, something which is expected for these types of trip purposes. Moreover, trips for health are shown to be less temporally dependent throughout the week, which is expected as people often do not decide when they make hospital visits. Leisure and shopping trips are found to be show a tendency to occur on weekends and after-work hours in these profiles.

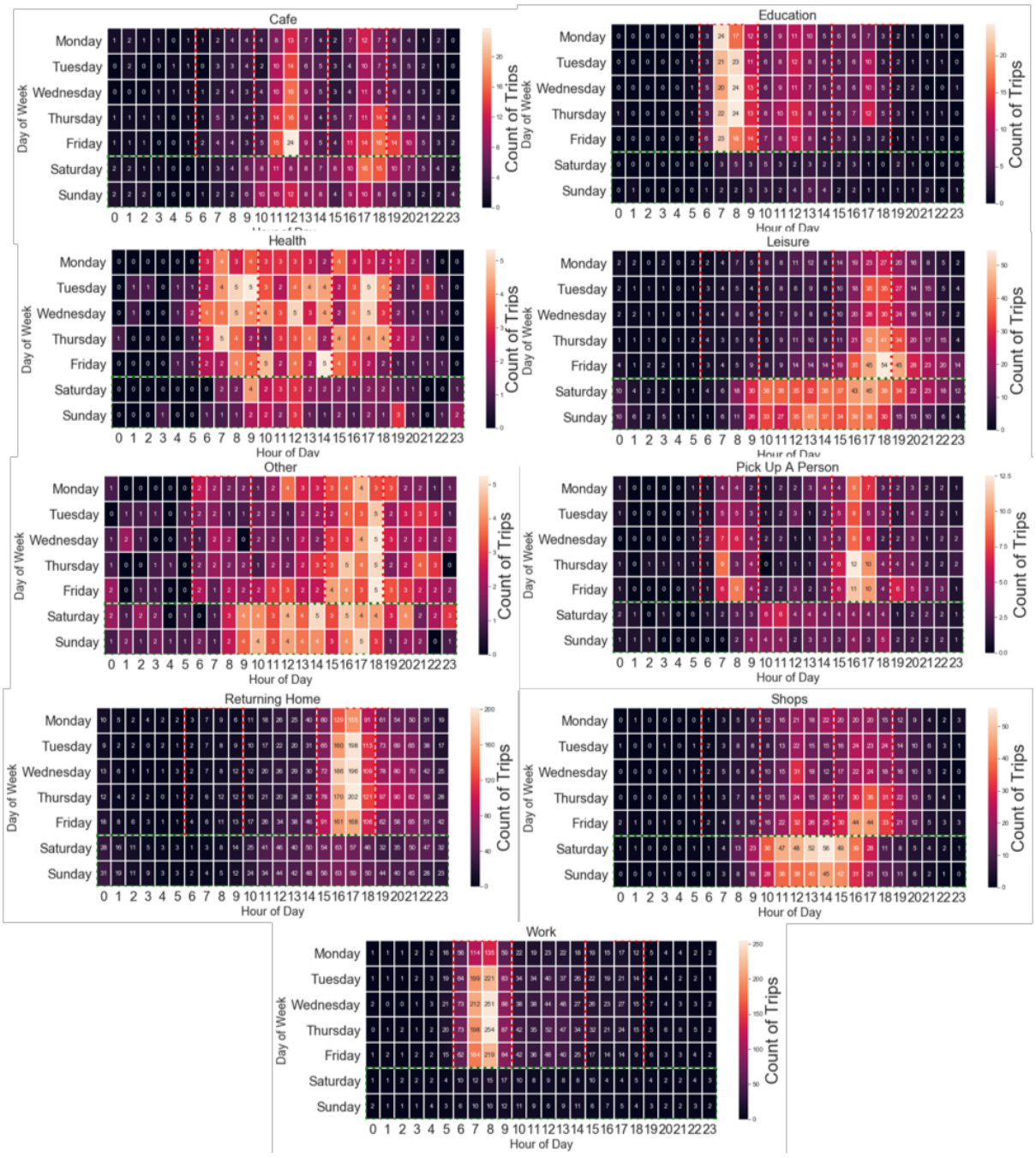

Figure 4.19 *Calendar plot showing the temporal profile for each trip purpose class of the count of trips recorded per hour as average per day of the week.*

A temporal deviation in weather patterns is found during the study period. As shown in **Figure 4.20**, the temperature is shown to generally decrease with more rain falling towards the latter half the study. We expect that change in weather to have had an effect on the

modes of transport chosen by the survey participants (Gong *et al.*, 2018), and thus may hinder the ability of our classification models to generalise (Xie *et al.*, 2016). Correspondingly, we find moderately strong and statistically significant (p-value>0.005) negative correlation between cycling (-0.36ρ) and walking (-0.44ρ) usage and temperature, evaluated using a spearman's rank correlation co-efficient.

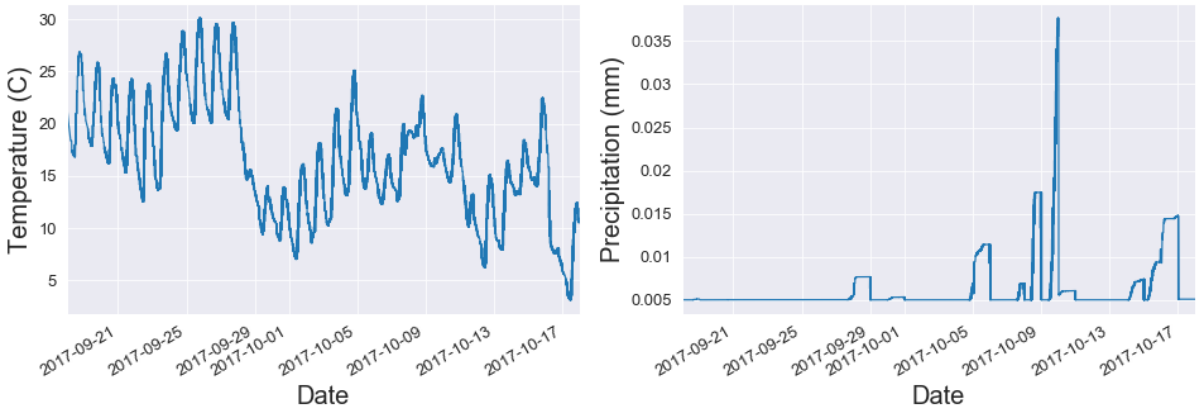

Figure 4.20 *Time series plot showing the average temperature (in Celsius) and precipitation (mm) recorded during the study period.*

Finally, we assess the degree of temporal stationarity in the data. A clear diurnal pattern can be identified from the temporal decomposition of the MTL Trajet trips at 24 hour lags, and it appears that more trips have been recorded later in the study period, suggesting a temporal non-stationarity (see **Figure 4.21**). This assumption of non-stationarity is proven to be statistically significant (p<0.005) by ADF tests presented in **Table 4.6** in 8 out 9 of the trip purpose classes.

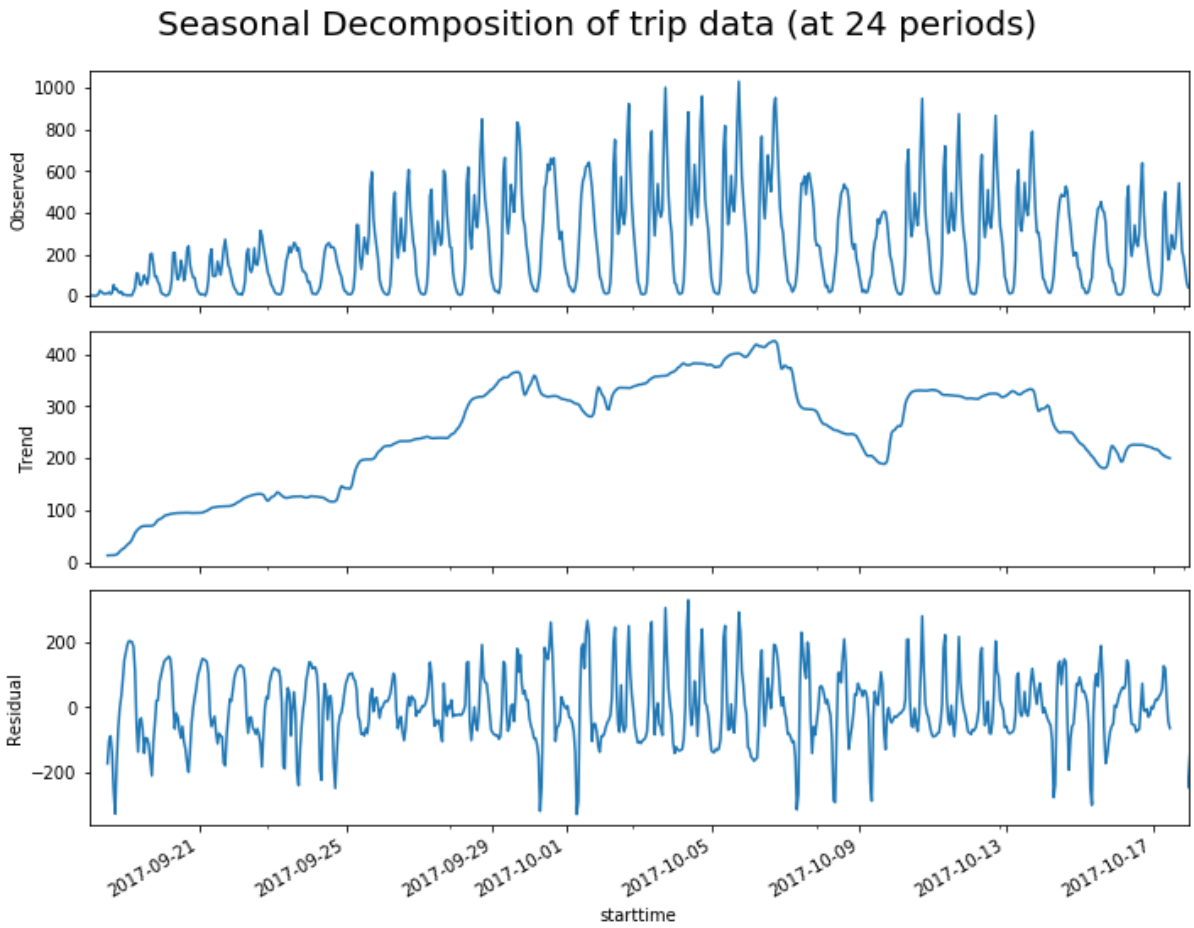

Figure 4.21 *Temporal de-composition of the count of trips recorded by the MTL Trajet travel survey at 24-hour lags.*

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

Trip Purpose	*ADF Statistic*	*p-value*	*n*
All	-2.72	0.069	73,029
Cafe	-2.73	0.067	3,189
Education	-2.86	0.049	2,830
Health	-4.13	**0.000**	1,061
Leisure	-1.86	0.351	9,379
Other	-2.49	0.116	1,219
Pick a person up	-2.86	0.049	1,592
Returning home	-2.85	0.050	27,128
Shops	-1.96	0.301	8,554
Work	-2.25	0.185	19,241

4.2.1 Spatial trends

Statistically significant, at the 99.5th confidence interval, spatial positive autocorrelation is found across the study in each unique trip purpose as discovered by their global Moran's I statistics (see **Table 4.7**). Positive autocorrelation is re-affirmed by LISA maps of each unique trip purpose. Areas of high spatial association are seen on the island of Montreal in all trip purpose classes (Anselin, 1995; **Figue 4.22**).

Table 4.7 *Global Moran's I tests by trip purpose (significant below 0.005 shown in* ***bold****).*

Trip Purpose	*Moran's I statistic*	*p-value*	*n*
Cafe	0.573	**0.000**	3189
Education	0.587	**0.000**	2830
Health	0.548	**0.000**	1061
Leisure	0.544	**0.000**	9379
Other	0.552	**0.000**	1219
Pick a person up	0.562	**0.000**	1592
Returning home	0.619	**0.000**	27128
Shops	0.592	**0.000**	8554
Work	0.591	**0.000**	19241

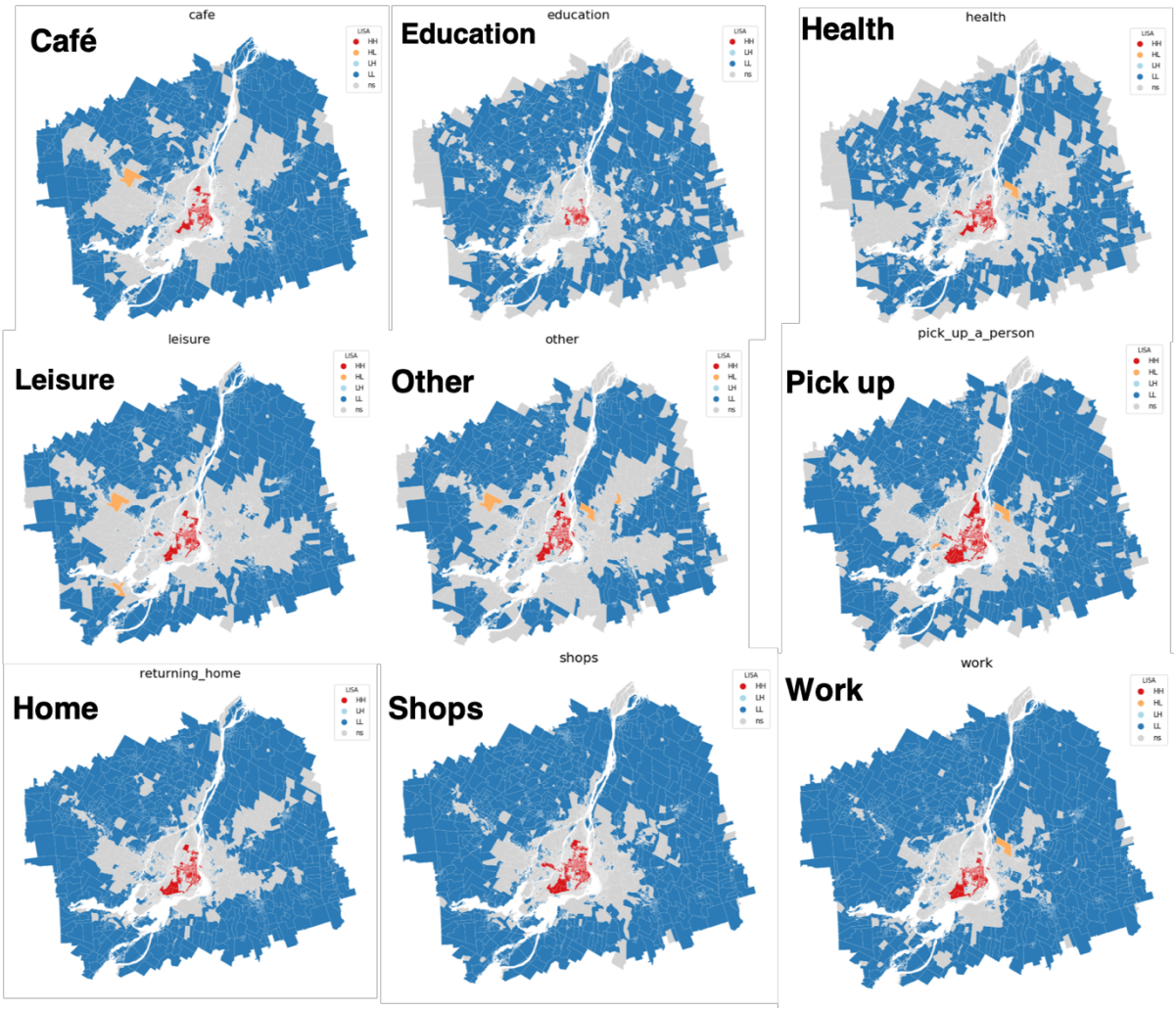

Figure 4.22 *Local indicator of spatial association (LISA) maps of local Moran's I of trip origin and destination points for each trip purpose class.*

4.3 Trip Purpose Classification Models

4.3.1 Model hyper-parameter tuning

After a using a grid-search techniques to look for the best parameters to use in the SVM and MLP models the best model parameters are shown in **Table 4.9**.

Table 4.9 *Results from the hyper-parameter tuning of the models.*

Model	*Best parameters**
Support Vector Classifier	***alpha:*** *0.01* ***Cost function:*** 0.1 ***Gamma:*** 0.01 ***Kernel Function:*** Radial-basis ***Modelling Strategy*****:** One vs One
Multi-layer Perceptron	***Hidden Layer Shape:*** 50,50,50 (3-layer) ***Iterations:*** 500 ***Solver:*** Limited-memory BFGS

* as determined by training on all the MTL Trajet data (n=71,801)

4.3.2 Classification results

After training an initial Random Forest and evaluating feature importance of the model inputs, we remove 10 out of 18 features based on the threshold of 0.05 (*3.5.4*; **Figure 4.23**). Notedly, temporal cluster labels are found to be the most important (*feature importance=0.11*), whereas the spatial clusters are removed. The only spatial feature remaining in the models after other features are removed is the destination land use – which, conceptually, makes sense if we are trying to discern *where* people are going and for which purpose.

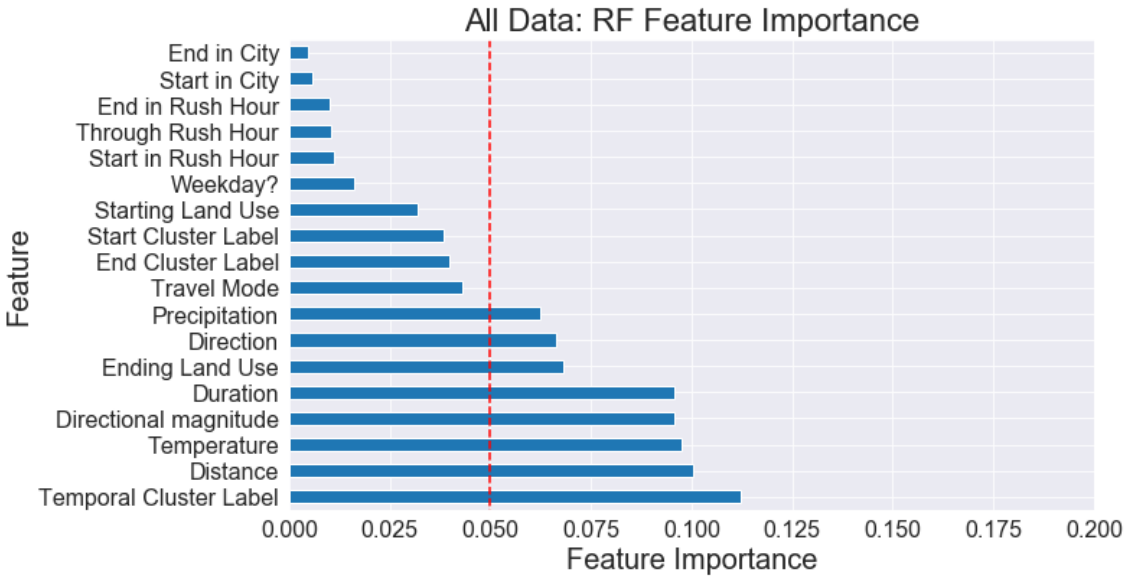

Figure 4.23 *Feature importance from a Random Forest Regression model of the entire MTL Trajet dataset (the red line indicates model features which will be removed).*

Overall, the accuracy of the three trip-purpose classifiers were found to vary in performance between 48.4–69.3% after 5-fold cross-validation (**Table 4.9**). We see that the MLP and RF only make predictions on 56.4% and 60.8% of the data, respectively. One reason for this is that these multi-class classifiers so they evaluate all the trip-purpose classes at once, as opposed to the SVC which evaluates the classes one-vs-one (see **Table 4.9**). Notably, the SVC is found to performs the least accurately.

Table 4.9 *Overall accuracy in within the classifier models.*

Classifier	*5-fold cross-validation*		*Trips where prediction is made (%)*
	Average Accuracy (%)	*CV Range (%)*	
Random Forest	67.2	0.9	56.4
SVC	48.4	1.2	100.0
MLP	69.3	1.5	60.8

When broken down by individual trip purposes in **Table 4.10** (as shown visually in **Figure 4.24**), work and returning home trips were overwhelmingly the most accurately trip-purpose. However, this is likely as this was the predominant class in the dataset perhaps leading to overprediction. MLP was shown to have the highest precision rate, with about half of the trip purposes above 0.5. This indicates that this model was slightly more confident when making prediction for these classes. Additionally, RF was found to have the highest recall for Work and Returning Home trips, indicating this model was the most outright accurate with the classification of these.

Table 4.10 *Results from the classification broken down by class of trip purpose (values above 0.5 are shown in **bold**).*

Random Forest Classifier				
	Precision	*Recall*	*F1-score*	*Support*
	0	0	0	245
Work	**0.76**	**0.86**	**0.80**	4271
Shops	0.40	0.16	0.23	1045
Returning Home	**0.66**	**0.91**	**0.77**	5548
Health	0	0	0	159
Leisure	0.34	0.12	0.18	1288
Education	0.36	0.06	0.10	515
Cafe	0.20	0.01	0.02	404
Overall Accuracy			**0.67**	13375*
Support Vector Classifier				
	Precision	Recall	F1-score	Support
Pick up a person	0	0	0	510
Work	0.43	**0.62**	**0.50**	6247
Shops	0	0	0	2758
Returning Home	**0.51**	**0.84**	**0.63**	8879
Health	0	0	0	332
Leisure	0	0	0	2998
Education	0	0	0	955
Cafe	0	0	0	1026
Overall Accuracy			0.48	23695*
Multilayer Perceptron				
	Precision	*Recall*	*F1-score*	*Support*
Pick up a person	0	0	0	282
Work	**0.75**	**0.87**	**0.80**	4553
Shops	**0.56**	0.07	0.13	947
Returning Home	**0.67**	**0.95**	**0.78**	6232
Health	0	0	0	173
Leisure	**0.52**	0.06	0.11	1307
Education	0.47	0.01	0.02	576
Cafe	0	0	0	351
Overall Accuracy			**0.69**	14421*

* Based on where the classifier actually made a prediction

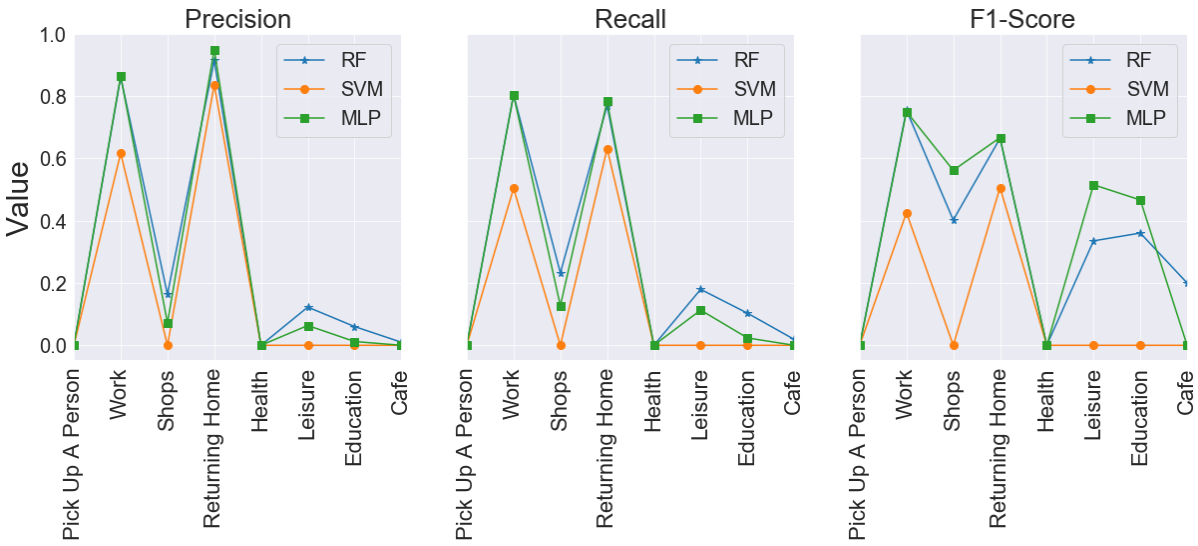

Figure 4.24 *Comparison of precision, recall and f1-score across the three types of classifier*

We discover that the classifiers are all accurate in predicting the trip purpose for 6,857 trips, but are all inaccurate in a further 10,317. When considering that a total 13,378 of correct trips are identified by any one of the models, the fact that only ~7000 trips have been identified by all of them suggests that each of the models have mapped onto different non-linear patterns within the data. We compare the amount of accurate trips predicted for each one of the models in a matrix in **Figure 4.27**. For example, where RF True +SVM intersect this indicates that these two models share 575 trips that were correctly predicted in both, but not in the NN. Where, RF True + RF intersect this means that only the RF has correctly predicted a trip. As shown in this figure, the SVM dominated trips that only it correctly classified.

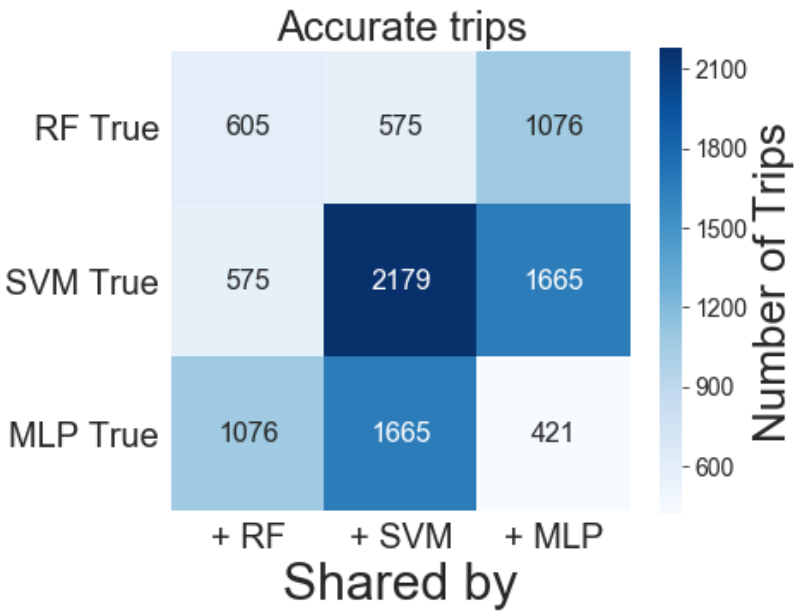

Figure 4.27 *Matrix showing the amount of correctly identified trips in each one of the models compared to the others (RF True +RF means only RF is correct).*

A comparison of random over- and under-sampling techniques for the trip-purpose classification models are shown in **Figure 4.26**. Noticeably, the RF model is shown to become much more accurate (from 67.2% to 80.5%) with a random over-sampling technique.

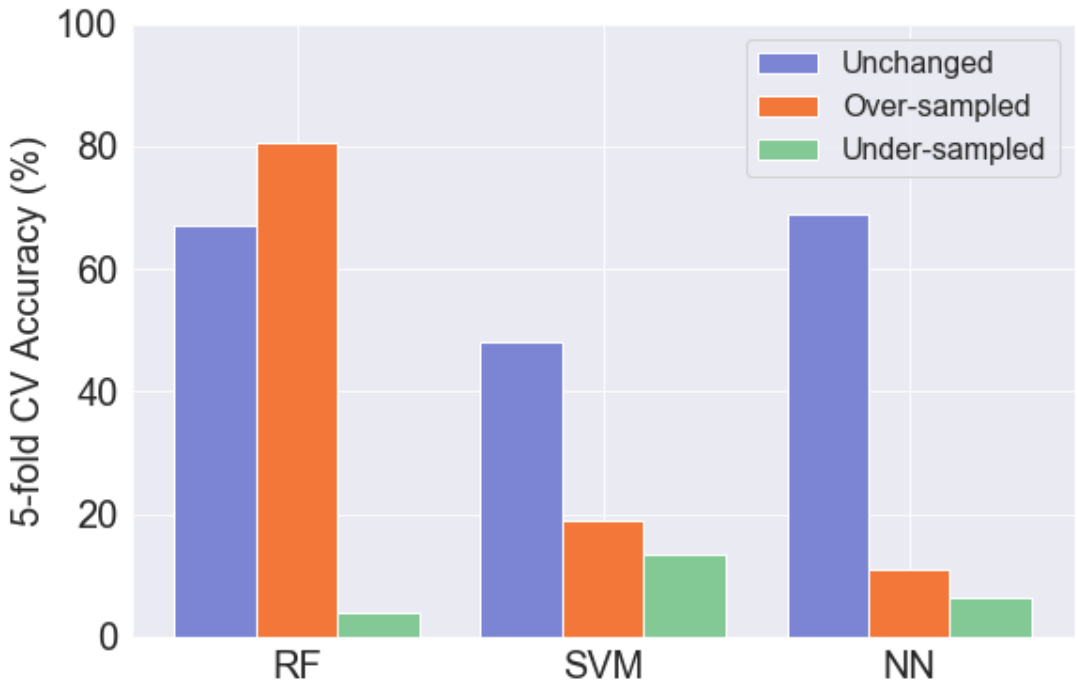

Figure 4.26 *Bar plots comparing the 5-fold cross-validated accuracy of unchanged, random over- and under-sampling techniques.*

Finally, we see some difference in the spatiality of the destination of trips which were both classified correctly and incorrectly, which are visually represented in **Figure 4.29**. As shown, the models were accurately able to identify trips purpose around the Island of Montreal and less accurately able to discern trip purpose at further regions of the study area.

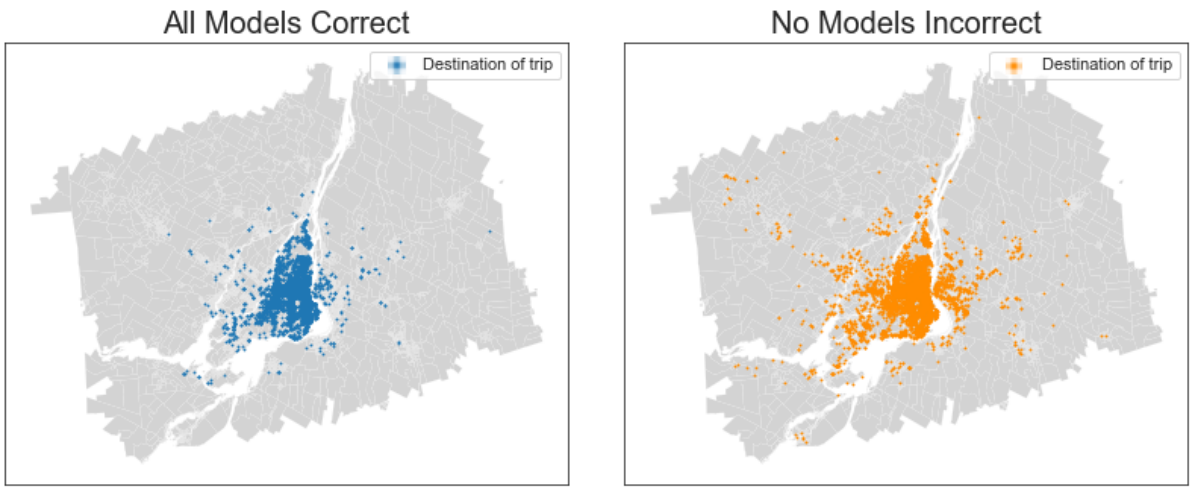

Figure 4.28 *Destination of trips that are classified correctly (left) and incorrectly (right) by all the classifiers.*

Some degree of temporality in the trips that were both accurately and inaccurately predicted, highlighted in **Figure 4.29**. Here, the SVM model is shown to become relative more accurate at forecasting the user trips towards the latter part of the study compared with the RF and MLP.

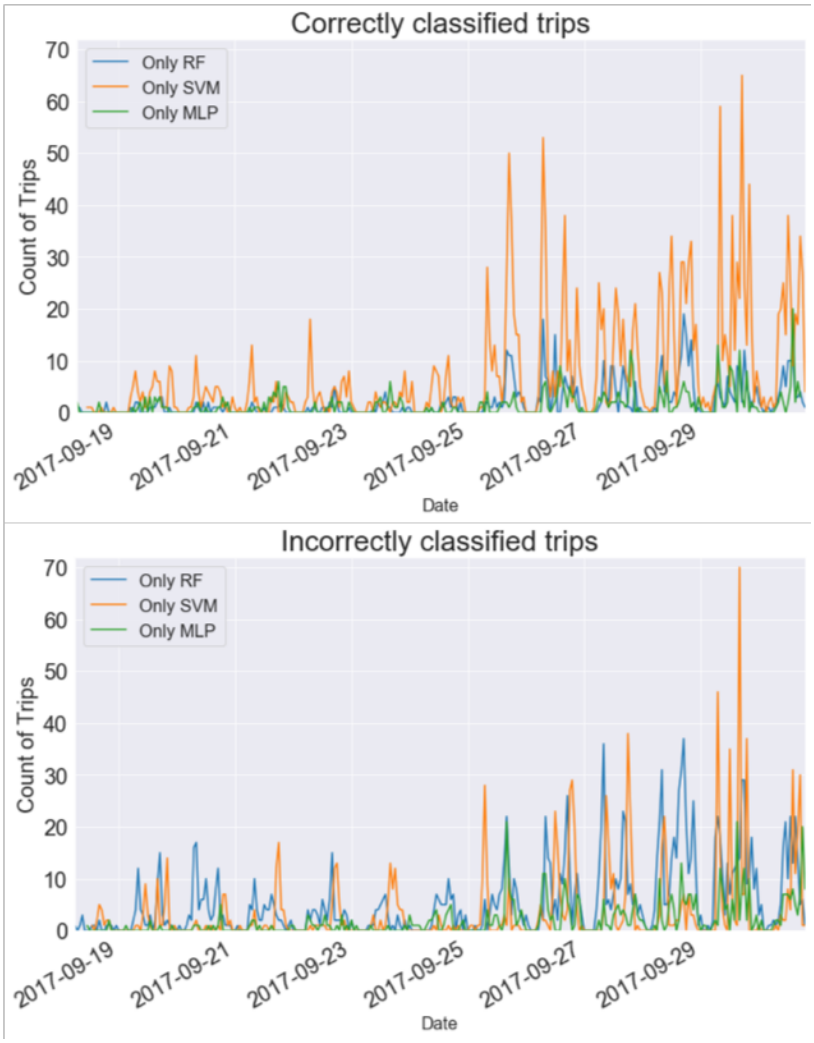

Figure 4.29 *Amount of correctly (top) and incorrectly (bottom) classified trips by each of the classifiers across the study period (18th September 2017–18th October 2017).*

Chapter 5. Discussion

5.1 Evaluation of research objectives

5.1.1 Main research question: Can we effectively classify trip purpose?

Overall, the classifiers struggled with comprehensive predictions across the unique trip-purpose classes within from the 2017 MTL Trajet. We do find that the models were quite effective in the classification of trips for the purposes of *work* and *returning home*, so we can infer to some extent about the types of urban movement patterns that these classes exhibit. Indeed, trips detailing both these classes were found to be very regular in both space and time.

The models struggled with the classification of shopping and leisure activities, something also found within the literature (Attard *et al.*, 2016). One factor for this, may simply be that a multi-class classification model is not an effective strategy for studying trip purpose of such heterogeneity. For example, we do not separate purposes that are time invariant and time variant and this may have been very problematic for the classifiers.

Instead a better strategy may have been to create broad categorisation for some of the trip purposes and individualised for other models. Indeed, it is likely that the space-time controls on each of these purposes will be vastly different and ignoring this may have led to a degree of omitted variable bias (OVB).

Correspondingly, we omit spatial cluster labels, city vs non-city and rush hour for the classification. These may be more important in the prediction for certain trip purpose. For example, it is likely that a completely different set of dynamics govern *how* and *why* people travel within the suburbs of Montreal vs downtown i.e. people being able to walk to shops downtown versus having to drive in the suburbs. Notably in 4.1.4, a greater proportion of leisure, shops and café trips are outside of rush hour and the city.

Moreover the discovery of effective classification in *work* and *home* trips is not something that serves as a new insight in mobility research as we expect to be able to characterise these (Meng *et al.*, 2019). Further, this trend may simply be a function of these trip purposes being the predominant class within the data set.

We also make an argument that the study area itself is unfeasible for effective classification as we see indications of both spatial and temporal non-stationarity across the study region and study period. This make the modelling procedure somewhat redundant, and may be an explanation as to why we see more errored trips in the suburbs of Montreal and towards the latter part of the study period (**Figure 4.28+29**).

5.2.2 Sub-Question: Which indicators were the most useful?

Broadly from the results of the analyses we do not overwhelmingly find one primary indicator for trip-purpose, although temporal clusters are found to be most important determined by *feature importance*. Notably, these highlighted temporal dependencies in the various purpose classes and were broadly found to be more important in the classifier performance of time-invariant trip purpose classes (*work*, *education*).

Direction and magnitude of direction (4.1.3) were also discovered to be an important predictor. Indeed, we see some purpose classes that have some identifiable directional dependence (work and returning home – as people travel to and from suburbs). And some activities with directional independence (cafés – as people may simply head to nearest and not head for a specific café).

5.2.3 Sub-question: Which models performed the best?

There is clear difference in the types of trips that each classifier was able to identify. Notably, each of the models discovered different trends suggesting they each able to mapped onto different non-linear trends within the data. The SVM and MLP were similar in terms of the trips that they predicted which the RF could not. These models which rely on the conversion of feature space into higher dimension to find trends, may have found non-linear patterns that inherently probabilistic methods (such as RF) may not have. Then again, we have no way of comprehensively knowing this as both SVM and MLP are 'black boxes'.

We find over-sampling to improve the performance of the RF (which reached an accuracy close to 80%), but unexpectedly the over- and under-sampling was ineffective for the SVM and MLP. Arguably, this may have been due to these model being been underfitted and so would have perhaps benefitted from further hyperparameter tuning (Semanjski *et al.*, 2017).

5.3 Uncertainty

The modelling set-up itself focuses on movements in Montreal, something which is specific to the city and its unique network topology. As such, we cannot be overly confident in transferring any findings from this report to other cities. Indeed, we can argue that the spatial and temporal trends from the results are 'frozen' in time and space.

Further, it is inaccurate to assume that what is examined in Montreal at the time of the study period can even be reapplied to Montreal at different points in time (i.e. to Winter or 5 years in the future or past), let alone to another city (Gong *et al.*, 2019). For real world decision-making in urban environments, we cannot comprehensively use too much of the information from the MTL Trajet to analyse movements outside of the realm of the study area.

Due to the types of model used (i.e. machine learning methods are non-linear) we still have a lack of understanding over the unique govern principles of why people make trips – a major gap also noted in other research of trip purpose classification (Meng *et al.*, 2019).

5.3 Further research

Although no explicit metric was discovered to be overwhelmingly important in classification models in this study, more research is needed to evaluate the potential of a wider range of metrics which could be used in combination with each other. Proposedly, accounting for the spatio-temporal interdependencies within the MTL Trajet could be used for to more effect that way that space and time separately (as in this report; Aslger *et al.*, 2018).

Finally, more work is needed to account for space-time, arguably the use convolutional neural networks may be used to solve this problem, as they could represent the MTL Trajet trips as spatial images. Additionally, combining a CNN with models which account for temporal memory (i.e. CNN–Long-Short Term Memory; Shi *et al.*, 2015) may be used to represent the trips as videos which these networks can study patterns in.

6. Conclusion:

In conclusion, we present an in-depth analysis into the feasibility of using space and time indicators in trip purpose classification modelling and find that they offer some degree of explanation in seemingly chaotic trip purposes.

Despite this, the modelling approach used in this report only focuses on Montreal and only for one month in September. And to this extent the research is *frozen* in time and *limited* by space. We can thus assume, this modelling procedure may have a completely different result for other cities. Moving forward, it is clear that trip purpose classification models will need include more contextual information if we hope to correctly identify key indicators of travel purpose. But, indicators themselves will need to be individualised to specific activities and cities.

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 13th August 2019].

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.

Howell, S. (2018) Montreal City Guide [Online]. https://adventure.howstuffworks.com/montreal-city-guide1.htm Available at: [Accessed 1st August 2019].

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

Kim, Y., Pereira, F. C., Zhao, F., Ghorpade, A., Zegras, P. C., & Ben-Akiva, M. (2015). Activity recognition for a smartphone and web based travel survey. In *22nd International Conference on Pattern Recognition* (p. 9).

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, X., Chen, Z., & Wang, H. (2015). Convolutional LSTM Network : A Machine Learning Approach for Precipitation Nowcasting, 19 Sep 2015, 1–12.

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.

Tobler, W (1970) “A computer movie simulating urban growth in the Detroit region”, *Economic Geography*, 46(2): 234-240.

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 anonymised data that is public available from the City of Montreal's Open Data Portal, this study was judged to be minimal risk due as informed by 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 at https://github.com/Thomasjkeel/MSc_Dissertation under the '*utils*' sub-directory.

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

Python scripts and *Jupyter Notebooks* which detail the procedure of the analyses carried out in this report are available online at https://github.com/Thomasjkeel/MSc_Dissertation. Note that, README.md detail the specifics.