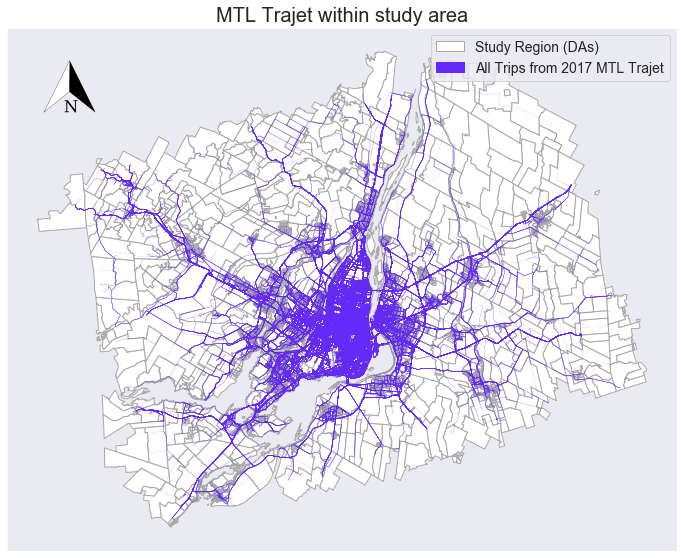
# Chapter 3. Methodology

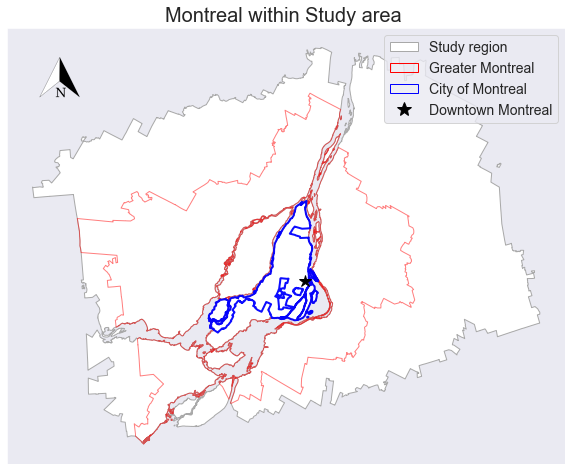
## 3.1 Study Area

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 MTL Trajet 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). The city of Montreal itself contains 19 sectors (or arrondissements) across 431.50 km2 (166.60 mi2) (WPR, 2019). The total of area of the Greater Montreal region is 4,259 km2 (644 mi2) (Chevalier *et al.*, 2018). The extent of the City and Greater region are shown in **Figure 3.2**.



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

To allow the analysis of this project, all geographically-referenced data was 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.

## 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. The Open Source Routing Machine (OSRM) has been used on the GPS trace so the route aligns with features of the Montreal road network (Patterson, 2016). 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 mode and purpose allowed for trips in the MTL Trajet 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

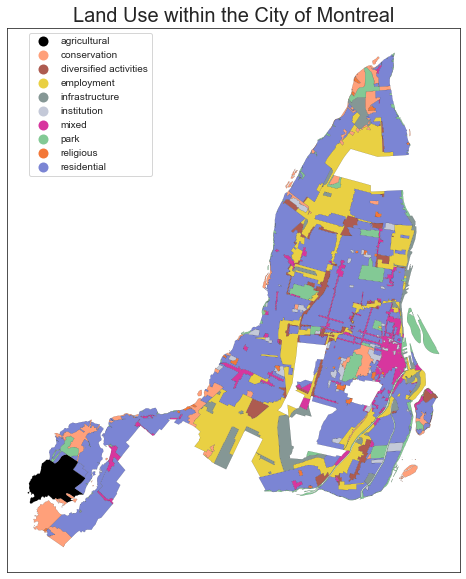
The time signature for the start and end of each trip has been converted from Coordinated Universal Time (UTC) to Eastern Daylight Time (EDT). This being the time zone that Montreal falls within and for the purpose of analysis using Python’s *datetime* library. The duration of each trip was calculated 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. Data from the *City of Montreal’s 2014 Plan d'urbanisme* was collected from the Montreal Open Database for land use (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 desination.

**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 weather, 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.

Furthe,r 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.

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

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

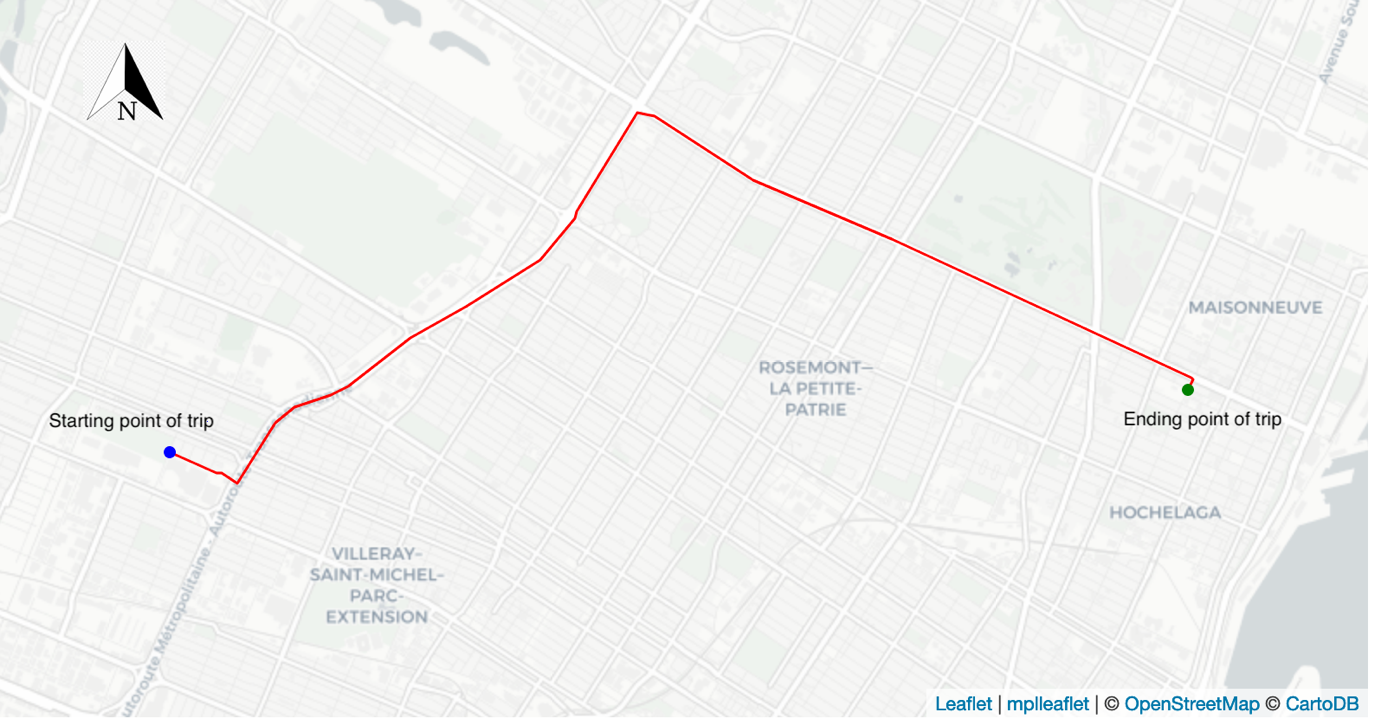
|  |  |  |  |
| --- | --- | --- | --- |
| *Section* | *Times* | *Days* | *Hours (each day)* |
| *Rush hour* | 6:00–10:00 & 15:00–19:00\* | Monday – Friday | 8 |
| *Off-peak* | Times outside *Rush hour* | Saturday – Sunday | 16 |

\* 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 (Anistrophy and Isotrophy) of each givencategory of trip purpose within the MTL Trajet data.

**Figure 3.X**, shows The algorithm calculates this to be E trip



**Figure 3.X** Example of an eastbound trip across Montreal (*trip-id=* *150744*)

To achieve this, each trip (in a LineString format) was first re-projecting 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:

(1)

where refer to the coordinates in decimal degrees of the first and second points of a pair respectively. Euclidean distance in degrees () was also calculated for each point pair using:

(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 Clusters

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 of a spatial and temporal cluster. The use of clusters as model inputs, as opposed to raw 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, further, to speed up model training times (Montini *et al.*, 2014).

- [Clusters] not only suggests characteristics of the pattern itself but also of its background processes involved across the city (Yamada & Thill, 2010)

For use in a k-means clustering algorithm, we extract only the starting and ending coordinates of each individual trip. As each trip is essentially an interaction between an origin and destination, it could thus be proposed that the route taken between them is less important within any model (Murray *et al.*, 2012). 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).

Temporal clusters have been created from the data in this report using a Latent Dirichlet Allocation (LDA) model and a methodology adapted from Liu & Cheng (2018). LDA is an probabilistic topic identification technique commonly used in the classification of topics in large bodies of unstructured text (Blei *et al.*, 2003; Doll, 2018). As LDA is a topic modelling algorithm, the information from the MTL Trajet regarding the day, time and purpose is first converted into ‘temporal words’ (after Liu & Cheng, 2018). For example, a trip to work beginning at 7.15am on Monday becomes a sentence containing two temporal words: ‘Monday\_7, work. The LDA model can discover patterns within a given collection of these sentences and discern words that have a high probability to cluster (e.g. work and Monday\_7).

LDA built with 50 passes we try to minimise perplexity and enhance coherence (Kumar, 2018)

Coherence suggests words within the topics are similar (Kumar, 2018)

perplexity of model continues to drop passed 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).

Look for similar temporal characteristics within the data.

Weighted importance -> probability that this word is associated with this class

The collection of all the sentence created from the MTL Trajet trips are used to train an LDA model with a given amount of topics (or temporal clusters) using 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 (clusters) for the data (after Liu &Cheng, 2018). Finally, the characteristics of each of the topic identified by the LDA are used to assign each trip in the MTL Trajet to a temporal cluster label. 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 the assess feasibility of modelling of each class of trip purpose, this section sets out the methods 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 class. Local Indicator of Spatial Association (LISA) maps have been 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 purpose class (Anselin, 1995). It is hoped that these LISA maps can be used to infer the level of spatial dependence in the origin and destination of the different types of trip purposes.

Values, above the 99.5% confidence interval (p>0.005) will be considered as statistically significant, after new standard introduced by Benjamin *et al.* (2018).

To examine the temporal structure within the trip purpose classes, methods have been employed to investigate temporal dependence, stationary and correlation.

For temporal dependency, the data has been grouped by hour and day of week for each purpose class. These are then plotted in a ‘calendar’ format (7-days\*24-hours) which allow us to capture daily and hourly temporal trends of each purpose (after Arribas-Bel & Tranos, 2017).By 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 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). Finally, to determine the degree of temporal correlation between the trip purpose classes we use a spearman’s rank correlation matrix to compare the count of each purpose per hour to all other purpose classes.

### 3.4.3 Outlier detection

Due to unknowns about the data and user input, we can only use trip distance and duration to inform the outlier removal. Data has already been cleared for mis-represented GPS through OSRM (Ville de Montréal, 2019).

As such, trip distance and duration will be used in conjunction with each other to inform the outlier removal process with trips between 60 and 10,800 seconds (1 minute and 3 hours) and 50 and 100,000 meters (0.05 km to 100 km) being removed from the analysis. Note, the MTL Trajet has already been cleaned for trips with errors in their speed/acceleration (Patterson & Fitzsimmons, 2017b). These thresholds have been decided after initial testing and what is likely to skew the data and affect the ability of the classifiers.

* Is a form of knowledge discovery (KD)﻿.Anomaly detection is inherently challenging as it requires a clear definition of what is considered to be normal and abnormal (Li *et al.*, 2016)
* Within a city, this could be considered a reasonable range of values edge results may not be important for the general classification

## 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 (e.g. the probability that a trip is for work based on time, location, etc.). In RF, a multitude of decision trees are used to make a classification with each tree having one vote as to which class they expect an input (or trip) to be part of (Montini *et al.*, 2014). Their ability to handle class imbalance have meant they have become the primary tool in trip purpose classification (Gong *et al.*, 2018)

Bagging, depth

As opposed to other form of classifiers used in this report, RF instinctively avoid overfitting even as more trees and branches are added to the model (Breiman, 2001).

They are the most popular form of model used in trip purpose classification (Gong *et al.*, 2018). Although, additional performance gains can be achieved by adjusting hyper-parameter of RF, no model tuning is carried out in this analysis (Segal, 2004).

Two distinct forms of Random Forest methods will be used in this report, a regular and multi-class structure.



**Figure 3.4** Adapted from Koehersen (2017)

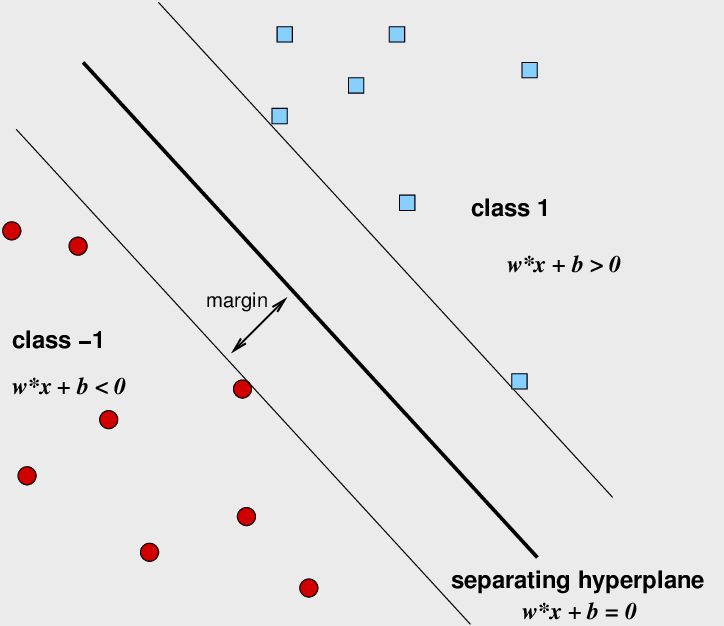
### 3.5.2 Support Vector Machine Classifier

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 model inputs 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: something referred to as the ‘kernel trick’. The threshold for this determines how accurate the separation of data classes by a hyperplane needs to be is determined by a cost function. Where cost functions are set too high, a model can suffer from overfitting (ref).

* Can handle high dimensions of feature spaces, making them useful for trip purpose classification, although can suffer from over fitting (Zhu *et al.*, 2014)

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.

* Cross validation to avoid over-fitting



**Figure 3.5** (adapted from: Aluja-Banet, 2009)

### 3.5.3 Multi-Layer Perceptron Classifier

The Multi-Layer Perceptron (MLP) is 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. In basic ANNs, data is sent into the hidden layer via links called weighted synapses. These synapses

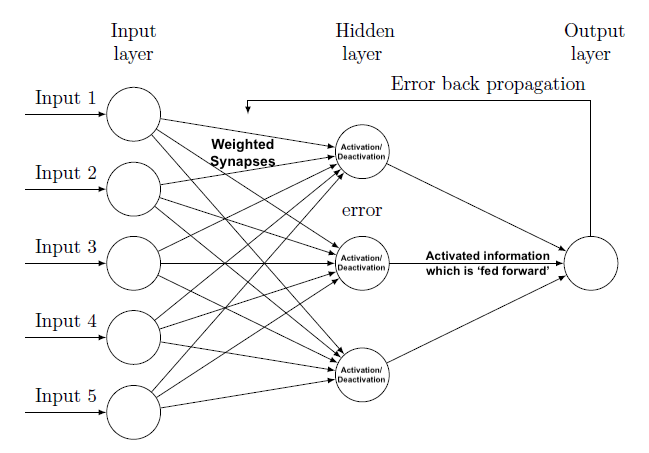
MLPs are considered to be ‘black-box’ models as it is unclear which weights

Activation function, weights and biases

(TDSB, 2016)

Can handle binary or multi-class, making them useful strategy for trip purpose classification (Xiao *et al.*, 2016).

We tune the model on a range of different hidden layer structures between 1-3 hidden layers. (5,10; 5,10,5; 10,)



**Figure 3.6** Structure of a generic Feed-forward Artificial Neural Network

### 3.5.4 Model Training

Each model will be trained using 3-fold cross validation and a training/testing data 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)

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 will be based on those input variables with a feature importance score of 0.05 and above. The *Scikit-Learn* library is used to apply an One-Hot Encoder to all categorical model inputs (listed in **Table 3.5**). Specifically, this encoder creates dummy binary variables for each unique category. 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.

The SVM and MLP models will have undergo hyper parameters tuned using a grid-search approach, whereby a range of different parameters will be used.

## 3.6 Limitations:

### 3.6.1 Methodological

The major drawbacks of the methodology fall into X categories, problems with both the temporal and spatial aggregation of trips/data and problems with .

Firstly, with aggregation

Significant issues also exist relating to the aggregation of space and time into clusters, as what we see at one scale, we might not at another (Ecological fallacy). Indeed, most processes in a city may be subject to an ‘edge effect’, whereby space and time just outside unit have an effect but are not recorded. Something which is

Also, use of spatial boundaries such as the City of Montreal being defined as a hard border and Land Use (binary variable) suffers from the modifiable areal unit problem (MAUP; Openshaw, 1984). Discounts that people move through these barriers, as such there is an edge effect (Kwan, 2018).

Interactions/Neighbourhoods in space-time are always expanding and contracting, such that what may be a neighbourhood at one time and space may not be in the future (Cheng *et al.*, 2012)

Further, temporal aggregation such as into rush hour and grouped by hour in the analysis suffers from the modifiable temporal-unit problem (MTUP; Cheng & Adepeju, 2014). Indeed, how you sample temporal information can have a dramatic effect the conclusions drawn/outputs (Zhao *et al.*, 2019). Temporal Edge effect.

There is also significant class-imbalance that exists within the data, the SVM and MLP less suited to this imbalance than RF (Shi *et al.*, 2018; Yazidadeh *et al.*, 2019)

“An important limitation is the impact of an ‘edge effect’ whereby regions or cells just outside the study regions, will influence the variables, but not be recorded (Openshaw, 1984) “

- Although, additional can be achieved by adjusting hyper-parameter of RF, no model tuning is carried out in this analysis (Segal, 2004)

- Significant class-imbalance exists in the MTL Trajet data affecting the ability of models to classify the data (see figure with histogram of classes)

- Various forms of spatial and temporal edge effect i.e. what is occurring outside the study regions ().

*On Neighbourhood effect ->* People often traverse neighbourhoods and boundaries throughout one day (Kwan, 2018)

- trips below 60 seconds and 100 meters being removed from the analysis.

- regularisation of space

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

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

### 3.6.2 Data

- General spatial error with GPS (drift and jump; Bantis & Haworth, 2017)

- Representativeness is problem present in all forms of VGI, especially this one

Uncertainty exists within the data itself as, as with all VGI I (Elwood *et al.*, 2012), it is difficult to falsify data from MTL Trajet in this analysis. As such, there may be mis-inputted data which can affect the accuracy of the classification models (Shi *et al.*, 2018). This studied is tied to the initial data handling and cleaning of the MTL project, so unknown unknowns exist.

No indication of which users are making which trips (as the data is completely anonymised) (Yazdizadeh *et al.*, 2019)

Also, computationally limited based on the size of the dataset, hyper parameter has thus been minimised for this analysis

Ultimately, data used in the classification only applies to study period (i.e. one month) and only to Montreal. In terms of the temporality of the data, It has been found that weather and season affect the way people travel (Xie *et al.*, 2016), thus applying the models from this analysis to the same study area at different parts of the year may be problematic.

- Compared with data from professional vendors, spatial UGC [VGI] faces greater authenticity issues, such as mistaken data due to carelessness or inadequate (Shi *et al.*, 2018)

\*\*Problems with training and testing on different parts of the year -> ﻿However, the “feasibility and effects of choosing these data from dif- ferent periods of the year are still unknown” Gong *et al.* (2018) [i.e. can’t apply to other parts of the year]

\*\* multi-label vs binary classification (in this we do multi-label for NN and RF -> thus can make a prediction of multiple or no classes)

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]

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].

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