# 3. Methodology:

## 3.1 Study Area

The study area chosen for this project spans across the Greater Montreal region in Eastern Canada (**Figure 3.1**). A shapefile containing Canada’s 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 subset of DAs was created where there was an spatial overlap between the DAs and the GPS tracks of respondents to the 2017 MTL Trajet survey. As illustrated in **Figure 3.2**, the resulting study area details a total of 7046 amount of dissemination areas across the Greater Montreal region.

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

**Figure** **3.2** 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 3.3**.

To allow the analyses of this project, all geographically referenced data were re-projected into the Statistics Canada Lambert (or NAD83), which is a projection Canadian-centric projection with a 1 metre unit (EPSG, 2019). The re-projection of the data was carried out using Python’s *geopandas* library (ref).

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

## 3.2 Data pre-processing:

This study makes use of one main and two supplementary forms of data to aid its analysis. A description of the data and the methods used to both retrieve and handle them is covered in this section.

### 3.2.1 MTL Trajet

Data detailing the results of the *2017 MTL Trajet* survey carried out within Montreal, Canada between 18th September 2017 – 18th October 2017, was retrieved from the Montreal Open Database (Ville de Montréal, 2017). This data is in a GeoJSON format and details 185,285 unique trips from 4,425 unique respondents (MTL Trajet, 2017). Each trip 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 (**Table 3.1**).

This study makes use of the geo-routed version of the raw GPS points Open Source Routing Machine

**Table 3.1** Description of the key 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 mode of transport used for a trip | String | 74,218 |
| *purpose* | The purpose of the activity 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* | WGS84 spatial reference system | 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 such 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.*

For this study, the MTL Trajet data has been translated from its original French to English. The translated version of the unique categories of the mode and purpose of the trips is shown in **Table 3.2**. Note that, the MTL Trajet app allowed respondents to choose any combination of travel mode categories per trip, however 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 starttime and endtime variable of each trip has been converted from Coordinated Universal Time (UTC) to Eastern Daylight Time (EDT) for the purpose of this analysis using Python’s *datetime* library. The duration of each trip was calculated in seconds by taking the difference between these two time signatures. Finally, 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.

### 

Outlier removal:

Distance and time will be used in conjunction with each other to inform the outlier removal process, with trips below 60 seconds and 100 meters being removed from the analysis. This is not founded in any literature, but it could be suggested that this is appropriate

### 3.2.2 Supplementary data

2-m surface temperature and precipitation 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 X). This data was first loaded into Python using the *iris 2.0* library (Met Office, 2018) before being re-formatted and output into csv. The purpose of this was to supplement the information from the trips, as it has been found in the literature that weather has an strong influence on transport mode choice (Dubos-Golain *et al.*, 2017).

### Land Use Data

Land use data from the City of Montreal’s 2014 Urban Land Use Plan (Plan d'urbanisme) was collected from the Montreal Open Database (Ville de Montréal, 2014). The data contains 10 unique categories of Land Use which are mapped in **Table 3.3** **Figure 3.5**. The purpose of this data is to add spatial context to the trips within the city of Montreal. As shown in **Figures 3.4+3.5**, the land use categories are fairly unbalanced with the majority of land use being residential and employment (27% + 18%, respectively).

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

|  |  |  |
| --- | --- | --- |
| *Land Use Category* | *Description* | *Total Area (%)* |
| Agricultural | – | 6.38 |
| Conservational | – | 8.85 |
| Diversified activities | – | 6.81 |
| Employment | – | 16.67 |
| Infrastructure | Including transport infrastructure (stations, railway lines, airports) | 8.12 |
| Institution | Major institutional Facility including Governmental | 5.20 |
| Mixed | – | 8.19 |
| Park | – | 9.09 |
| Religious | – | 2.84 |
| Residential | – | 27.86 |

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

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

### Development of space-time metrics:

A number of both spatial and temporal metrics have been created for the MTL Trajet dataset to aid the analysis carried out in this report. Binary labels were created from the data to indicate whether a trip occurred inside or outside of the City of Montreal (by using the *intersects* method within *Shapely*; after Zahabi *et al.*, 2017), whether the trip had begun or ended in a rush-hour period (see **Table 3.4**) and whether the trip had begun or ended on a weekday or weekend (after Liu & Cheng, 2018). This was decided upon 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 governing spatio-temporal dynamics change throughout the day and across the city.

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

|  |  |  |  |
| --- | --- | --- | --- |
| *Section* | *Times* | *Days* | *n (each day)* |
| *Rush hour* | 7:00 – 9:00 and 17:00 – 19:00 | Monday – Friday | 6 |
| *Off-peak* | Times outside *Rush hour* | Saturday – Sunday | 18 |

*Direction (unfinished):*

For this study, we calculate the cardinal direction of each trip (i.e. W, N, E, S, etc.) to investigate the directional dependence (Anistrophy and Isotrophy) of giventransport modes and purposes (ref).

The mean direction was calculated for each trip by first converting each trip’s route into its individual decimal degree points pairs and calculating the bearing between them. The bearing (θ) in degrees was calculated using the following:

(1)

where refer to the coordinates in decimal degrees of the first and second points of the pair.

Euclidean distance used for mean direction:

*magnitude = math.sqrt(((pnt2[0] - pnt1[0])\*\*2 + (pnt2[1] - pnt1[1])\*\*2))*

After this is calculated for each pair of points in each trip, the mean direction is calculated from the collection of these points using:

Calculations for mean direction in degrees:

*V\_east = magnitudes \* np.mean(np.sin(from\_dir \* np.pi/180))*

*V\_north = magnitudes \* np.mean(np.cos(from\_dir \* np.pi/180))*

*then..*

*mean\_dir = np.arctan2(V\_east, V\_north) \* 180/np.pi*

*mean\_dir = (360 + mean\_dir) % 360*

*mean\_dir = np.mean(mean\_dir)*

*Calculations for mean distance (magnitude) in degrees:*

*C = (1. / len(from\_dir)) \* (np.sum(np.cos(from\_dir \* np.pi/180)))*

*S = (1. / len(from\_dir)) \* (np.sum(np.sin(from\_dir \* np.pi/180)))*

*then…*

*resultant\_magnitude = (C\*\*2 + S\*\*2)\*(1./2.)*

This results in a mean direction (in decimal degrees) and mean magnitude for each trip. The overall direction of all trip is shown in a directional windrose diagram in **Figure 3.6**. Here, the mean direction of the trips is towards the NNE and SSW directions, similar to the morphology of the island of Montreal (i.e. it is a NE–SW city).

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

Final description of column within the MTL trajet data set shown in **Table 3.5**.

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

|  |  |  |
| --- | --- | --- |
| *Column* | *Description* | *Format* |
| *Original* | | |
| id\_trip | See Table X |  |
| mode | See Table X |  |
| purpose | See Table X |  |
| starttime | See Table X |  |
| endtime | See Table X |  |
| geometry | See Table X |  |
| *Spatial explanatory variables introduced for this study* | | |
| distance\_m | Total distance of trip |  |
| start\_city\* | Binary variable of whether the trip starts in the City of Montreal |  |
| end\_city\* | Binary variable of whether the trip ends in the City of Montreal |  |
| direction | Mean direction of trip in decimal degrees |  |
| magnitude | Magnitude of direction of trip |  |
| carddir | Cardinal direction of mean direction (i.e. NW) |  |
| *Temporal explanatory variables introduced for this study* | | |
| duration | Number of seconds elapsed for a trip |  |
| start\_rushhour\* | Binary variable of whether the trip starts in rush-hour |  |
| end\_rushhour\* | Binary variable of whether the trip ends in rush-hour |  |

\* see 3.3.1

### 3.2.5 Outlier Detection (unfinished):

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)

Outlier removal (from 3.2.1):

Distance and time 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. This is not founded in any literature, but it could be suggested that this is appropriate. Note, the MTL has already been cleaned for trips with errors in their speed/acceleration (ref). These thresholds have been decided after initial testing and what is likely to skew the data. Within a city, this could be considered a reasonable range of values.

These edge results may not be important for the general classification

## Exploratory Spatial-Temporal Data Analysis

This section highlights the methods carried out to investigate space, time and space-time signatures in the data. It is hoped that the identification of these forms of trends will help inform the modelling process (detailed in 3.4).

### 3.4.1 Spatial Methods

To examine the spatial signature of MTL Trajet routes across the study area the data has been transformed from a network format and aggregated into dissemination areas. This has been achieved by creating an algorithm to calculate an intersection area (using the *geopandas ‘Spatial Join*’) between a given trip and the underlying dissemination areas (see **Figure 3.7**). For a given trip, dissemination areas within the intersection are given a value of 1 (red) and all other areas are given a value of 0 (grey). For the spatial analysis in this report, data has been grouped using *geopandas groupby* function into individual modes and purposes of travel (see **Table 3.1** in section 3.2.1 for types) and then visualisations have been produced from each showing the total intersection count in each dissemination area across the entire study period (see X.X).

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

KMeans clustering algorithm was then applied to each one of these.. used to identify the characteristics of the background processes involved across the city (Yamada & Thill, 2010).

In the interest of computational time KMeans was preferred over other spatial clustering algorithms.

Further, to determine the spatial autocorrelation each purpose Global and Local Moran’s I statistics are calculated for each purpose and mode using Python’s *Pysal* library (Rey & Anselin, 2007). For this, a Queen’s case contiguity spatial weight matrix has been computed from the study area’s dissemination areas used. The Local Moran’s I has been used in the production of Local Indicator of Spatial Association (LISA) maps for each unique purpose and mode.

### Geographically weighted regression (unfinished)

Geographically weighted regression between weather (temperature and precipitation; 3.2.2) and mode/purpose in given areas. A model will be built for each individual mode and purpose. This form of regression makes use of the queen’s case contiguity matrix to account for the spatial lag. It is hoped that this will give an idea of sensitivity of transport in certain regions to weather i.e. Downtown will likely not be as effected by rain.

To Add:

- routes passing into given land use zones (% per purpose)

- LISA (and queens case)

- Moran’s I

- frequency plots and show figure of line through dissemination area

- directionality

- KDE

LISA:

- Merge the data into the dissemination areas

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

### 3.4.2 Temporal Methods

To examine the temporal signature of the trips across Montreal, the data has been grouped by hour and day of week for each purpose and mode. We can then examine a calendar plot (i.e. 7 days\*24 hours) for the mean amount of trips occurring with a given purpose/mode per hour (e.g. see **Figure 3.8**). By examining these calendars we can determine the time-variant and time-invariant properties of given modes of transport and purposes in the trip.

After Zhang *et al.* (2019) look at the ﻿relationship between passengers’ movement patterns and social-demographics by using smart card (SC) data with a household survey. Exploring] ‘how’ (including ‘when’ and ‘where’), ‘who’ and ‘why’ travel in public transit



**Figure 3.8** Count of trips calendar plot

*Detrending for later modelling:*

An Augmented Dickey-Fuller test statistic will be applied to each purpose/mode to evaluate the stationarity of each across the study period (18th September 2017 – 18th October 2017). Detrending (by removing the running mean) will be applied if non-stationarity is discovered in any of the variables.

### 3.4.2.1 Temporal Clustering

Latent Dirichlet Allocation (unfinished)

“Liu & Cheng (2018) conjoins socio-economic data to transit patterns to interpret behaviour

using temporal clusters from LDA and temporal words (Liu & Cheng, 2018)

Liu & Cheng (2018) Looks at who constitutes each temporal cluster (which socio-economic groups)”

A methodology for creating temporal clusters by using LDA from the MTL Trajet data has been adapted from Liu & Cheng (2018). The methodology starts by stringifying the data temporal aspects of the data OR creating temporal words. An example of this is a car trip to work that occurred at 7am on Monday would be transformed into “Monday\_7, work, car”.

Python’s *Genism* and *Natural Language Toolkit* libraries are used to create the corpus of all of these temporal words and build the LDA model on this corpus. The LDA can discover patterns within the corpus and discover terms that cluster i.e. work and [weekday]\_7.

It is hoped that, in identifying the make-up of each of the these clusters we can get an idea of the co-clustering of given periods of the day and certain modes or activities.

Log Perplexity and Coherence scores will be calculated after fine-tuning the model (i.e. adjusting the amount of clusters and topics per cluster.

The methods used to deconstruct the temporal aspects of the data involve grouping by

ToAdd:

* ACF + PACF?

- Time-series

- grouped hour time series

- Groupby hour and day

- temporal calendar

- Detrending?

- Differencing and ADF

Maybe:

* autocorrelation

### 3.4.3 Spatio-Temporal Methods

- directionality throughout day

- space-time calendar (after Arribas-Bel & Tranos, 2017) for given regions of Montreal

- space-time interaction tests in pysal

- SatScan – for Spatio-temporal clusters

- space-time evenness grid

- cross-correlation and Coefficient of determination between given regions/DAs of Montreal

“We aim to create conditions for ‘effective’ space-time forecasting” (Yue & Yeh, 2008).

Merging of other pertinent data (i.e. land-use)

ST-KDE:

See Wei *et al.* (2018) (Also uses grid-ified data)

To-Add:

* [On visualisation of big geodata] We should carefully generalize, e.g., emphasize the important while removing the unimportant, (Li *et al.*, 2016)
* Reason for: Insight into those spatial and temporal trends can improve the performance of Intelligent Transportation Systems (ITS).(Tayyab *et al.*, 2014)

## 3.5 Modelling:

### 3.5.2 Classification (unfinished)

Model Selection and tuning:

This allows HDBSCAN to find clusters of varying densities (unlike DBSCAN), and be more robust to parameter selection. (https://github.com/scikit-learn-contrib/hdbscan)

Dbscan doesn’t scale well

Before classification:

* Significant class-imbalance exists in the MTL Trajet data (show histogram of classes) so data may need to be under/over sampling.
* Normalise the data between 0, 1
* Comparison of classification techniques
* Grid-search

Types of classification to test:

RF:

“It takes few parameters to create a successful model and the structure of decision tree ensembles avoid overfitting instinctively.”

Ensemble method

﻿Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

SVM:

One vs One

“In one vs one you have to train a separate classifier for each different pair of labels. This leads to *𝑁*(*𝑁*−1)2 classifiers. This is much less sensitive to the problems of imbalanced datasets but is much more computationally expensive.”

<https://stats.stackexchange.com/questions/91091/one-vs-all-and-one-vs-one-in-svm>

NN:

* Basic feed-forward neural network
* RNN for time
* CNN for space (requires gridification of data)

After classification:

* Discuss feature importance
* After Ren *et al.* (2019): do both spatial and temporal distribution of errors

Spatial regression between key areas

difficult to quantify space-time clusters → At what point does a cluster of crimes become a hotspot? (Li *et al.,* 2016)

After Ren *et al.* (2019) does both spatial and temporal distribution of errors

## 3.6 Limitations:

### 3.6.1 Data (unfinished)

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

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

- Difficult to falsify MTL Trajet as is the case with all VGI (Elwood *et al.*, 2012)

- mistaken data due to carelessness, something which especially true if the user is making frequent trips on the MTL Trajet app (Shi *et al.*, 2018)

- This study relies on the initial data handling and cleaning of the MTL project, so unknown unknowns exist. Although, this is something that cannot be rectified as the access to the original data is not available

- Only for 1 month at one specific part of the year (i.e. very specific to mid-sept and Montreal)

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

- 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)

### 3.6.2 Methodological (unfinished)

- Vital to consider all forms of uncertainty in a spatial big data investigation, not something that is thoroughly considered by the methodology and hence the visualisations (Shi *et al.*, 2018)

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

- Mode choice is partially pinned to the weather throughout Montreal (Xie *et al.*, 2016)

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

- Also MAUP and MTUP (Openshaw, 1984; Cheng & Adepeju, 2017).

- Scaling issue and the ecological fallacy

- Higher accuracy of deep learning classification relies on much larger training data sets with labelled classes than conventional classifiers (Shi *et al.*, 2018)

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

- hyper parameter (computational time)

- regularisation of space

- [if STKDE] -> MTUP, MAUP and gridding of data (the regularisation)

- [for LISA markov] assumes that transitions are independent to space (Clark and Rey, 2017)

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

\*\* after Xiao *et al.* (2016) could have taken equal proportions of each purpose instead of cross validation