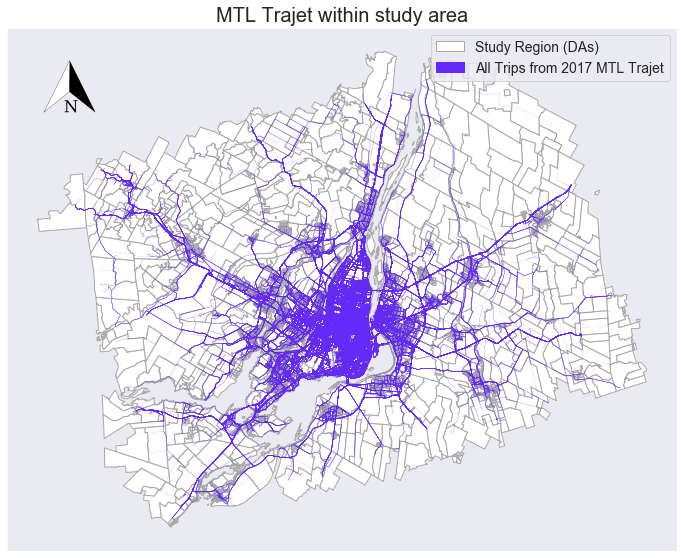
# Chapter 3. Methodology

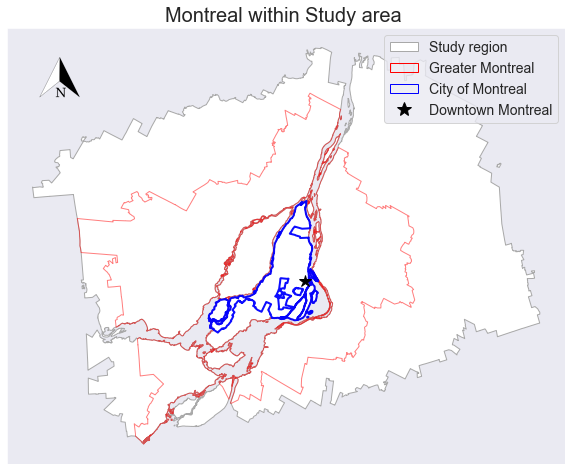
## 3.1 Study Area

The study area chosen for this project spans across 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 the 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 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.

## 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 – 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 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.*

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, 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

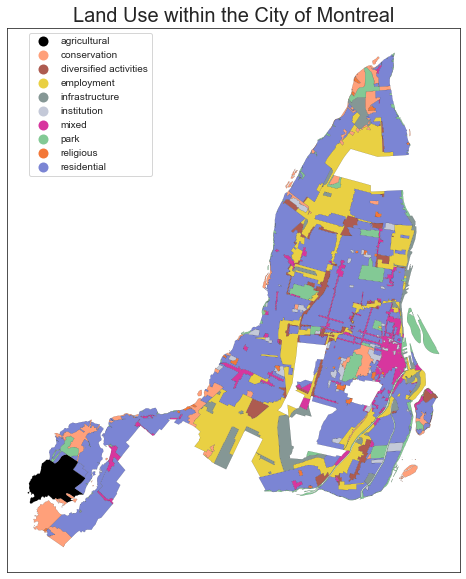
The time signature for the start and end of each trip has been converted from Coordinated Universal Time (UTC) to Eastern Daylight Time (EDT), the Time zone that Montreal falls within, for the purpose of this analysis using Python’s *datetime* library. The duration of each trip was calculated in seconds by taking the difference between these two time signatures. 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 – 18th October 2017. For 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 10 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 follows what is used in the literature (see 2.1) and is primarily to add contextual spatial information to where trips begin and end within the city of Montreal. A *spatial join* is carried out using *Geopandas* to find the category of land use that each trips within the MTL Trajet start and end in.

**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 of this was to supplement the information from the trips, as it has been found in the literature that weather can influence which activites people partake in and even 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 calculating using the *intersects* method from Python’s *Shapely* between each trip and a shapefile of the City of Montreal.

Further binary labels were created to distinguish whether a trip had 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. 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 data 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 2-50, 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 7.15 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).

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.

### 3.4.4 Outlier Detection of Model Inputs:

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.

### 3.4.5 Overview of Model Inputs:

Normalised between 0 and 1

These edge results may not be important for the general classification

After sections 3.2-3.3, a total of 17 predictor variables are used as inputs in the classification models (see **Table 3.5**).

**Table 3.5** Description of the key explanatory variables used within the trip purpose classification models.

|  |  |  |  |
| --- | --- | --- | --- |
| *Column* | *Description* | *Data Type* | |
| *General Explanatory variables* | | |
| 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* | | |
| Trip Distance | Total distance of trip (m) |  | |
| 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* | | |
| Trip Duration | Seconds between start and end of a trip | Ratio | |
| 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)

## 3.4 Exploratory Data Analysis of Trip Purpose and Model Inputs

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 the explanatory variables used to classify them (see **Table 3.5**). It is hoped that these methods serve to assess the feasibility of modelling of each class of trip purpose within the MTL Trajet data (Tayyab *et al.*, 2014).

### 3.4.1 Exploratory Spatial Data Analysis

To assess the level of spatial autocorrelation within each class of trip purpose class in the MTL Trajet, local and global Moran’s I statistics are calculated. To achieve this, the start and end coordinates from each trip have been aggregated into the 7,046 underlying DAs of the study area (see 3.1) using a *Spatial Join* method within *Geopandas*. A Queen’s case contiguity spatial weight matrix using Python’s *Pysal* library has then been computed from this for use in Global and Local Moran’s I (Rey & Anselin, 2007). Local Indicator of Spatial Association (LISA) maps have been created from the values of Local Moran’s I indicating 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 indicate where there are regions across Montreal where different types of trip purposes begin and end.

We then investigate the dispersion of the start and end points of each trip purpose by calculating Standard Deviational Distance using *Pysal*. This statistical measure gives us an indication of how far trips for given purposes are spread across Montreal (Rey & Kang, 2019). Finally, to determine the degree of bi-variate spatial correlation between each given purpose and the model inputs we use the spearman’s rank correlation coefficient. This form of non-parametric correlation coefficient is preferred after initial Kolmogrov-Smirnov tests find long tails in the distribution of each trip purpose.

### 3.4.2 Exploratory Temporal Data Analysis

- ADF, correlation with weather, seasonal decomp, temporal profile, correlation

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. Using this 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.

Capturing long and short term with calendar (Arribas-Bel & Tranos, 2017)

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.

Another part of the ESTDA is the anomaly detection, as we want to discover what is considered normal and abnormal in space time (Li *et al*, 2016)

## 3.5 Classification models

To evaluate the effectiveness of trip purpose classification models, we select three distinct classifiers that will be built using the metrics developed by this analysis:

1. Random Forest Classifier

2. Support Vector Machine

3. Multi-Layer Perceptron

The set-up and description of each are outlined in this section.

### 3.5.1 Random Forest Classifier

* NO NEED TO Normalise the data between 0, 1
* Testing of both a multi class and regular 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.

Two forms of Random Forest methods will be analysed, a regular and multi-class structure.

### 3.5.2 Support Vector Machine Classifier

* Normalise the data between 0, 1
* One vs One and One vs All
* “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>

### 3.5.3 Multi-Layer Perceptron Classifier

* Normalise the data between 0, 1
* Basic feed-forward neural network
* MLP

### 3.5.4 Model Training

Random Forest Classifier for feature importance, above 0.05

Normalise and factorise

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

Train 66% test 33%, this approach was chosen because

Each method will have undergo hyper parameters tuned using a grid-search approach, whereby a range of different parameters will be used.

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

Further we divide the data into four subsets from which all models are built upon. The purpose of this is to suggest

These binary labels will form the basis of 5 unique subsets, detailed in **Table 3.6**, from which classification models will be built.

|  |
| --- |
| Overall |
| City |
| Non-City |
| Rush Hour |
| Off-Peak |

After Classification:

* Feature importance

### 3.5.5 Model Outputs

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

Specifically, LISA maps indicate disconformity of the residuals in the LSOAs i.e. where the

regression had overestimated (‘hot’ spots) and underestimated (‘cold’ spots) this relationship (Anselin, 1995).

## 3.6 Limitations:

### 3.6.1 Data

- 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

- 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 ().

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

- 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

challenges with representing big data in one map (Robinson *et al.*, 2017)

Zhao *et al.* (2019) - The temporal sampling interval (TSI), which is measured by the temporal interval between consecutive records, determines how well such data can describe human activities and influence the values of human mobility indicators.