Literature reveiw:

Flanagin & Metzger (2008) find that VGI, gives us information that only ‘locals’ know.

Finally, there is a

We select a defined area across the region of Montreal based on , so we ignore any consideration into any influence or ‘*edge effect*’ that area outside this study area have. Then again, we somewhat avoid a severe degree MAUP as the study area is selected at the maximum extent of the GPS trace from the MTL Trajet data and no physical trips are taken outside this region (**Figure 3.1**).

Although, some evidence of spatial clusters being used as explanatory variables (e.g. Montini *et al.,* 2014; Yazdizadeh *et al.*, 2019), less attention paid to temporal clusters. Other forms of mobility research, have also seen less consideration to temporal clusters, although a 2018 study by Liu & Cheng adapt a Latent Dirichlet Allocation model to better account for temporal structures in movements from smart card data.

Montreal has a relatively high share of transit ridership (for a North American city) Also multimodal public transport network (Eluru *et al.* 2012)

Eluru *et al.* (2012) survey to look at public transport and how people move around Montreal. Trying to encourage it

The city of Montreal itself is the largest city within Quebec and the second largest within Canada ()

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

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

*Montreal:*

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

-WPR (2019) -> 1.75 million people as of 2016. The island in total has an estimated population of 1.95 million. “The city proper has a population density of 4,517 people per square kilometer (11,701 residents per square mile)”. “By 2030, the Greater Montreal Area is expected to grow to 5.275 million”

-The city of Montreal, located within the Greater Montreal region is the largest city in Quebec with a population size of … . It is of particular interest to city transport research due to its unique road and public transport networks. There are a total of

“Montreal with its unique multimodal public transportation system consisting of bus, metro and commuter train offers multiple transit route alternatives to individuals commuting to downtown" Eluru *et al.* (2012)

an region containing around 4 million people (WPR, 2018; **Figure 3.1**).

Data collection [for survey data] driven by data availability or convenience of data collection rather than by domain knowledge, theory, or insight into the process(es) of interest. (An *et al.*, 2015).

[On visualising big data maps] New computational and technical paradigms for cartography are accompanying the rise of geospatial big data (Robinson *et al.*, 2017).

Big geospatial data visualisations do not always scale well – because they can become messy (Li *et al.*, 2016)

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

(Arribas-Bel & Tranos, 2017), (Bricka *et al.*, 2015)

Indeed, understanding the space and time structures in cities help us understand them better (Chen*g et al.* 2017).

New nexus, between data science and VGI (Burini *et al.*, 2017)

“Getting a deeper understanding of human mobility is a prerequisite for a broad range of possible studies on smart cities and related research areas”. (Xie *et al.*, 2016)

high quality traveller information for public transport is undoubtedly crucial for Government’s transport policies (Lyons & Harman, 2002)

Data collection [for survey data] driven by data availability or convenience of data collection rather than by domain knowledge, theory, or insight into the process(es) of interest. (An *et al.*, 2015).

Finally, crowd sourced data often needs to be repurposed to fit the needs of research (Li *et al.*, 2016). Therefore,

Developed onto of the success of the DataMobile Smartphone Travel Survey example from 2014 carried out by Concordia University in Montreal (Patterson & Fitzsimmons, 2016) 10 November - 5 December 2014. Close to 900 people participated in the survey [Only around the univirsity]. Based in Montreal

This has meant we have been able to identify processes of movement in more representative samples of populations (Miller & Goodchild, 2014).

For example, GPS tracts picked up from smartphone travel surveys may indicate detailed travel modes and prp

if we knew that people tended to cycle to cafés during lunch breaks, policy could be implemented to introduce bike racks near the cafés. Notably, VGI has already seen use in real time transport demand management (Bricka *et al.*, 2015)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Macro average | 0.34 | 0.27 | 0.26 | 13375 |
| Weighted av | **0.60** | **0.67** | **0.61** | 13375 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Macro average | 0.37 | 0.25 | 0.23 | 14421 |
| Weighted av | 0.63 | 0.69 | 0.61 | 14421 |

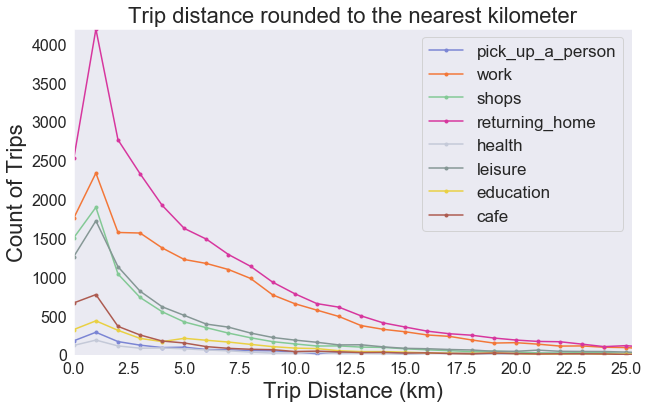
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Macro average | 0.12 |  |  | 23695 |
| Weighted av | 0.30 |  |  | 23695 |

After outlier removal

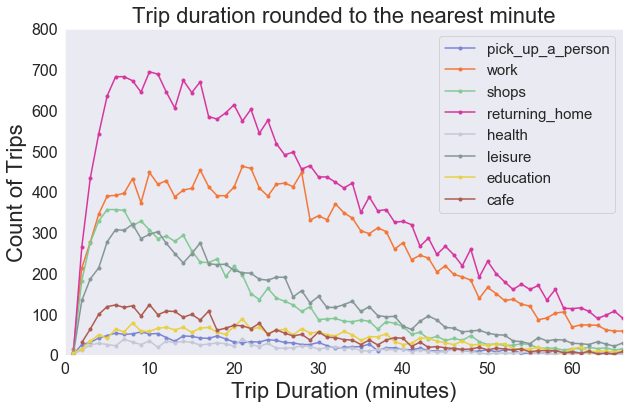
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **std** | **min** | **25%** | **Median** | **75%** | **max** | **N** |
| **Distance (m)** | 6634 | 9927 | 50 | 840 | 3147 | 8092 | 99810 | 177938 |
| **Duration (sec)** | 1537 | 1285 | 60 | 616 | 1204 | 2081 | 10799 | 177938 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wolf *et al.* (2001) | Distance; duration; land-use matching | Georgia, USA 3 days  ﻿March 23 through April 28, 2000 | 13 (156 trips) | 93% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probabilistic Methods: Multinomial Logit Model | | | | |
| Oliveria *et al.* (2014) | Duration; Mode; Land use; Personal locations proximity | Georgia, USA 2011 | 1,354 (﻿10,512 trips) | 70% (95-97% for work and school) |

****

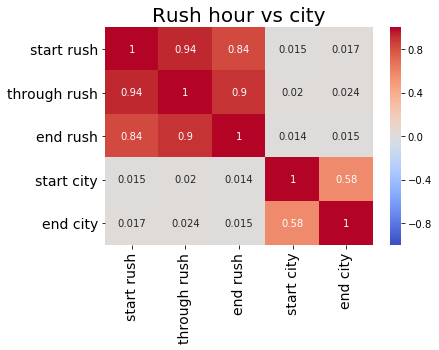
**Figure 4.X**

****

**Figure 4.X**

Overall, the rush hour and city labels are found to exhibit statistically significantly (*p-value<0.05*) positive correlation across the trips, as determined by a spearman’s rank correlation coefficient (ρ) (**Figure 4.8**). Notably, there is no discernable correlation

we see weaker positive correlation between trips that begin and end in the city.



**Figure 4.X** Correlation matrix (statistical significance below a 0.05 confidence interval indicated by white text; *n=73,029*)

KS test vs norm for grouped 1 hour

cafe KstestResult(statistic=0.5551535507713675, pvalue=3.287213099129875e-205)

education KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

health KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

leisure KstestResult(statistic=0.7251252221594695, pvalue=0.0)

not\_available KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

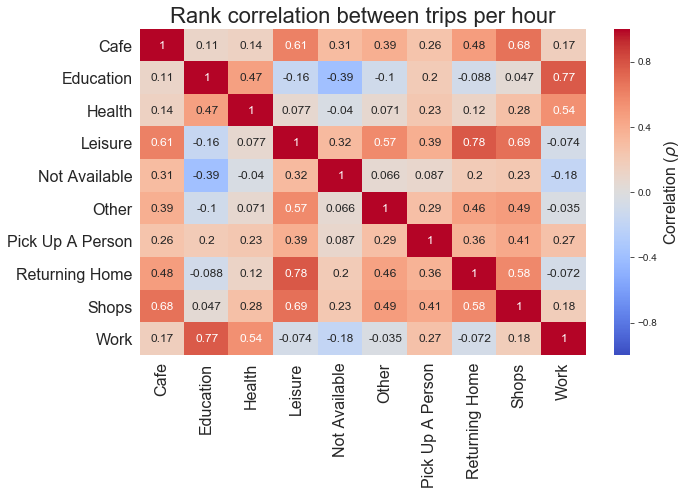
other KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

pick\_up\_a\_person KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

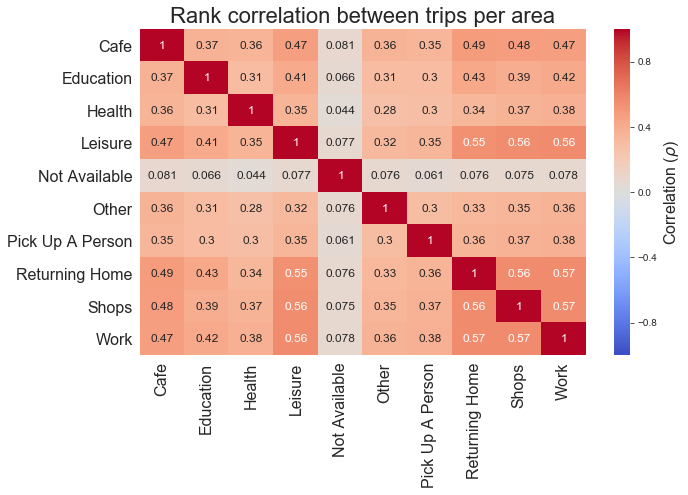
returning\_home KstestResult(statistic=0.8370232391566367, pvalue=0.0)

shops KstestResult(statistic=0.6061450521877981, pvalue=3.586226223248267e-249)

work KstestResult(statistic=0.749204542272784, pvalue=0.0)

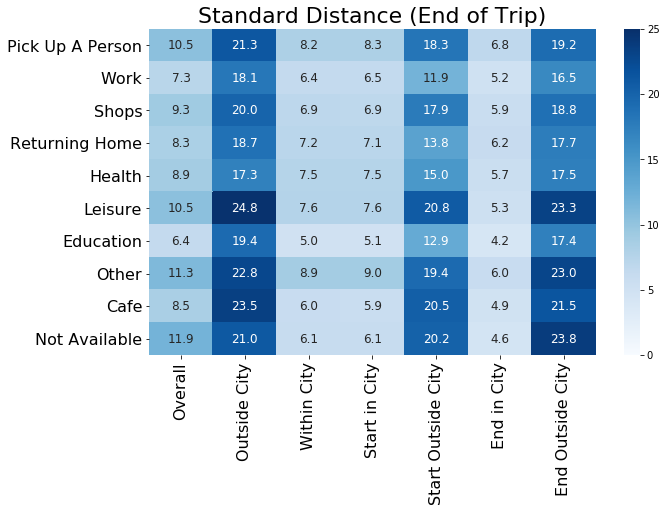
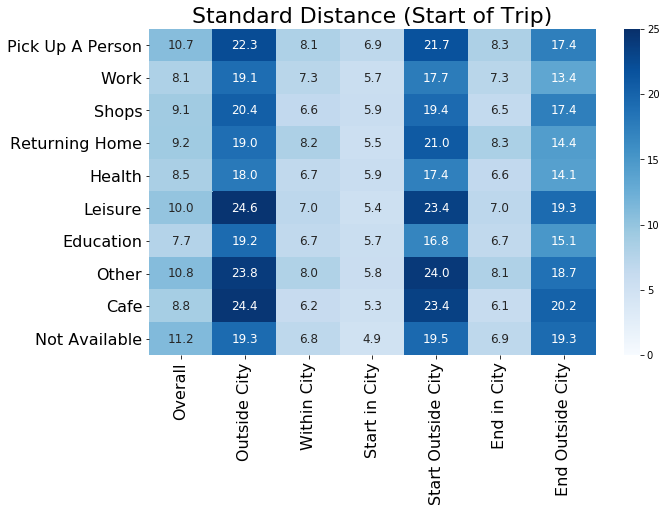


**Figure 4.X**



**Figure 4.X**

Standard distances between each unique trip purpose and



**Figure 4.17** Heatmap plots showing the standard distance between start and end points of the trips grouped by individual trip purposes.

For a measure of spatial dispersion within the model inputs, the Standard Deviational Distance is calculated using *Pysal* on all the start and end points of the trips. 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 spatial correlation between the trip purpose classes we use a spearman’s rank correlation matrix to compare the count of each purpose in each of the 7,046 DAs.

* This non-parametric correlation co-efficient is preferred after initial Kolmogrov-Smirnov test suggest the

New Purpose labels:

Based on correlation

"1":["shops","leisure","cafe","returning\_home"],"2":["education"],\

"3":["health"],"4":["pick\_up\_a\_person"],"5":["work"]}

1 47464

5 18950

2 2769

4 1574

3 1044

Based on discovery of trip purposes that share spatial and temporal dependencies we further subset the 10 unique trip purpose classes into . We also opt to remove trip classified as ‘Other’ and ‘Not Available’. It is hoped that these will improve the performance of the classifiers (Jahromi *et al*., 2016).

Table of New Groups:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Label | Purpose Types |  |  |  |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |

Extra:

The improvement of mobile networks and the geolocation abilities of smartphones in recent years, means we currently have an opportunity to study cities through data resulting from mobile GPS traces (Li *et al.*, 2016; Patterson *et al.*, 2019).

Mobile phones as sensors (ref)

Extra:

Respondents:

B & M: 1104

Oliv 1,354

Xiao 321

Mont 156

Kim 793

Yaz 6845

Zhu 10,372

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.

### 3.4.3 Spatio-Temporal Investigation of Model Inputs

- directionality throughout day

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

- space-time interaction tests in pysal

- 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).

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)

Sentences:

The casual links have been harder to prove in the literature

Owing to the…

To Add:

* This has affected probabilistic models over machine learning models (Oliveria et al., 2015)

Arguably, there is a need for more encompassing approach to the modelling inputs has allowed more generic model creation that can applied to a range of different purposes (as we do not know which purposes rely on which i.e. POI may be important for tourist purpose but not for work purposes).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gong *et al.* (2018) | POI; Socio-demographics; Temporal Features; Trip Duration; Weather | ﻿Hakodate City, Japan 2012–2013 | (﻿9981 trips) | n/a |

Spatial-Temporal:

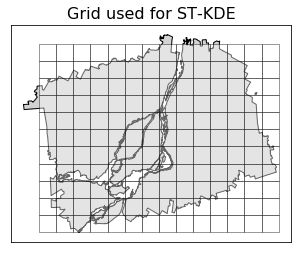
ST-K

Space-time Ripley’s K and spatial-temporal randomness -> end to start

[One STKDE Ripley’s K]

Space-time KDE end to start

ST-KDE



**Figure 4.X**

[Some sort of plot for KDE]

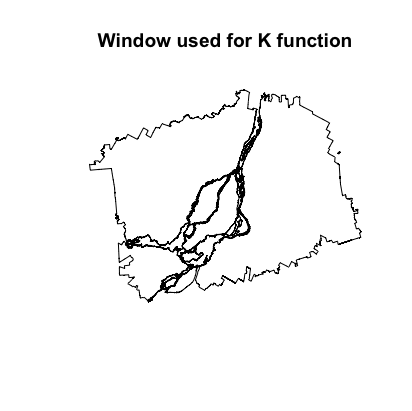
Join the ST-KDE back to the data -> if there a trip at a particular time and space (a grid region) can be flagged as in the given cluster (0,1)

11 columns of STKDE cluster

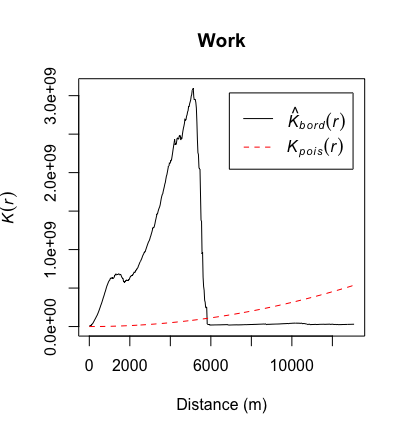
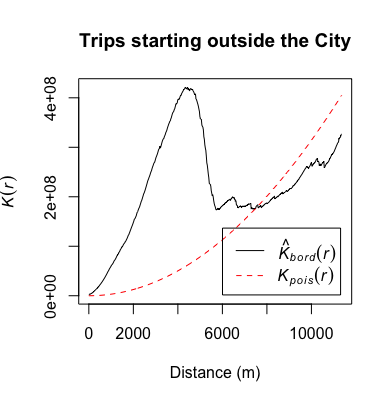
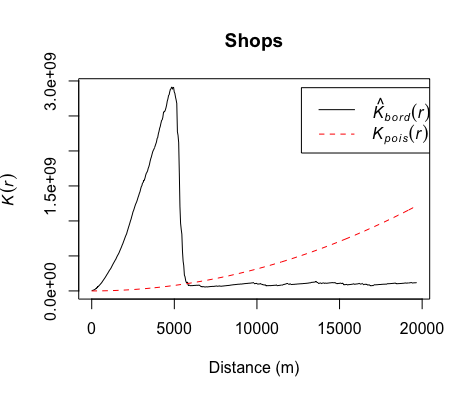
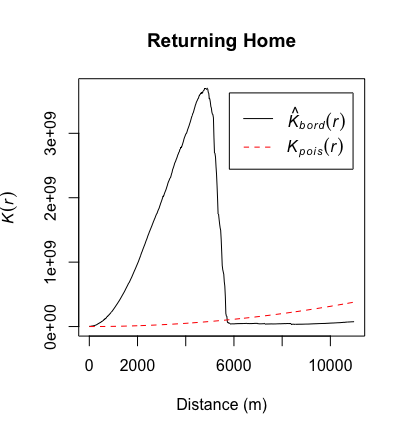
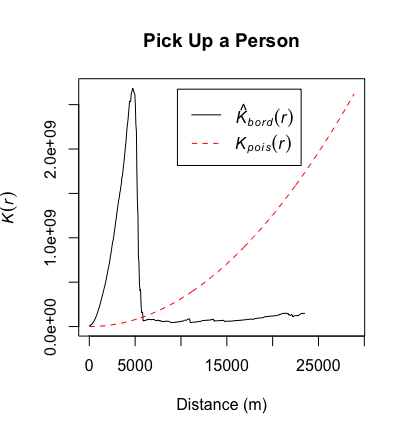
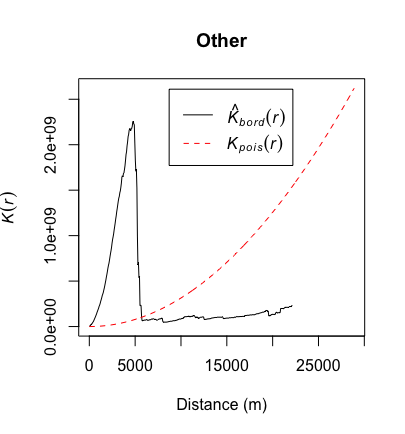
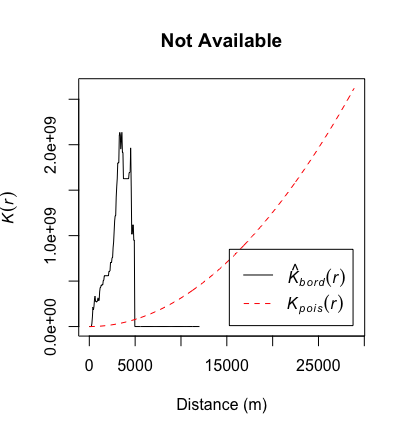
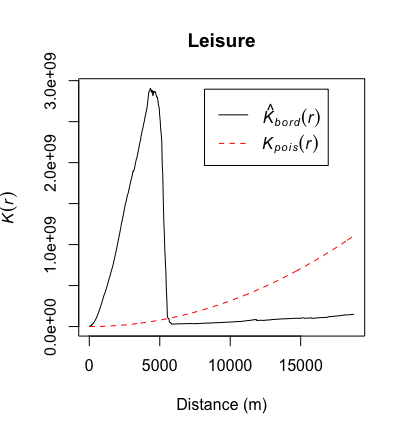
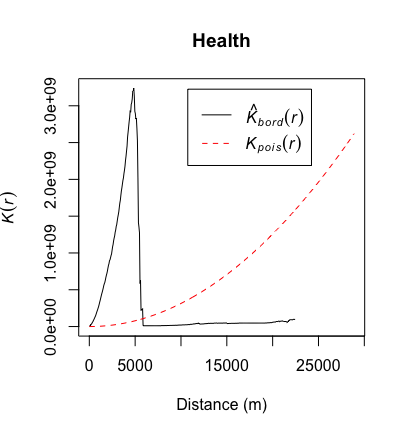
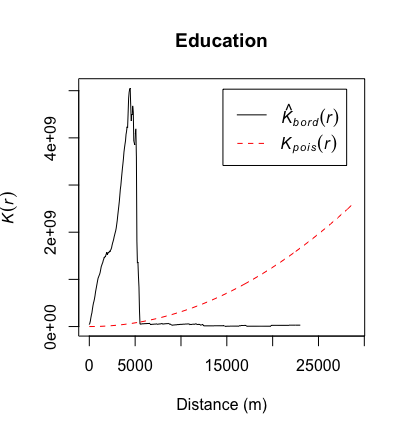
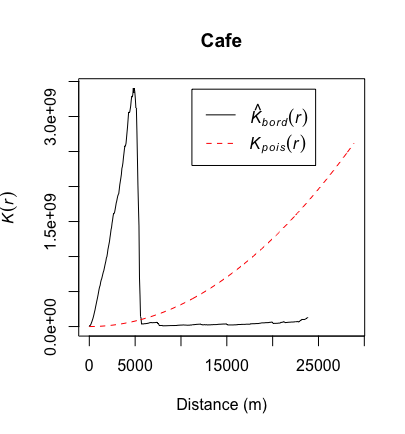
For a given trip, dissemination areas within the intersection are given a value of 1 (red) and all other areas are given a value of 0 (grey). For the spatial analysis in this report, data has been grouped using *geopandas groupby* function into individual modes and purposes of travel (see **Table 3.1** 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**)

Ripley’s K on start and end -> spatial randomness



**Figure 4.X** Window used for Ripley’s K function



“Using this window, a points pattern object is created from the Treasure Hunt location coordinates and estimations for Ripley’s K is calculated. This estimation of K is plotted in Figure 5 against a Poisson distribution (which indicates Complete Spatial Randomness of Treasure Hunt locations). As shown, the trend in estimated K is far higher than the Poisson distribtion implying there is clustering and spatial dependence of Treasure Hunt locations across London.”

We assess spatial dependency of trips from each type of trip purpose using Ripley’s K-function calculated using the R’s *spatstat* library. Specifically, the Ripley’s K can be used as a measure of spatial distribution of start and end points of the trips indicating whether there is Complete Spatial Randomness within the data (Dixon, 2002). These

Moreover, we detailed in **Table 3.6**. we divide the data into four distinct subsets from which all models are trained, The purpose of this is to evaluate the disparity in performance within the model when only looking at one broad spatial or temporal category.

**Table 3.6** Subsets used to evaluate classification model accuracy

|  |  |
| --- | --- |
| *Subset Name* | *Description* |
| City | Trips starting o ending in the city |
| Non-City | Trips starting and ending outside the city |
| Rush Hour | Trips starting or ending in the Rush Hour |
| Off-Peak | Trips starting and ending in Off-Peak |

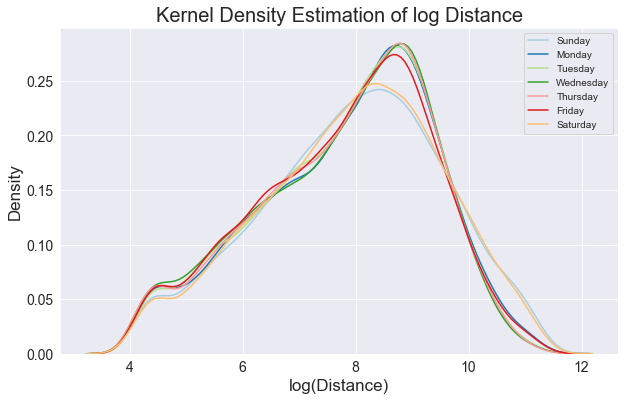
Individually, decision trees represent the probability of one of a set of outputs given an input condition, in RF classification models, decision trees are used in combination with (; find other ref)

Boostrap aggregation (or bagging)

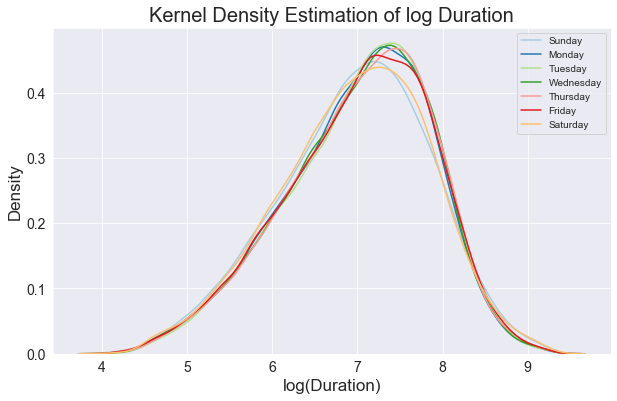
The data is hence, normalised for the purpose of the classification models using by taking the natural log. **Figure 4.3+4.4**

Log distance skew: -0.37300090670060865 kurtosis: -0.5147541361828902

Log duration skew: -0.38136680556957275 kurtosis: -0.21730048341457042



**Figure 4.3** 1-Dimensional Kernel Density Estimation plot of trip distances.



**Figure 4.4** 1-Dimensional Kernel Density Estimation plot of trip durations.

An ordinary least squares (OLS) regression is carried out to evaluate the change in log duration (y) as determined by the change in log trip distance (x). We see that there is coefficient of +0.386 log meters (coefficient tells you the size of the effect of x on y).

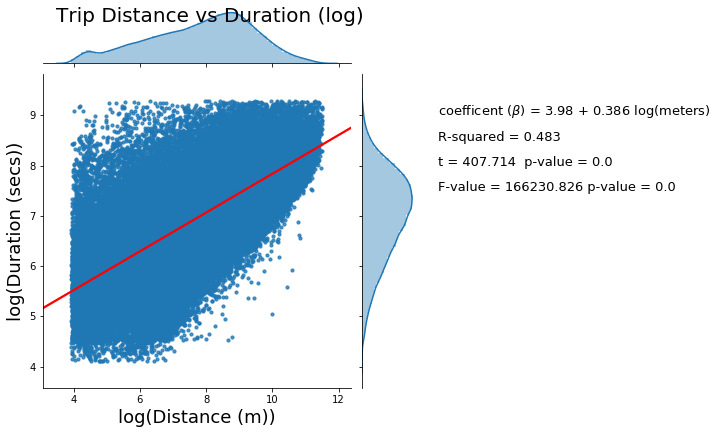
- relatively strong relationship, how well the line fits the data (R2=0.483)

- t-value is significant indicating difference in means,

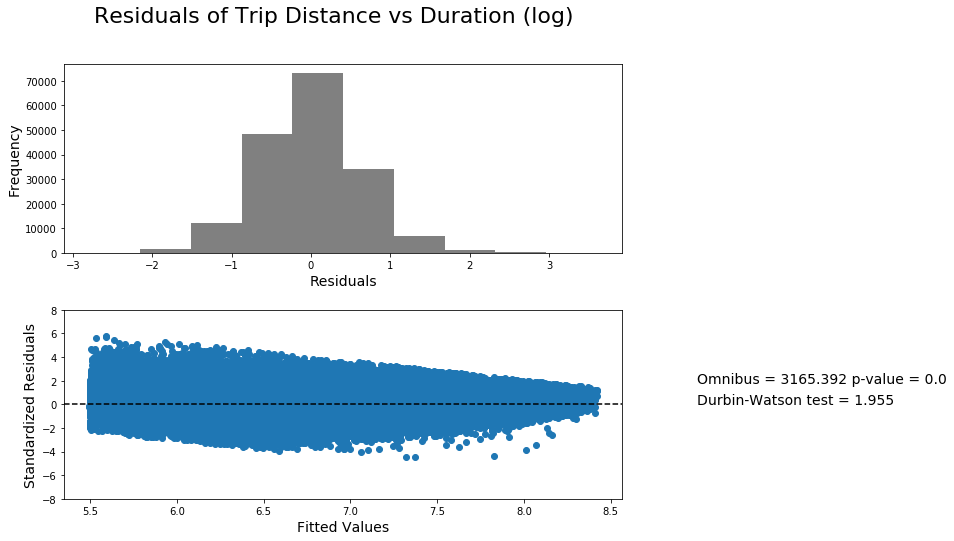
- F-value is probability the model is wrong (indicates the probability that all the coefficients in our regression output are actually zero)

- may need ref

Here, the residuals from the regression model were found to show non-normality (Omnibus below α=0.05) and



**Figure 4.X** OLS Regression results



**Figure 4.X** OLS Regression residuals

Breusch-Pagan test for Heteroscedasticity:

'B-P Test Statistic': 8757.14320669015, 'B-P Test p-value': 0.0, 'F-Statistic': 9210.327120693702, 'F-Test p-value': 0.0

Result shows: The data is heteroscedastic

“The Breusch-Pagan tests affirm that the residuals show no statistically significantly (p<0.05) heteroscedasticity, and thus are spatially random”

# Appendices

## Appendix 1 Notification not to apply for Ethical Approval

## Appendix 2 Python Script used for retrieval of Climatological data:

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

# Appendix 3 Mean Direction and Distance Calculations

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.)*

# List of Figures

**Figure 1.1** (A) Screenshot from the MTL Trajet app showing recorded GPS trace (source: Patterson, 2017a). (B) Example of prompt similar to one used in the 2017 MTL Trajet app (source: Patterson et al., 2019………………………………………………………………….………………..……….1

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

**Figure 2.2** Example of ‘Map Matching’ done by the Open Source Routing Machine when processing the raw GPS trace from user devices (Source: Hamouni, 2018)…………………..………2

**Figure 2.3** Example of GPS trace with location collection priorities within an Itinerum platform app (A); Example of the on screen prompt after an Itinerum platform app stops recording movement (B) (Source: Patterson et al., 2019)……………………………..………………………3

**Figure** **3.1** GPS routes from the 2017 MTL Trajet travel survey plotted within the study area………………………………………………………………………………………………………………………………………4

**Figure 3.2** Location of Montreal within the study area…..…………………………………………………….5

**Figure 3.3** Map showing land use categories within the City of Montreal (data from: Ville de Montréal, 2014).………..…………...…..………..…………...…..………..…………...…..……….………………...….5

**Figure 3.4** Example of an eastbound trip across Montreal (trip-id= 150744)………………………..5

**Figure 3.5**Example of temporal profile………..………..………..………..………..………..………..…………..6

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

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

**Figure 4.1** Boxplots (top), Kernel Density Estimation (middle) and Quantile-Quantile (bottom) plots showing the distribution of distance and duration of trips from the 2017 MTL Trajet travel survey………..………..………..………..………..………..………..………..………..………..……………8

**Figure 4.2** Bar charts showing the type of trip purpose and travel mode selected by respondents to the 2017 MTL Trajet survey………………….………………….………………….………………9

**Figure 4.3**Bar chart comparing the proportion of each unique trip purposes accounted for by each unique travel modes……………………………………………………………………………………………..10

**Figure 4.4**Map showing the mean direction of trip within each region of Greater Montreal……………………………………………………………………………………………………………………………11

**Figure 4.5** Circular contour plot showing the mean direction of trips for each trip purpose.12

**Figure 4.6** Bar chart showing the proportion of trips carried out during rush-hour and off-peak as grouped by purpose………………………………………………………………………………………………13

**Figure 4.7** Bar chart showing the proportion of trips carried out within and outside the City of Montreal as grouped by purpose…………………………………………………………………………………...14

**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)………………………………………………….15

**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…………………..…………………..…………………..…………………..…………………..………….16

**Figure 4.10** Line graph comparing sum of squared distances and silhouette scores of k-means clustering algorithm for k between 2-20……………….……………….……………….………………17

**Figure 4.11** Map of origin and destination points from the MTL Trajet trips coloured by cluster label across the study region…………………………………………………………………………………..18

**Figure 4.12** Bar chart showing number of trips per spatial cluster identified by the k-mean clustering algorithm……………………..……………………..……………………..……………………..………………19

**Figure 4.13** Line graph comparing coherence score and log perplexity of LDA models using a topic count of between 1-12.…………………..…………………………………..……………………………………20

**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**)……………………..……………………..……………………..…………………21

**Figure 4.15** Inter-topic Distance map between each of the Five Temporal Clusters identified by the LDA model………………………………………………………………………………………………………………22

**Figure 4.16** Count of trips associated with each temporal cluster identified for this analysis…………………………..…………………………..…………………………..…………………………..…………..23

**Figure 4.17** Line plot showing the amount of recorded trips taken from the MTL Trajet app between 18th September 2019– 18th October 2019 (weekends indicated in **purple**)………….24

**Figure 4.18** Average trip distance and duration represented as a percentage of the mean..25

**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…………………………….……….26

**Figure 4.20** Time series plot showing the average temperature (in Celsius) and precipitation (mm) recorded during the study period………………………………………………………………………..…..27

**Figure 4.21** Temporal de-composition of the count of trips recorded by the MTL Trajet travel survey at 24-hour lags………….………….………….………….………….………….………….………….………….28

**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…………………………………………………………….29

**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)…………………..30

**Figure 4.24**Feature importance from a Random Forest Regression model subset MTL Trajet dataset (the red line indicates model features which will be removed)………..…………………..30

**Figure 4.25**Comparison of precision, recall and f1-score across the three types of classifier……………….……………….……………….……………….……………….……………….……………….….…..31

**Figure 4.26**True vs predicted purpose confusion matrices for each classifier……………….……32

**Figure 4.27**True vs predicted purpose confusion matrices for each classifier (as % of total predictions)……………………………..……………..……………..……………..……………..……………..…………….77

**Figure 4.28** Bar plots comparing the cross-validation accuracy………………………………….……….77

**Figure 4.29** Matrix showing the amount of correctly identified trips in each one of the models compared to the others…………………………………………………………………………………………78

**Figure 4.30** Destination of trips that are classified correctly (left) and incorrectly (right) by all the classifiers across the study area……….……………………………………………………………………….…78

**Figure 4.31** Amount of correctly (top) and incorrectly (bottom) classified trips by each of the classifiers across the study period..…….………………………………………………………………………………78

# List of Tables

**Table 2.1**Overview of classification models used in the literature to predict trip purpose (POI=Points of Interest).……………………………………………………………………………..………………………..1

**Table 3.1** Description of the variables from data from the MTL Trajet survey before pre-processing………………………………………………………………………………..……………………..…………………..2

**Table 3.2** Categories of travel mode and purpose responses allowed for trips in the MTL Trajet travel survey………………………………………………………………………………………………………………3

**Table 3.3** Description and cover of Land Use categories within the City of Montreal……………4

**Table 3.4** Definition of Rush hourand Off-peak hours used in this study……………………………..5

**Table 3.5** Description of the key variables from the MTL Trajet survey after pre-processing..6

**Table 4.1** Outline of trips removed from the analysis……………………………………………………………7

**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)…………………………………………………….8

**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)………………………………………10

**Table 4.4**Results from the application of Rush-hour and City labels to the data….……….......12

**Table 4.5** Outline of trip purposes associated with each temporal cluster found by a 5-topic LDA model.…………………………………………………….…………………………………………………………………..13

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

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

**Table 4.8** Subsets used in the building of classification models………….………….………….……….14

**Table 4.9** Overall Accuracy in the models………….………….………….………………….………….…………15

**Table 4.10** Results from the classification broken down by class of trip purpose (values above 0.5 are shown in **bold**).….………….………………….………….……………………………………………..15