4000 words

What I am trying to show:

* Idea about the MTL Trajet
* What specs for purposes (i.e. which modes, where, etc.)
* Space, time and space-time trends
* Classification results (how well we can classify purpose and most important things)

# Results

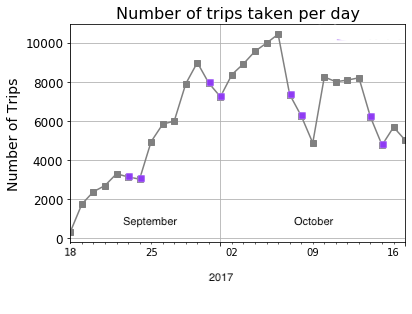
*Introduction:*

This section is divided into three sections, the first (4.1) examines general trends in the data (detailed in section 3.2) and identifies key areas of analysis, the second (4.2) uncovers space, time and space-time structures and interdependencies which are useful for modelling and the third details the results of the classification models used to classify purpose of the trips (4.3).

*4.1 Preliminary analysis:*

*Data statistics:*

A total of 185,285 trips were recorded across the study period (18th September 2017– 17th October 2017), but there is significant variation in the amount of recorded trips per day. As shown in **Figure X**, during the first 7 days of the study less than 4000 trips were recorded per day compared with more than 4000 trips in the remaining days (with the most amount of trips being recorded on Fridays). Here, less trips are recorded on weekends versus weekdays, other than the Monday 9th October, which was the day Thanksgiving was celebrated that year in Canada.



**Figure X** 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** ; data from PDO, 2017).

*Distance & Duration & Outlier Removal:*

After calculating the distances and duration of the individual trips (3.2.X), our analysis finds 7,594 trips that are less than 50 m or more than 100 km in distance and less than 60 seconds or more than 3 hours in duration, leaving 177,938 trips used for this analysis. As shown in **Table X**, the majority of these were from trips that were less than 50 m in length. These trips are potentially from cases where the app had switched on for slight movements, as it is noted in ref for this problem… (ref)

**Table X** Trips removed from the analysis

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

The resulting trips are shown to

We see that the standard deviation for distance is above the mean, suggesting that the data does not cluster around it.

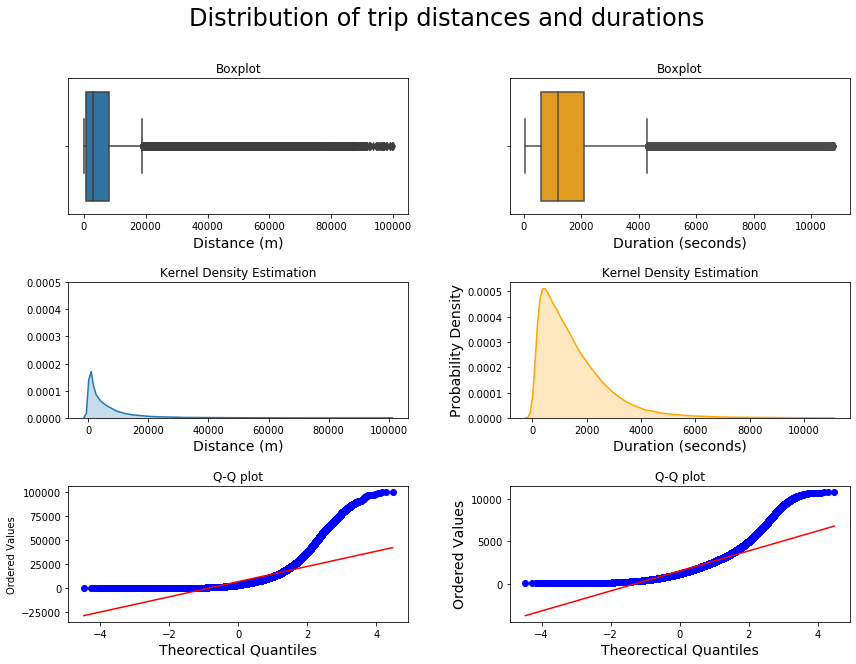
- 95% of the trips are around 25 km and 65 minutes

- relatively closer median to mean for duration

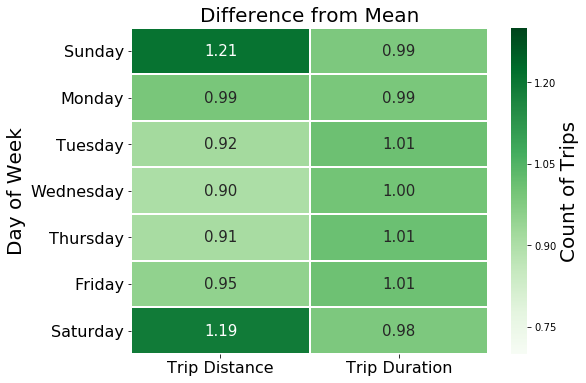
After outlier removal (converted to km and minutes; N=177,938)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *mean* | *STD* | *min* | *25%* | *Median* | *75%* | *95%* | *max* |
| *Distance (km)* | 6.63 | 9.92 | 0.05 | 0.84 | 3.14 | 8.09 | 25.25 | 99.81 |
| *Duration (min)* | 25.62 | 21.42 | 1.00 | 10.27 | 20.07 | 34.68 | 65.81 | 179.98 |

Both trip distance and duration are shown to be heavily positively skewed with both showing long tails in their distribution. A large disparity between the mean and median in this case may be indicative of a long-tailed distribution (see Table X.). Distance is more skewed with a longer tail. The data is hence, normalised for the purpose of the classification models



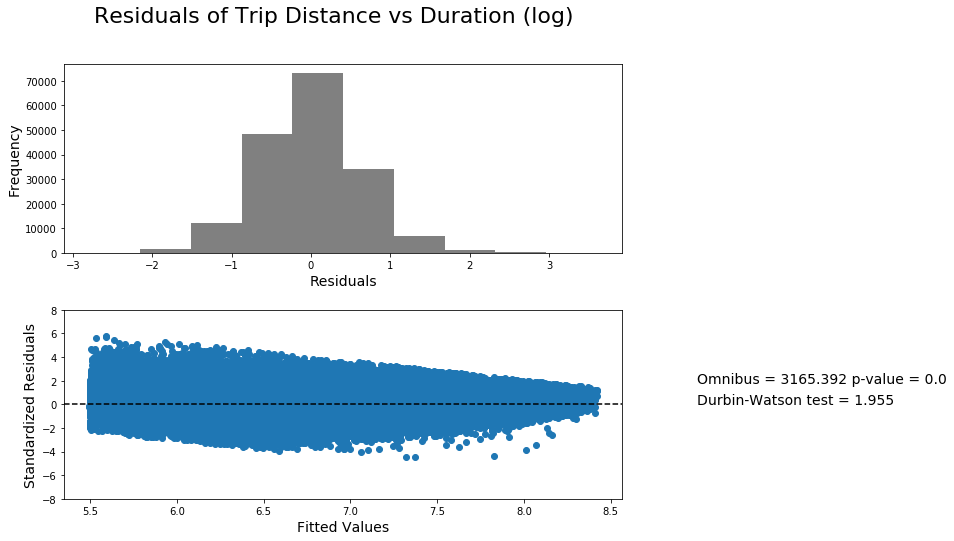
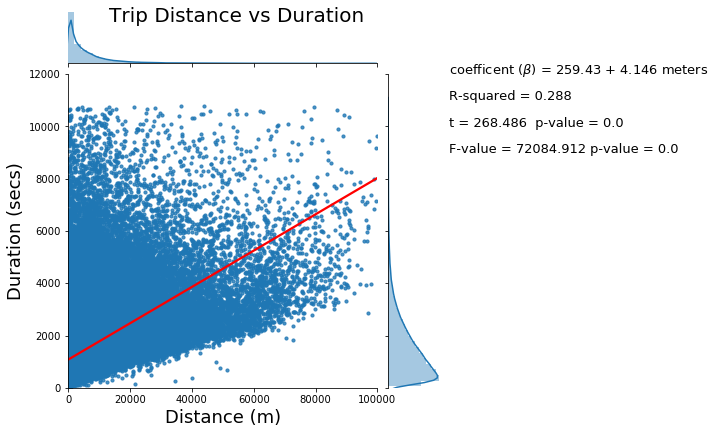
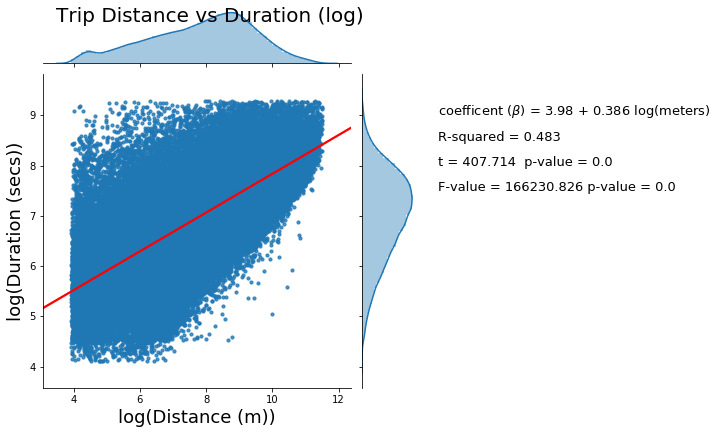
Across the week, on average, we see that more trips of longer distances compared to the overall mean are travelled on weekend versus during the week, arguably this could result from the influence of work hours, with people travelling further into rural areas during weekends. Notably, there is no deviation from the mean travel duration across the week, on average in the trips.



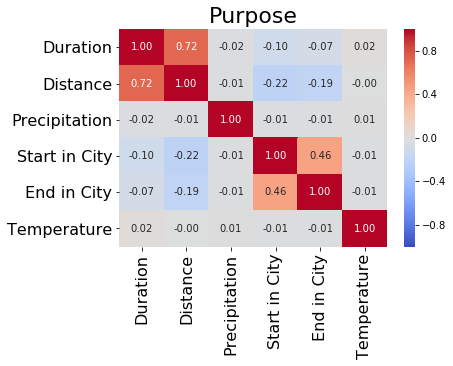
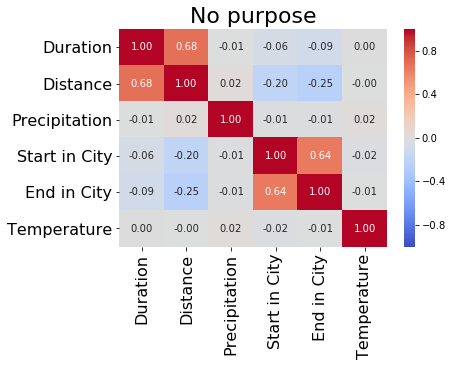
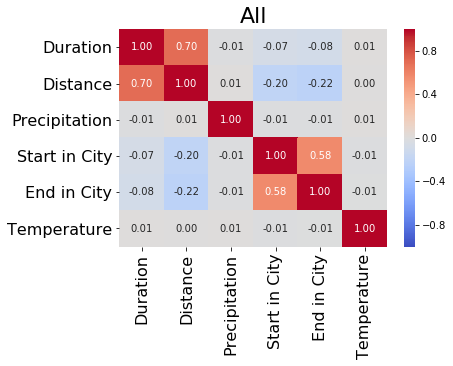
**Figure X** Distance and duration

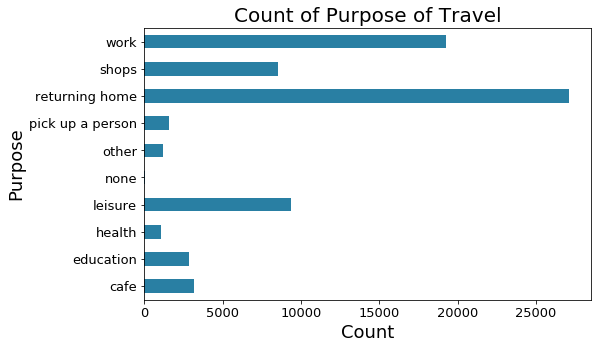
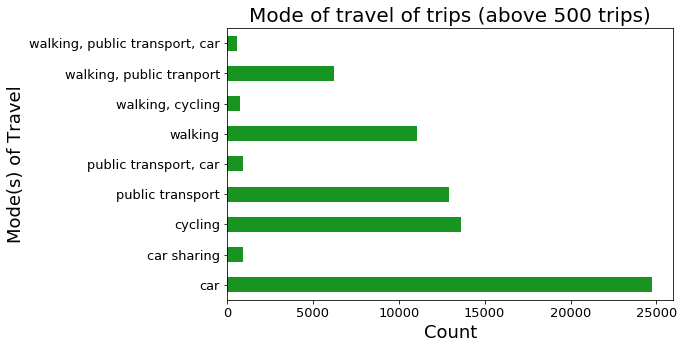
See: <https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/> for interpreting the regression

An ordinary least squares (OLS) regression is carried out to evaluate the change in duration (y) as determined by the change in distance (x)

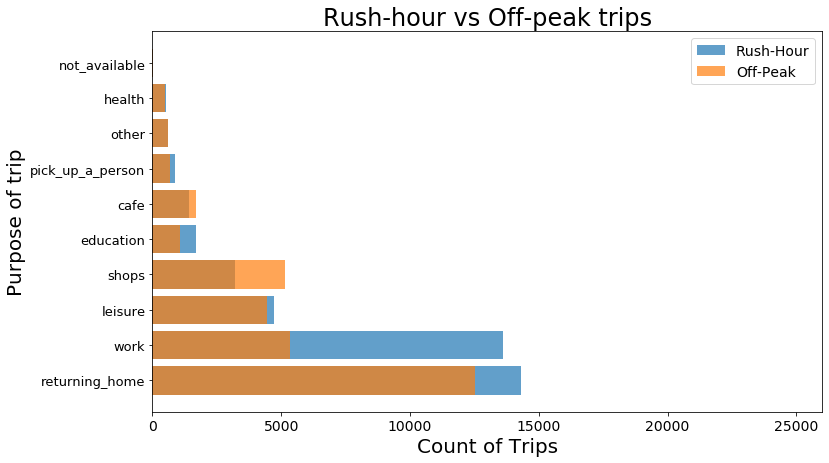
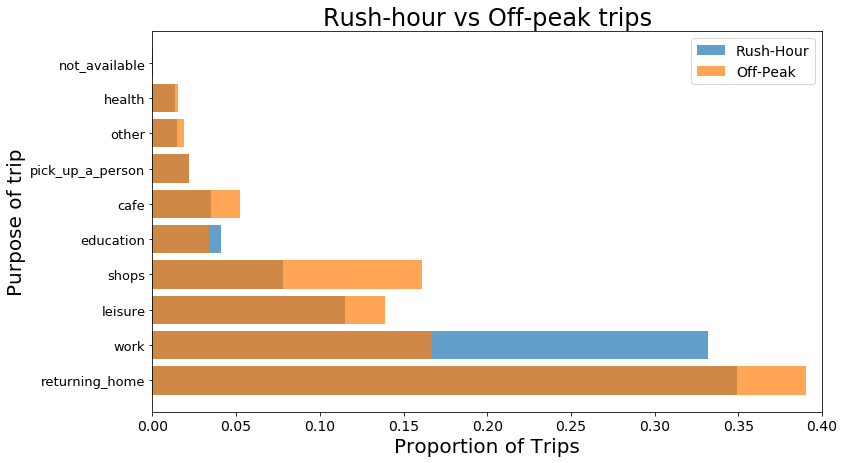
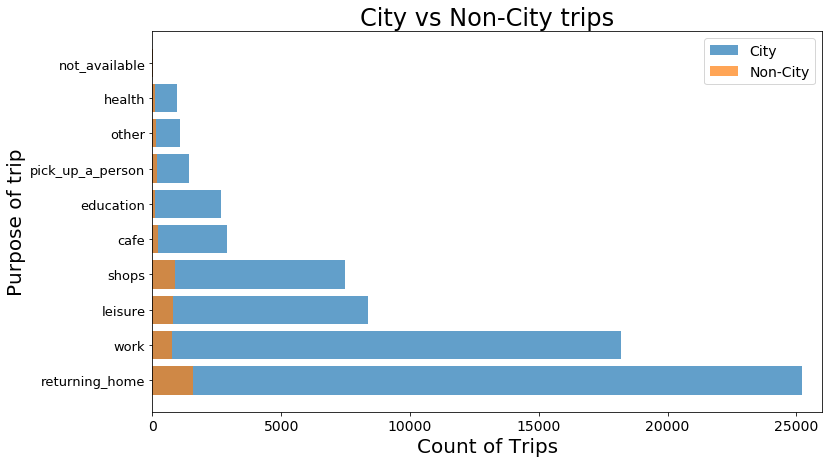
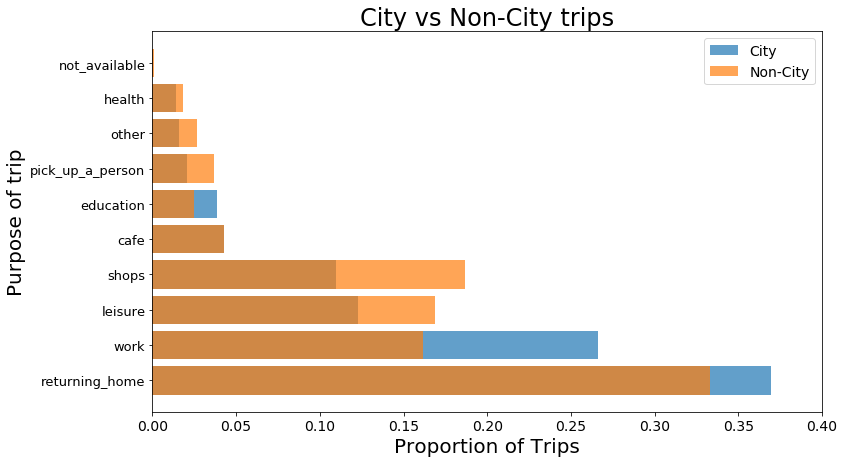


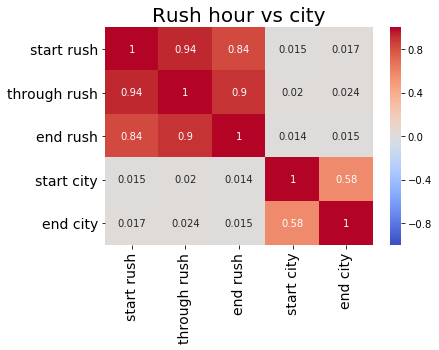
*Mode & Purpose:*





*City and Rush hour:*

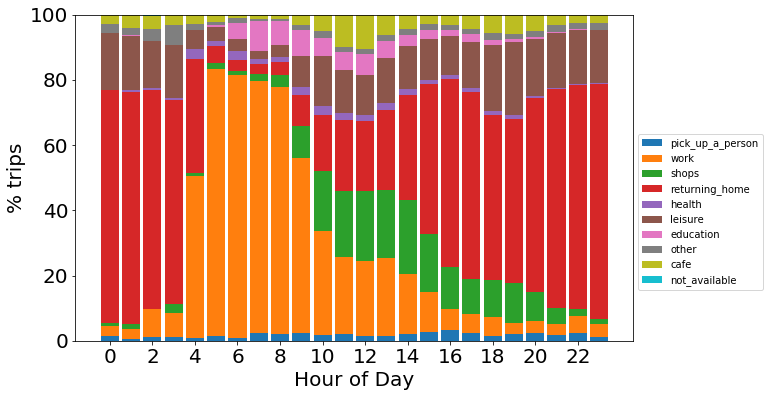
**

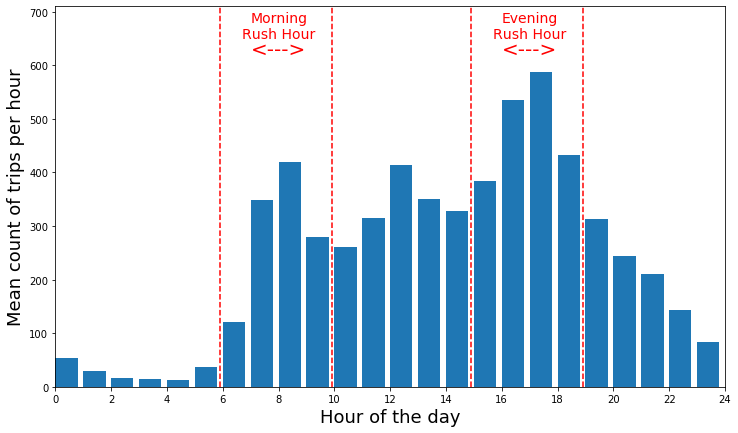


*4.2 Exploratory Space-Time Data Analysis (ESTDA):*

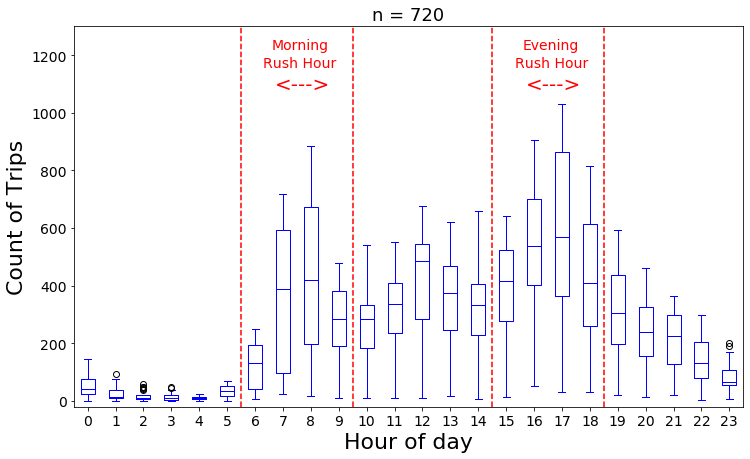
For this report, we divide this section into three main parts: spatial, temporal and spatial-temporal methods used to discern signal from the data.

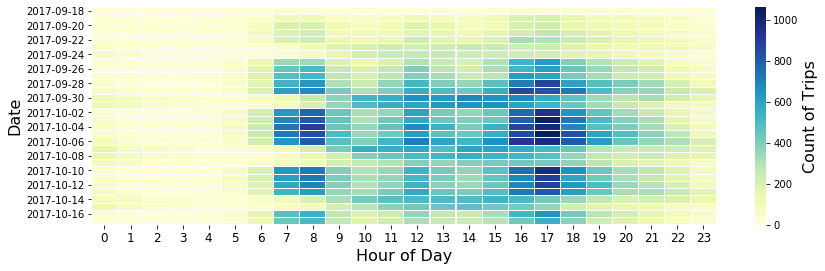
Temporal Analysis:

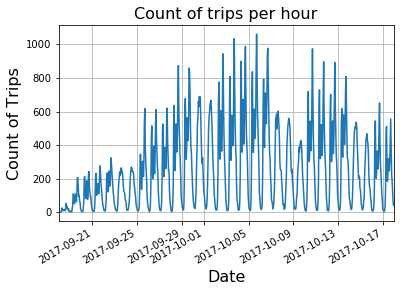




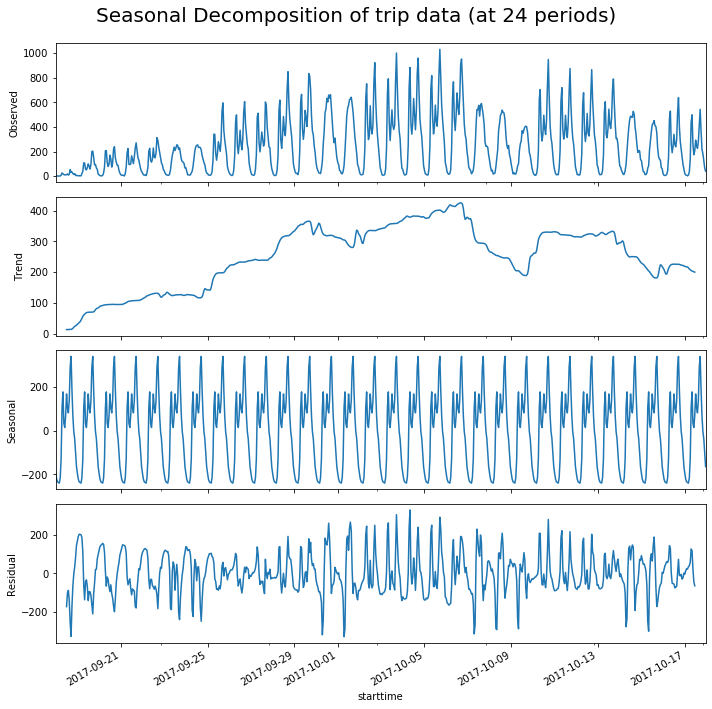
**Figure X** (n=720)

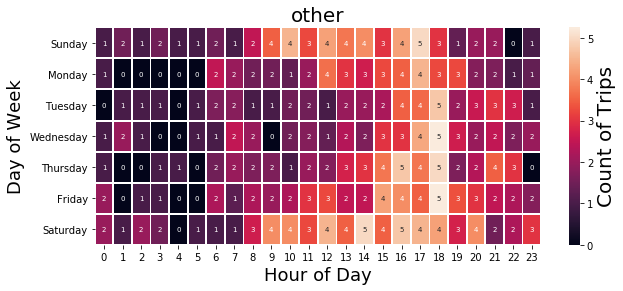


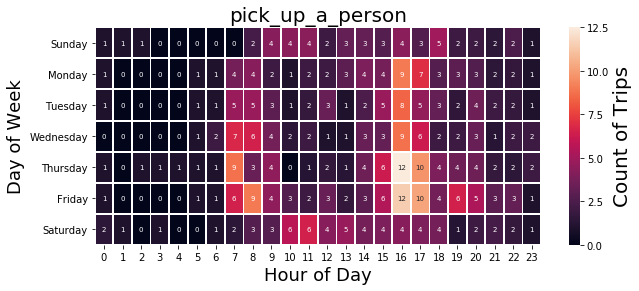


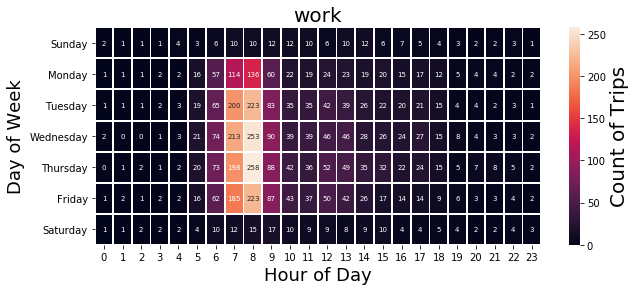
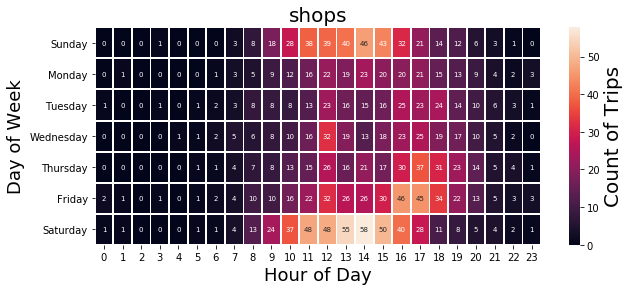
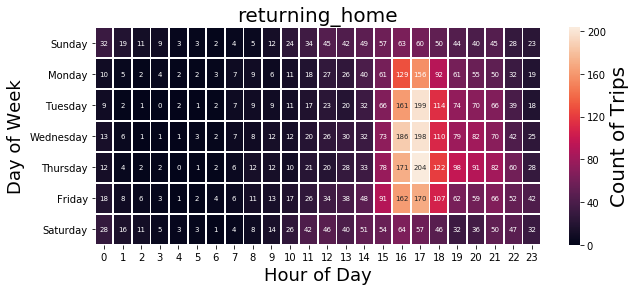
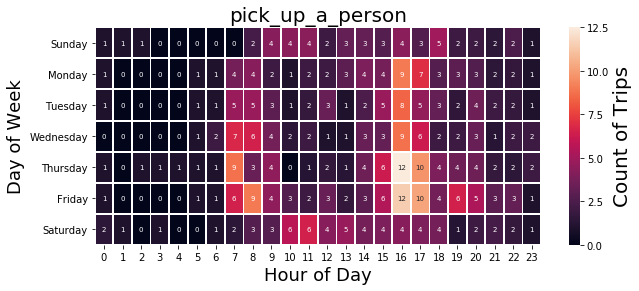
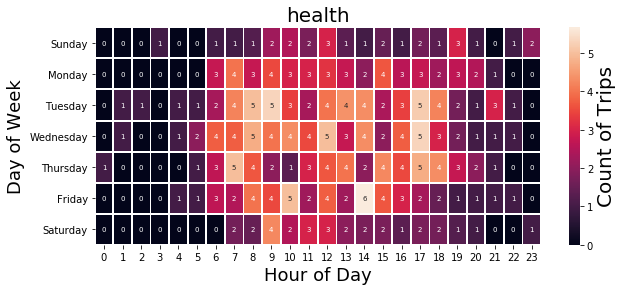
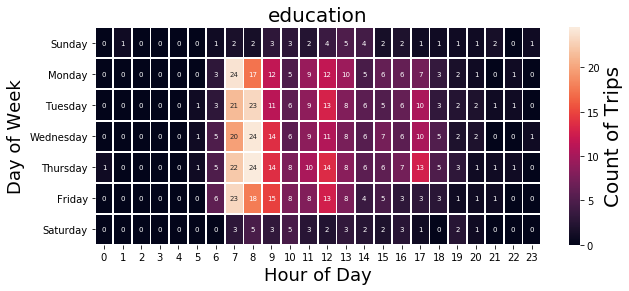
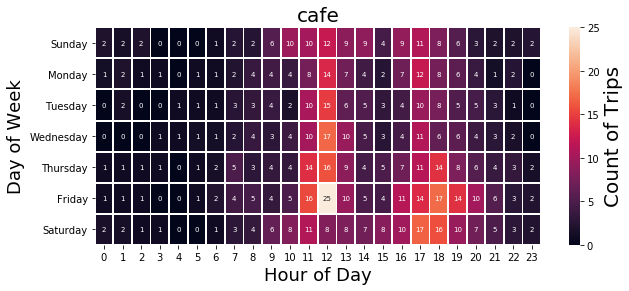


Clear diurnal and weekly a (Seasonal Decomposition)

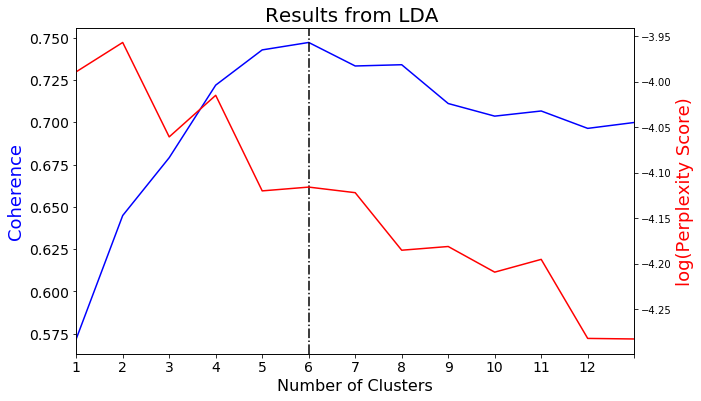


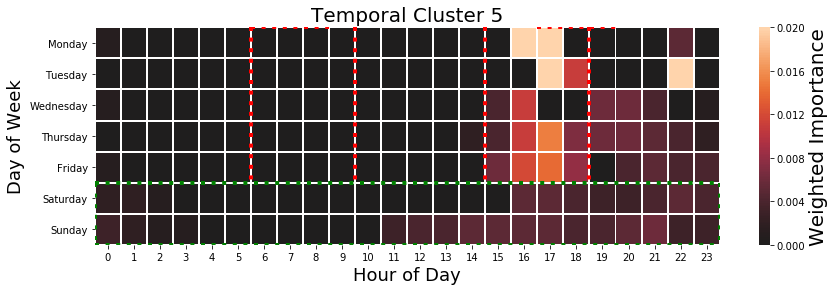
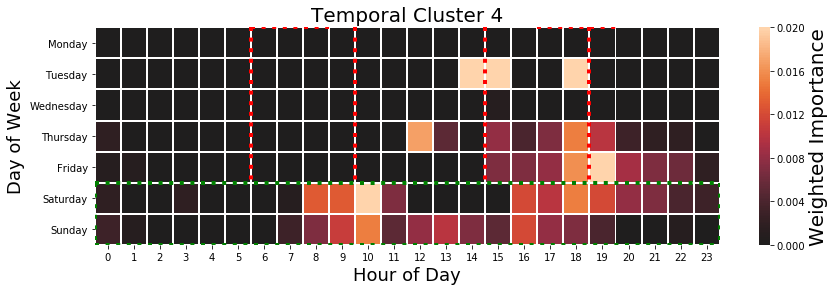
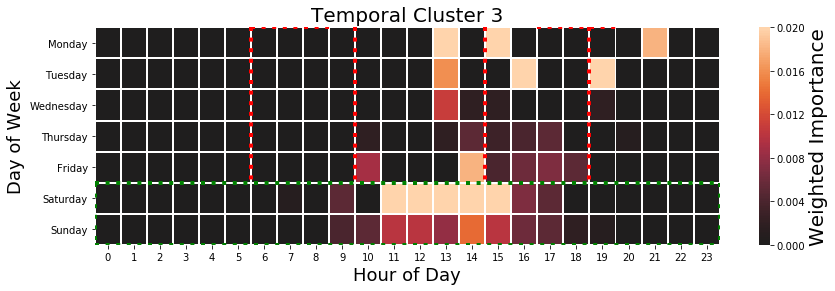
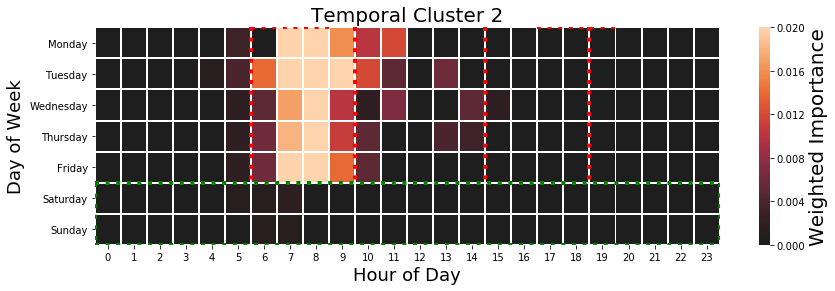
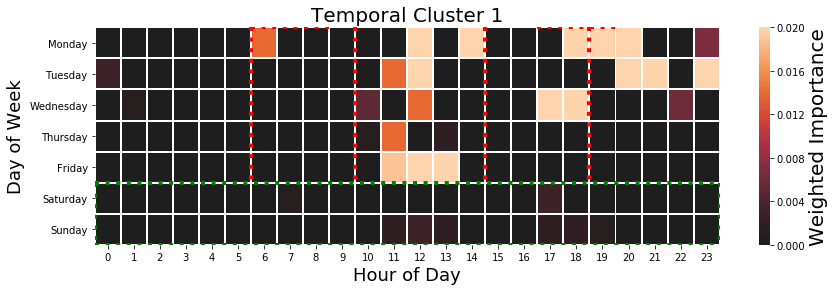




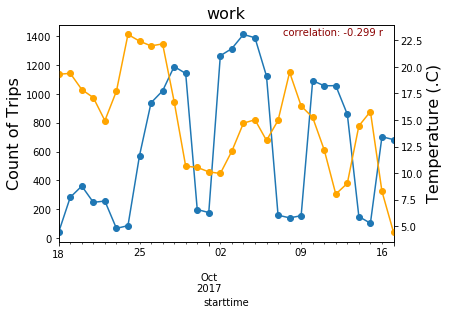
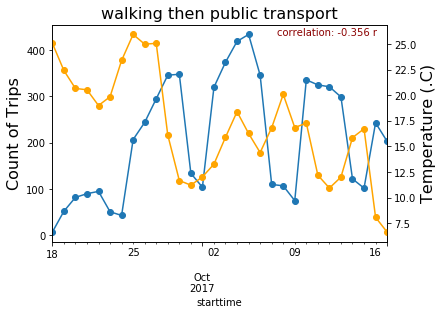
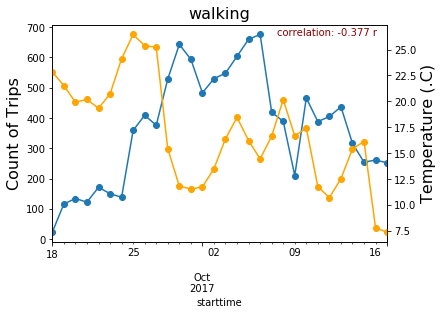
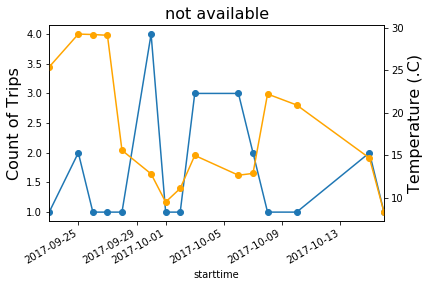
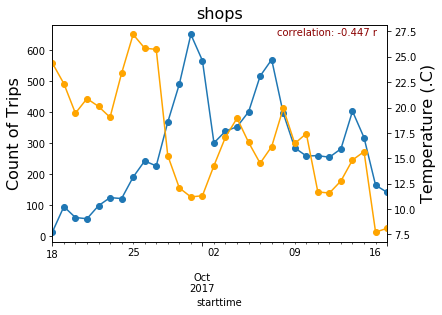
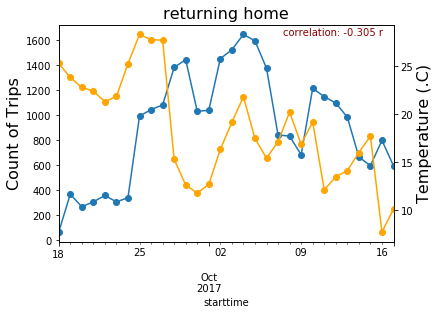
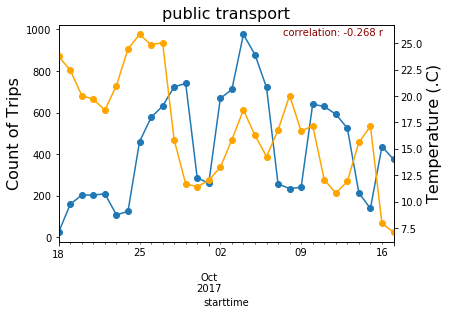
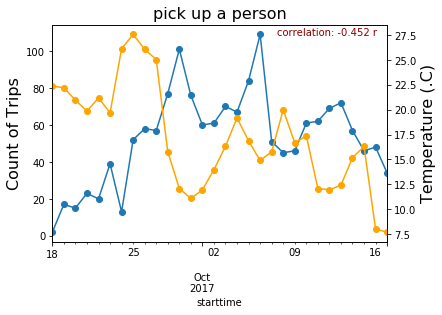
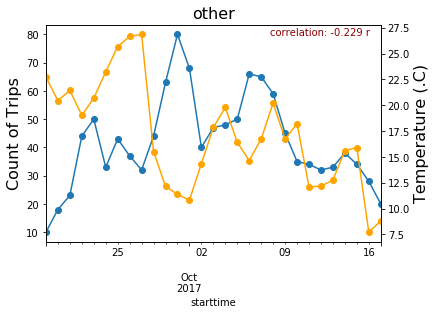
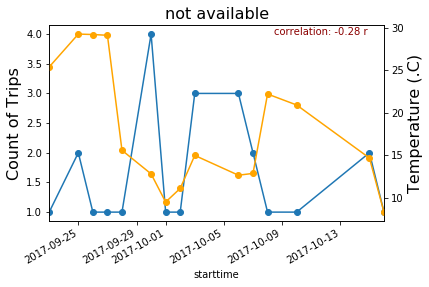
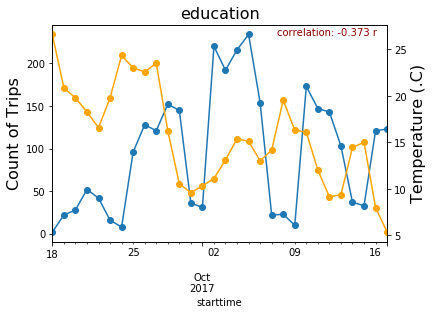
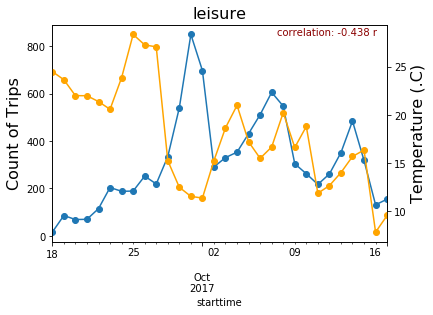
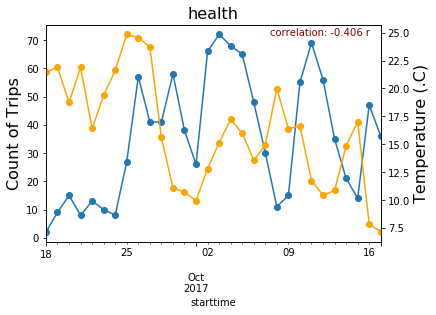
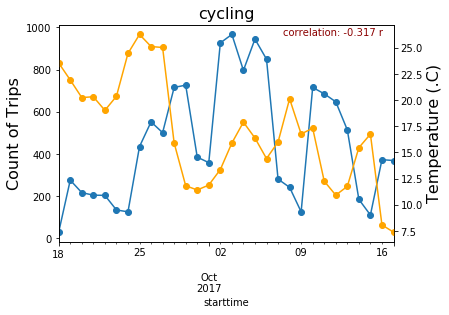
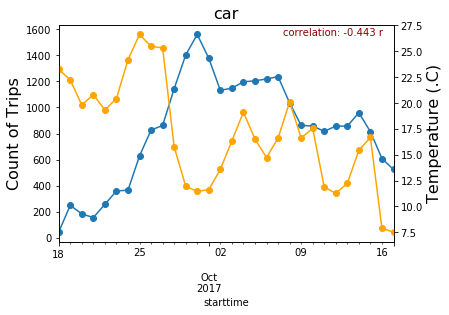
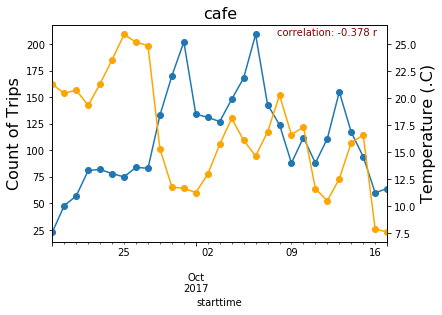


LDA (temporal clustering)





Comparison with weather

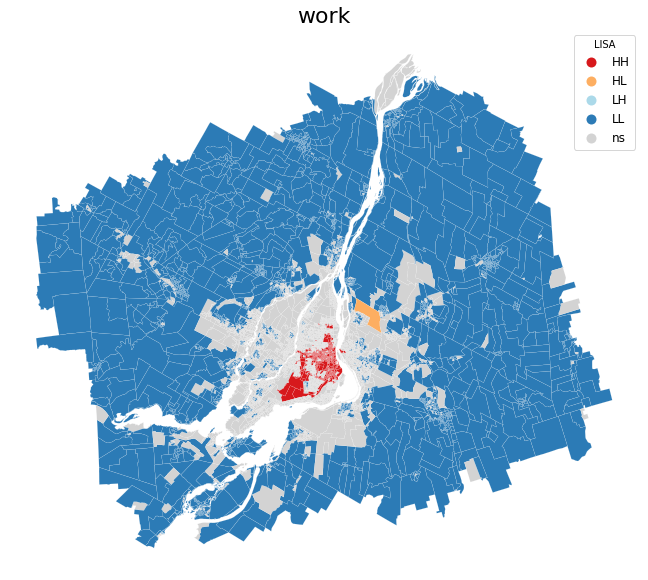
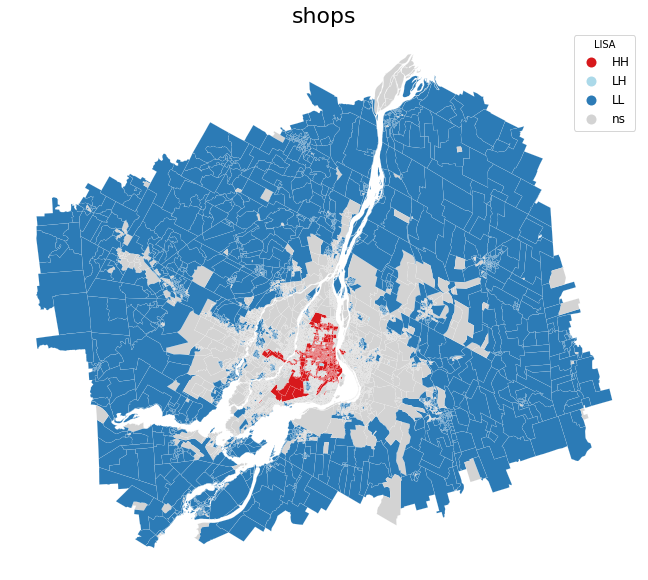
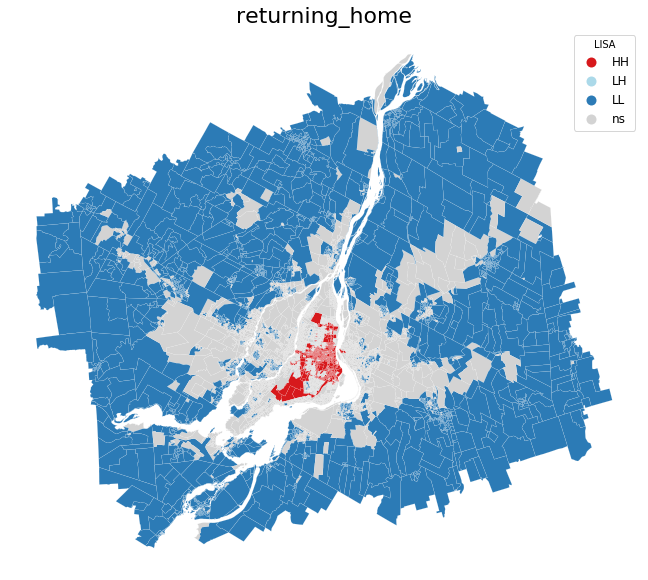
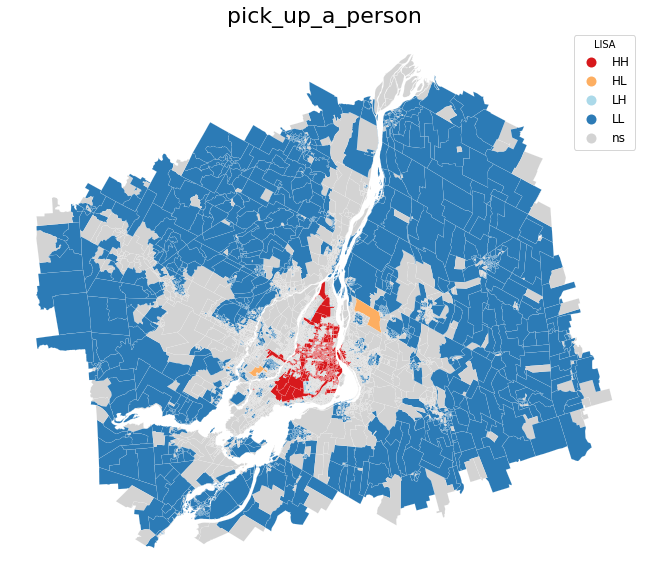
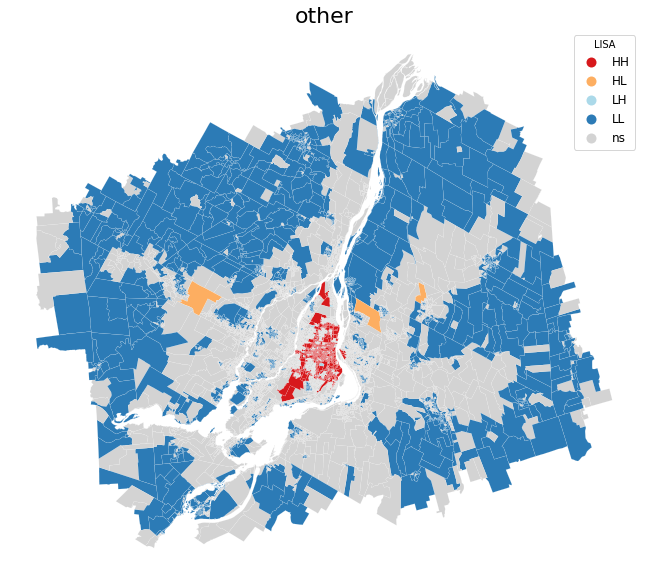
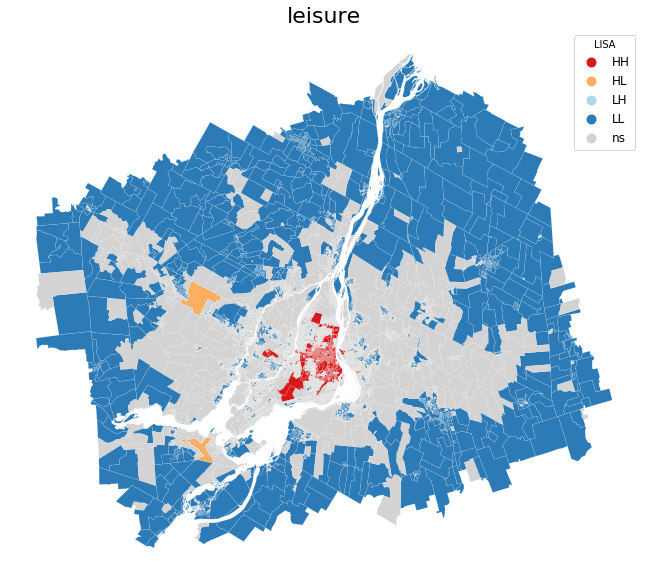
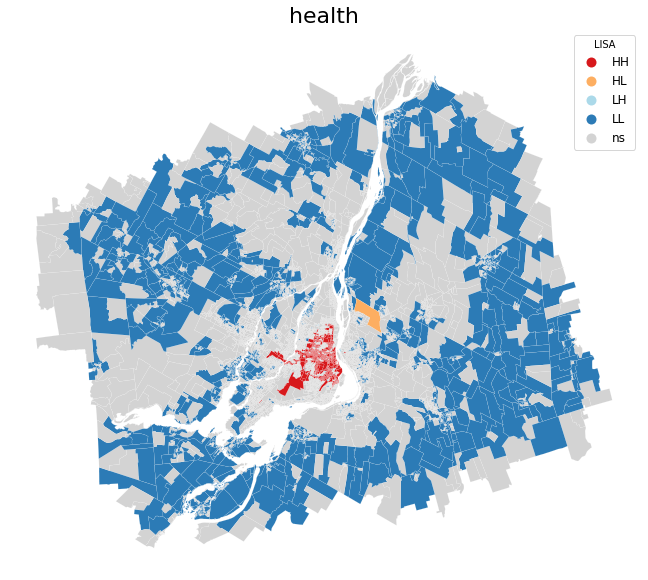
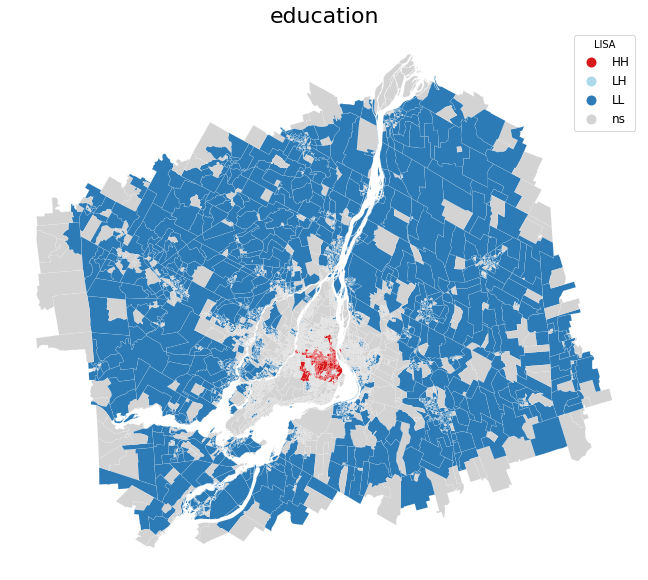
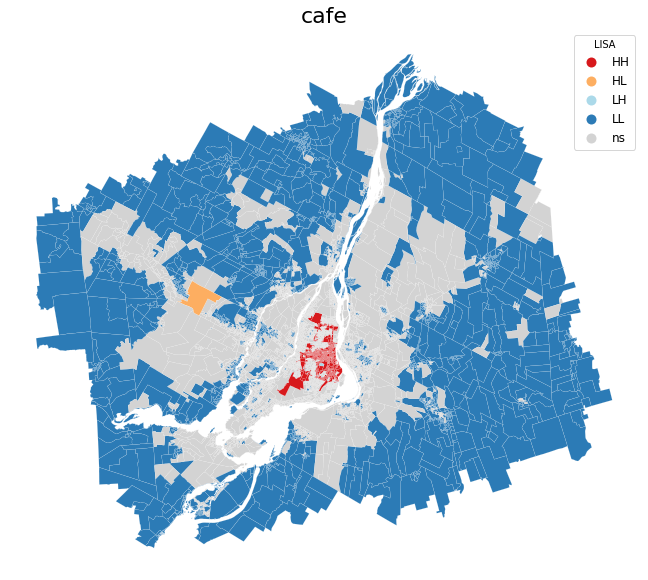


Augmented Dickey-Fuller Test (significant below 0.005 shown in **bold**)

|  |  |  |  |
| --- | --- | --- | --- |
| Purpose | ADF | p-value | n |
| All | -2.7261 | 0.0696 | 185285 |
| Cafe | -2.7386 | 0.0676 | 3189 |
| Education | -2.8689 | 0.0491 | 2830 |
| health | -4.1338 | **0.0009** | 1061 |
| Leisure | -1.8601 | 0.3511 | 9379 |
| Not available | -4.7958 | **0.0001** | 25 |
| Other | -2.4963 | 0.1164 | 1219 |
| Pick a person up | -2.8686 | 0.0491 | 1592 |
| Returning home | -2.8543 | 0.0509 | 27128 |
| Shops | -1.9669 | 0.3013 | 8554 |
| Work | -2.2594 | 0.1854 | 19241 |

Spatial Analysis:

Distribution:



Global Moran’s I

pick\_up\_a\_person 0.5629000152701666 0.0 0.001

work 0.5921787733479512 0.0 0.001

shops 0.5921624780237638 0.0 0.001

returning\_home 0.6169009566510006 0.0 0.001

health 0.547517193189426 0.0 0.001

leisure 0.5448749744913084 0.0 0.001

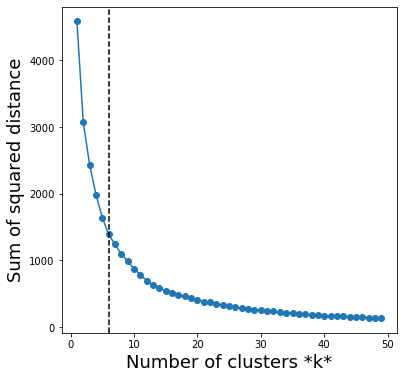
education 0.5872240653154563 0.0 0.001

other 0.5518371624439252 0.0 0.001

cafe 0.5733876521027782 0.0 0.001

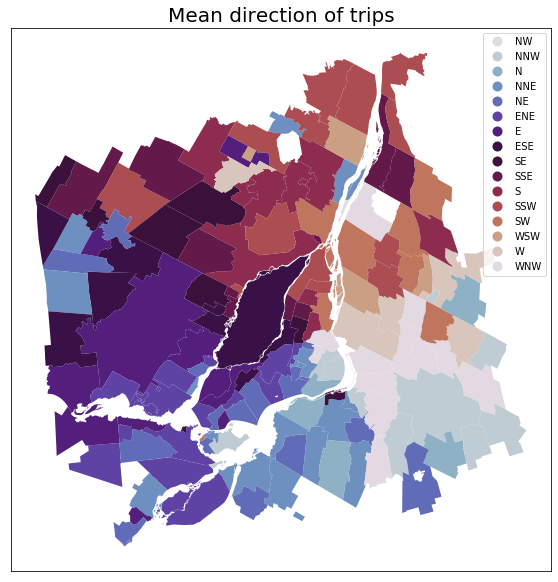
not\_available 0.4356105321751141 0.0 0.001

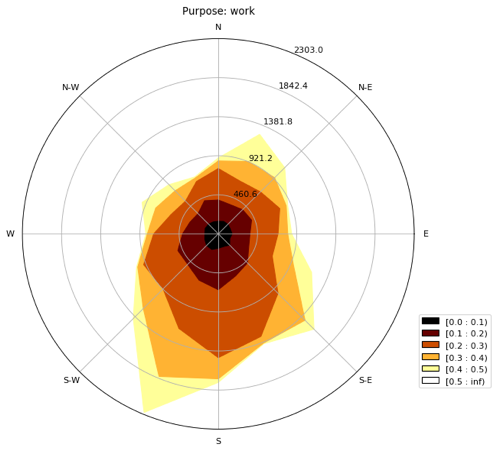
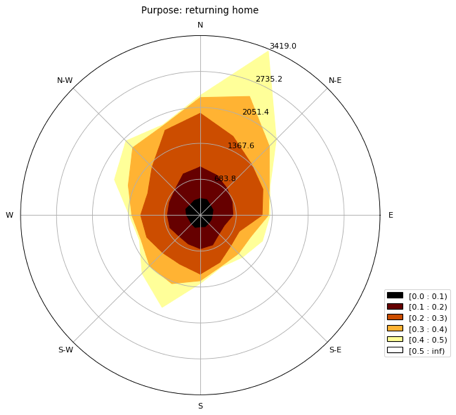
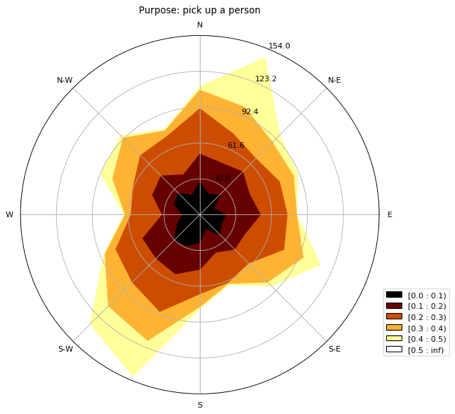
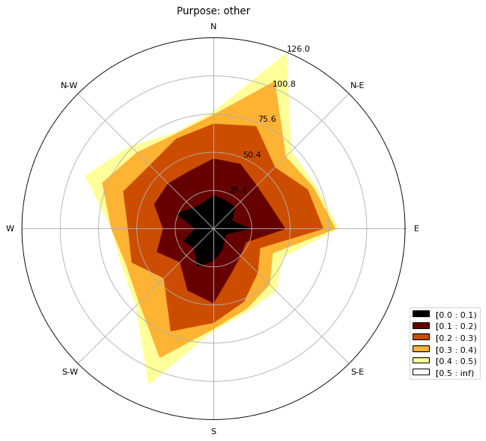
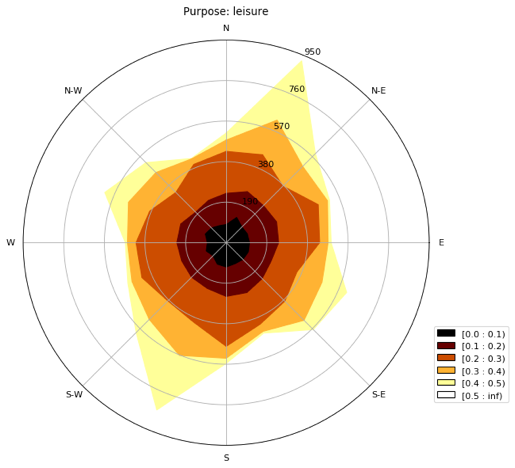
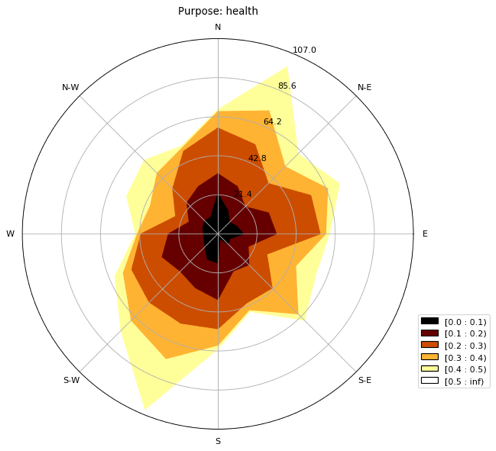
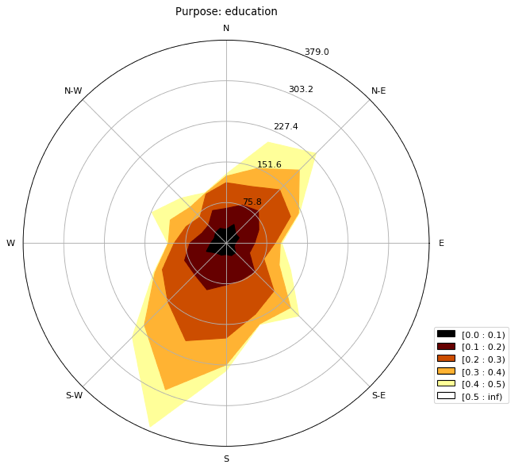
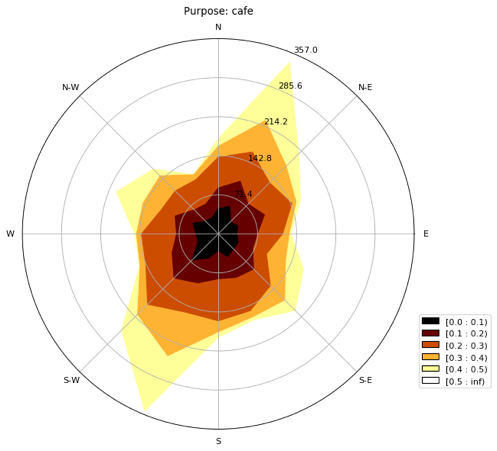
KMeans



Spatial Regression

Spatial-Temporal:





Clustering:

* STSS?

*4.3 Modelling:*

Examining the forecast-ability of the network: To build upon the structure and findings of the ESTDA.

*Classification of purpose of travel:*

Random Forest:

* Feature importance
* Plot residuals

SVC:

For the purposes of the classification models, the data is normalised for the purpose of the SVC and ANN

ANN, LSTM or RNN–CNN: