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)

# 4. Results

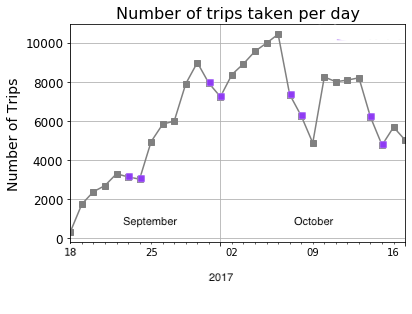
*Introduction:*

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 4.1**, 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 4.1** 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).

### 4.1.1 Distance & Duration:

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 4.1**, 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 4.1** Number of 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

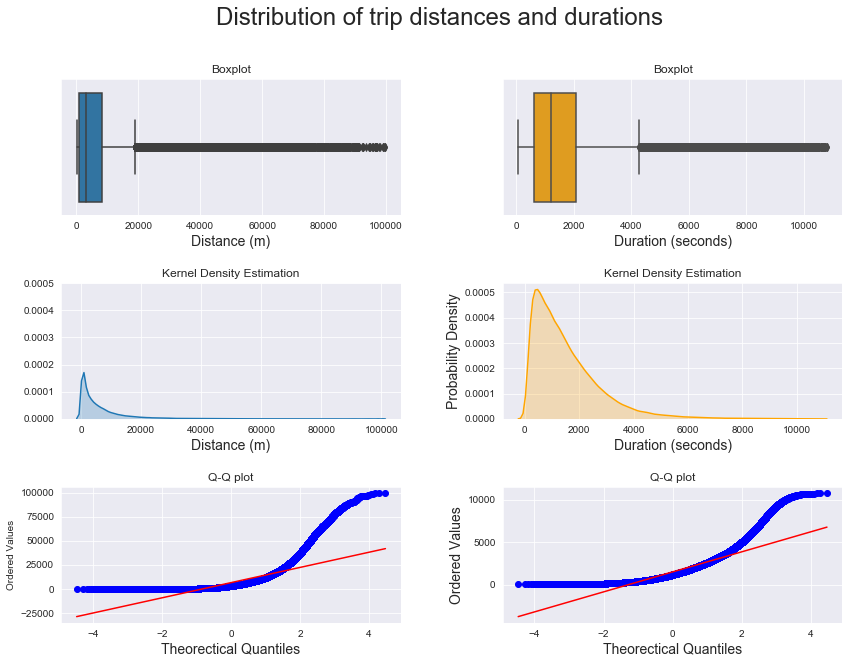
- relatively closer median to mean for duration

- Distance more heavily skewed and higher kurtosis

**Table 4.2** Summary statistics for Distance and Duration of trips (converted to km and minutes; *N=177,938*)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *mean* | *STD* | *min* | *25%* | *Median* | *75%* | *95%* | *max* | *kurtosis* | *Skewness* |
| *Distance (km)* | 6.63 | 9.92 | 0.05 | 0.84 | 3.14 | 8.09 | 25.25 | 99.81 | 15.216 | 3.355 |
| *Duration (min)* | 25.62 | 21.42 | 1.00 | 10.27 | 20.07 | 34.68 | 65.81 | 179.98 | 6.097 | 1.967 |

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 4.2**). Distance is more skewed with a longer tail. See **Figure 4.2**

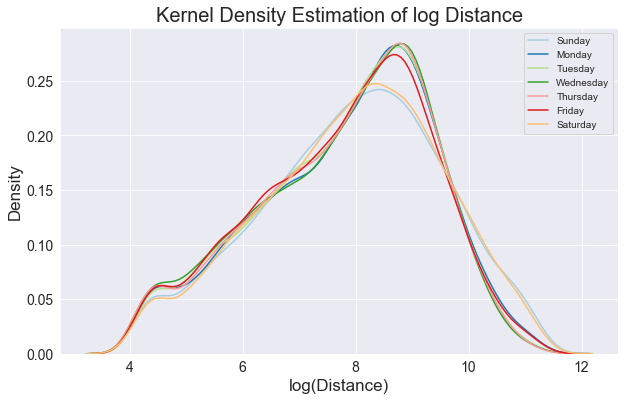


**Figure 4.2** Boxplots (top), Kernel Density Estimation (middle) and Quantile-Quantile (bottom) plots showing the distribution of trip distance and duration.

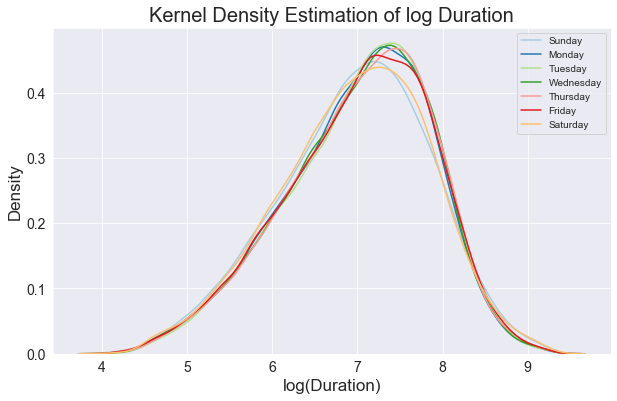
The data is hence, normalised for the purpose of the classification models using by taking the natural log. **Figure 4.3+4.4**

Log distance skew: -0.37300090670060865 kurtosis: -0.5147541361828902

Log duration skew: -0.38136680556957275 kurtosis: -0.21730048341457042

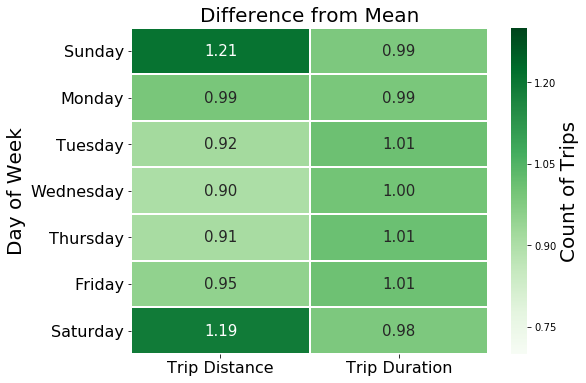


**Figure 4.3** 1-Dimensional Kernel Density Estimation plot of trip distances.



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

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 4.5**.



**Figure 4.5** Average trip distance and duration as proportion of the mean.

See: <https://www.graduatetutor.com/statistics-tutor/interpreting-regression-output/> for interpreting the regression

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).

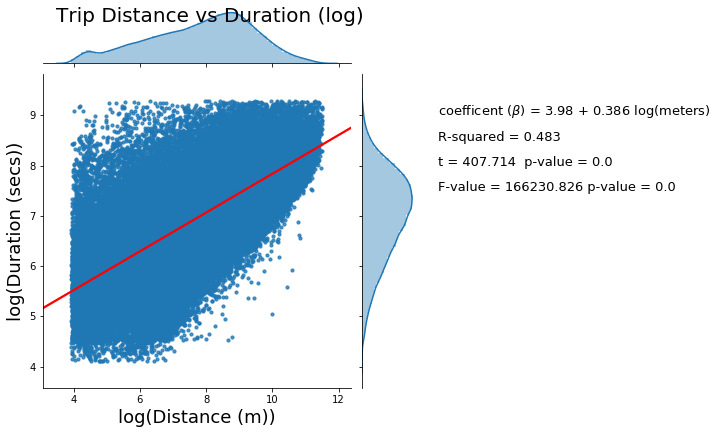
- relatively strong relationship, how well the line fits the data (R2=0.483)

- t-value is significant indicating difference in means,

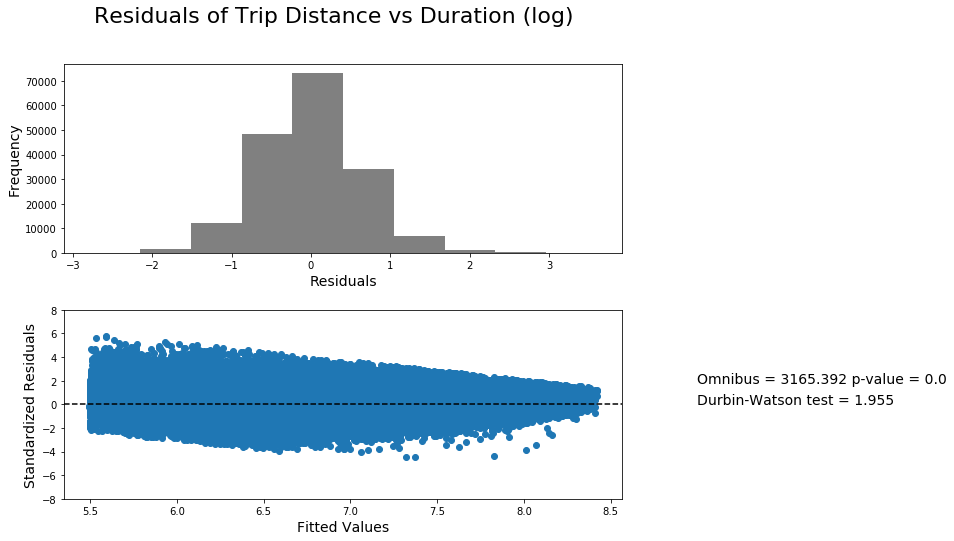
- F-value is probability the model is wrong (indicates the probability that all the coefficients in our regression output are actually zero)

- 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”

### 4.1.2 Mode & Purpose:

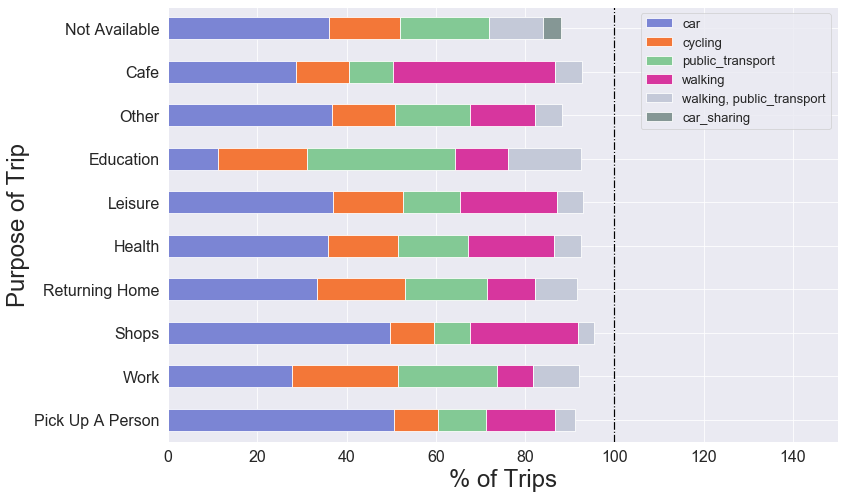
Of the 11 unique purposes (see Table X). The amount of each, varies across the 5 weeks of study

(Note, each week has 7 days apart from the last).

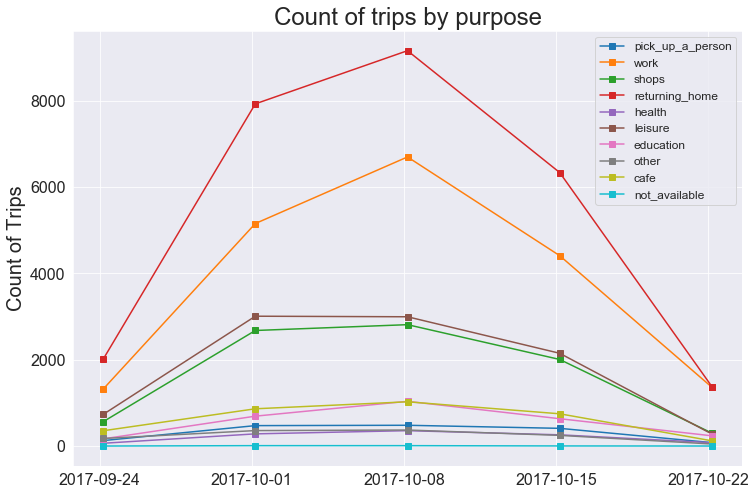
**Table X**

|  |  |  |
| --- | --- | --- |
| *Response to purpose of trip* | *n* | *Popular Transport Modes* |
| *Café* | 3,115 | 1. Walking 2. Car 3. Cycling |
| *Education* | 2,769 | 1. Public transport 2. Cycling 3. Walking, Public transport |
| *Health* | 1,044 | 1. Car 2. Walking 3. Cycling |
| *Leisure* | 9,167 | 1. Car 2. Walking 3. Cycling |
| *Not Available* | 25 | 1. Car 2. Public transport 3. Cycling |
| *Other* | 1,203 | 1. Car 2. Public transport 3. Walking |
| *Pick up a person* | 1,574 | 1. Car 2. Walking 3. Public transport |
| *Returning Home* | 26,819 | 1. Car 2. Cycling 3. Public transport |
| *Shops* | 8,363 | 1. Car 2. Walking 3. Cycling |
| *Work* | 18,950 | 1. Car 2. Cycling 3. Public transport |

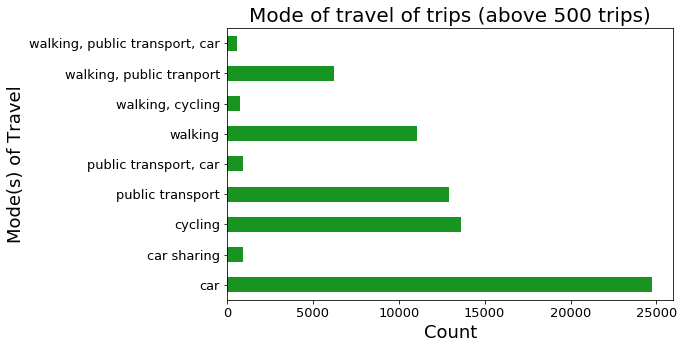
[Figure showing how people travel for given activities -> a stacked bar]



**Figure 4.X**

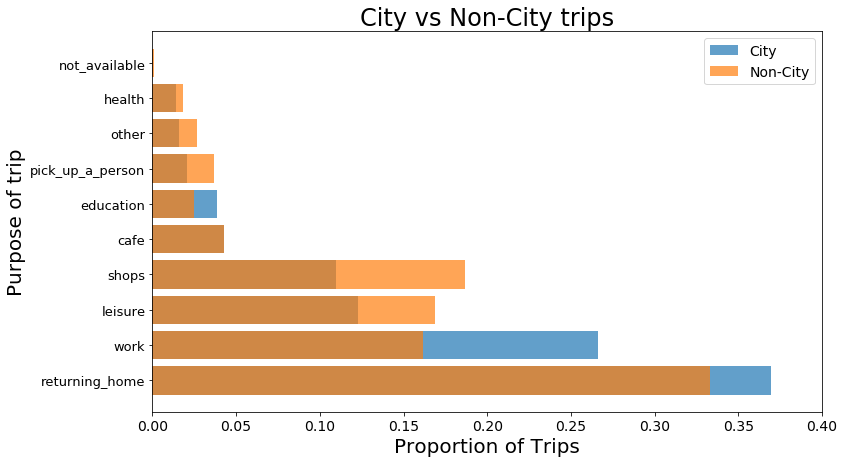


**Figure 4.X**

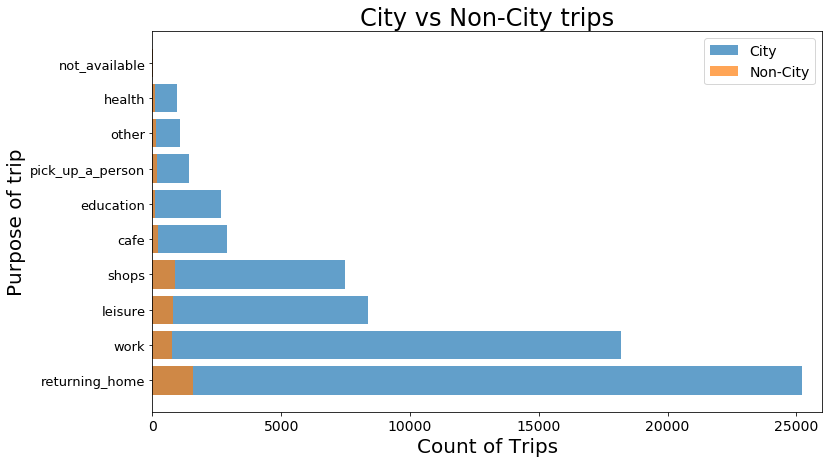


**Figure 4.X**

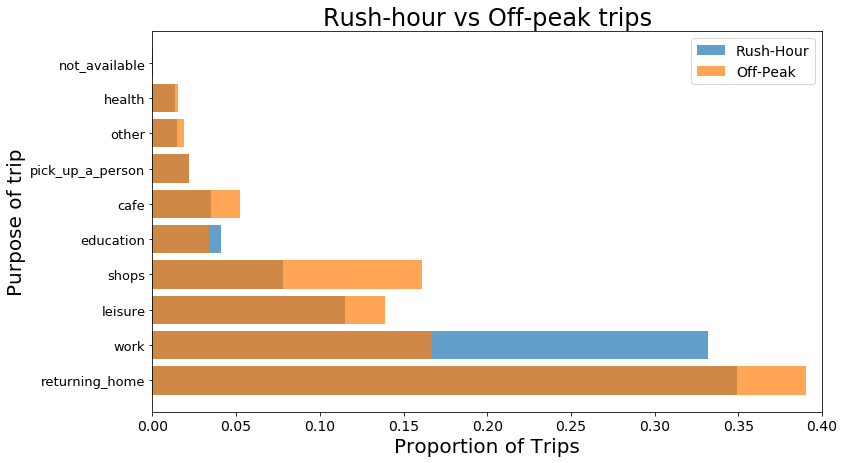
*City and Rush hour:*

**

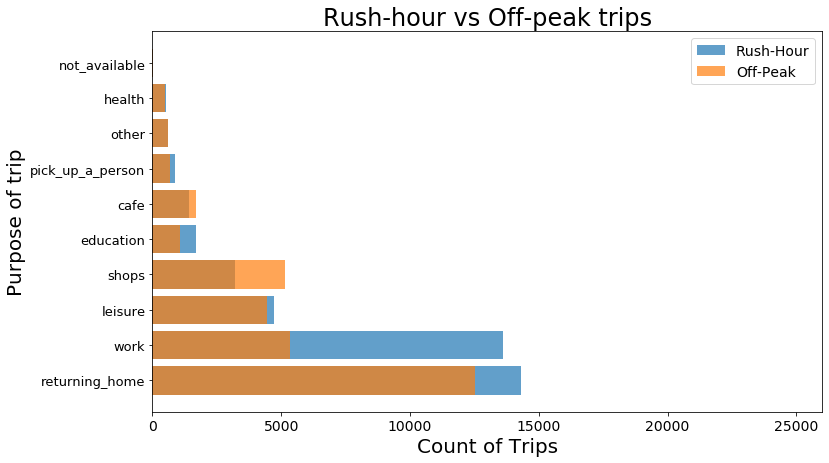
**Figure 4.X**

**

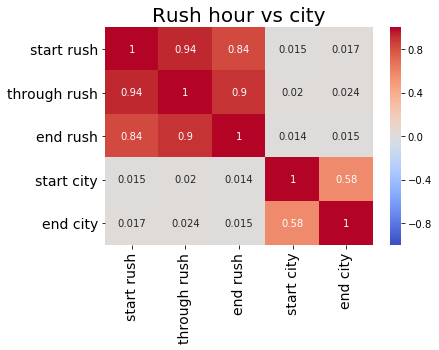
**Figure 4.X**

**

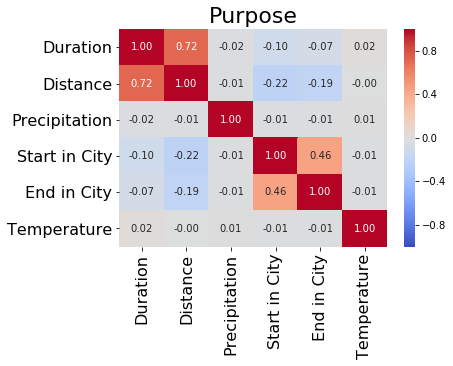
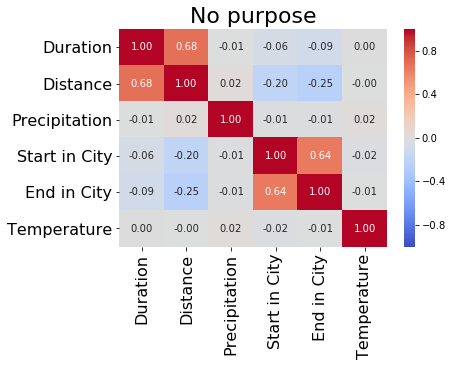
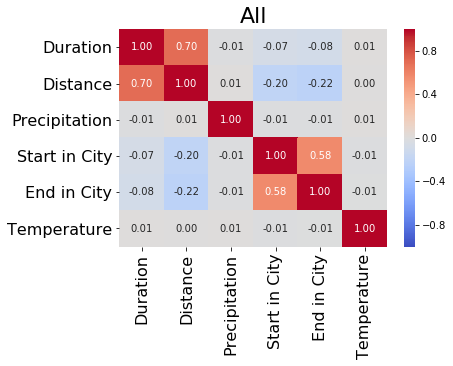
**Figure 4.X**

**

**Figure 4.X**



**Figure 4.X**



**Figure 4.X**

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

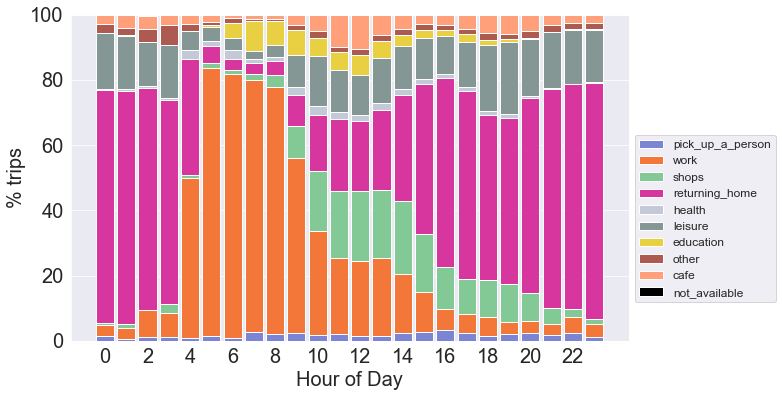
For this report, we divide this section into three main parts: spatial, temporal and spatial-temporal methods used to discern signal from the data.

Temporal Analysis:

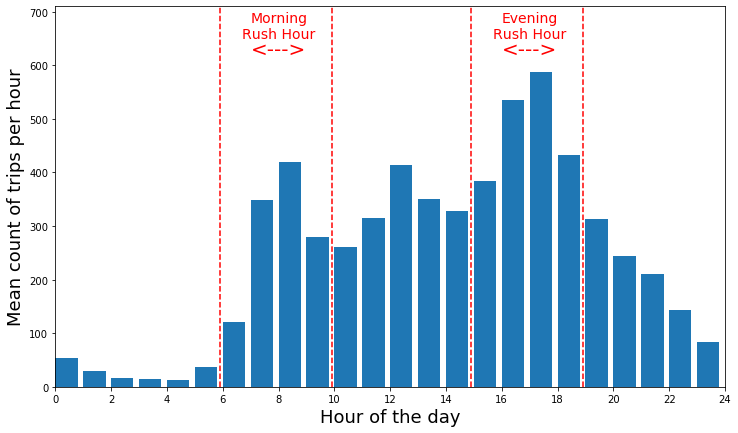
Daily,

Weekly,

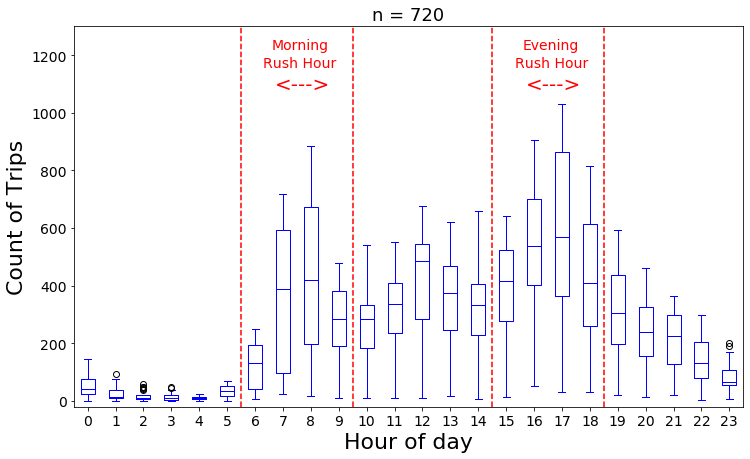
Seasonly



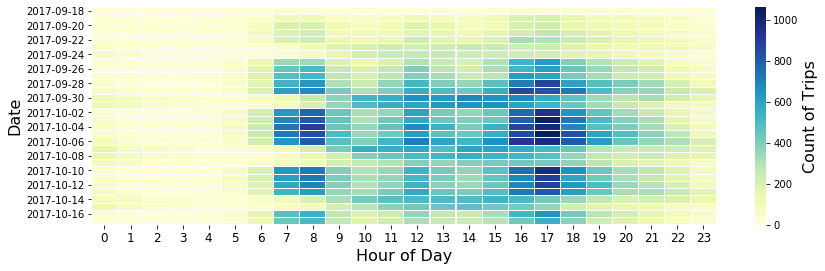
**Figure 4.X**



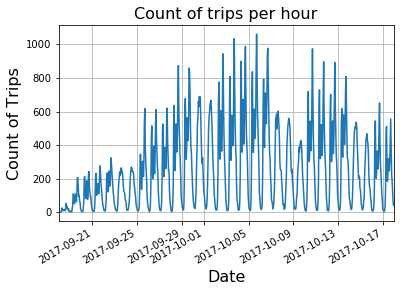
**Figure 4.X** (*n=720*)



**Figure 4.X**

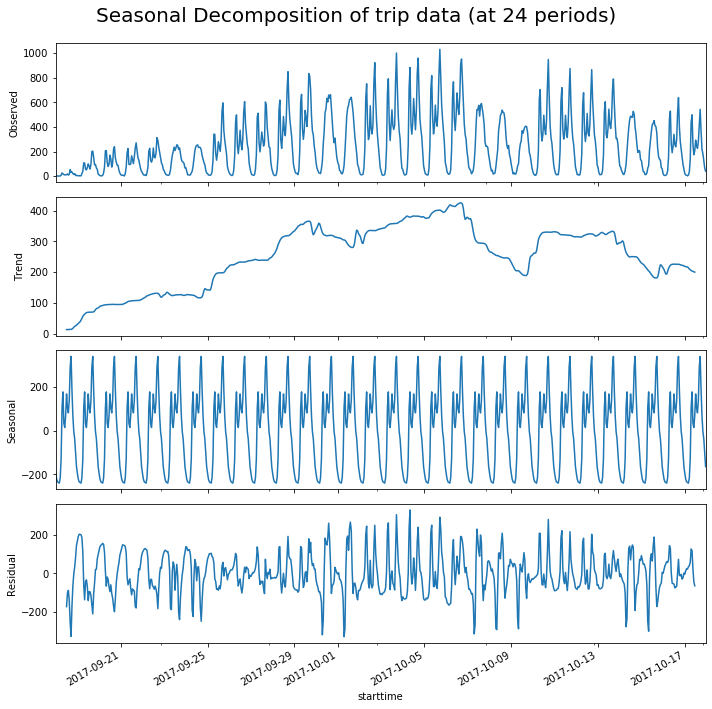


**Figure 4.X**

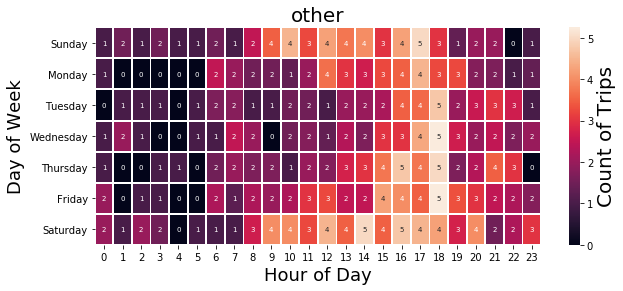


**Figure 4.X**

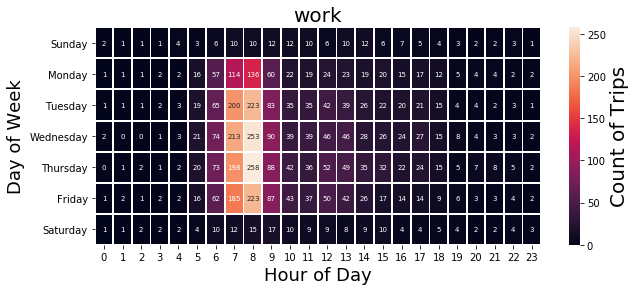
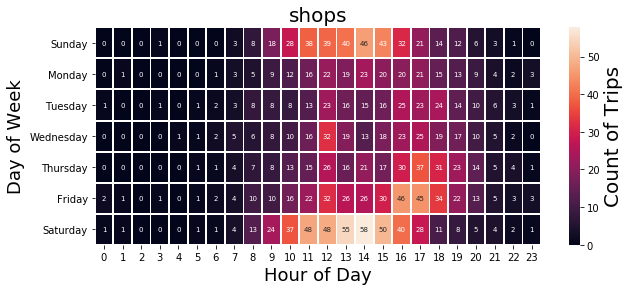
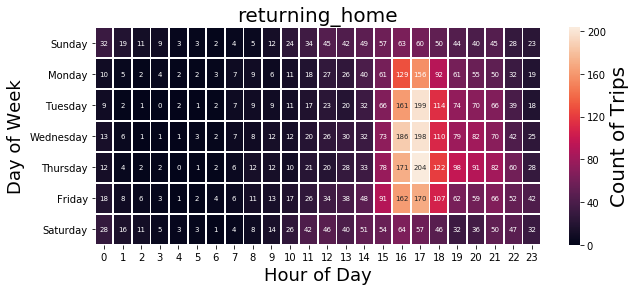
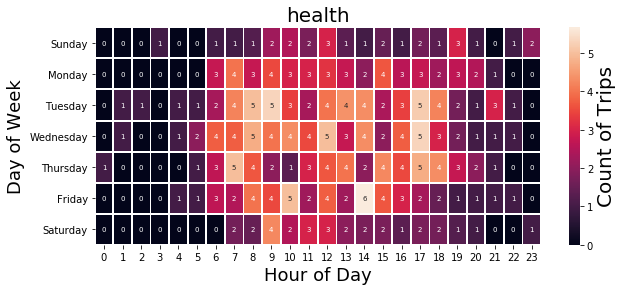
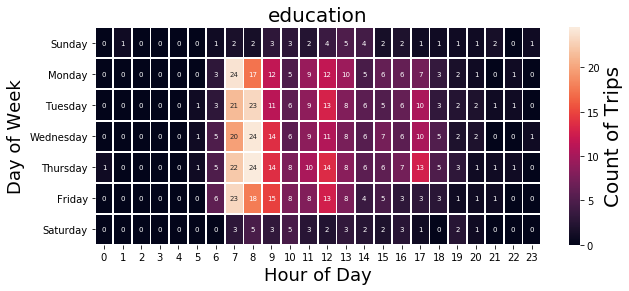
Clear diurnal and weekly a (Seasonal Decomposition)



**Figure 4.X**







Purpose Correlation:

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)

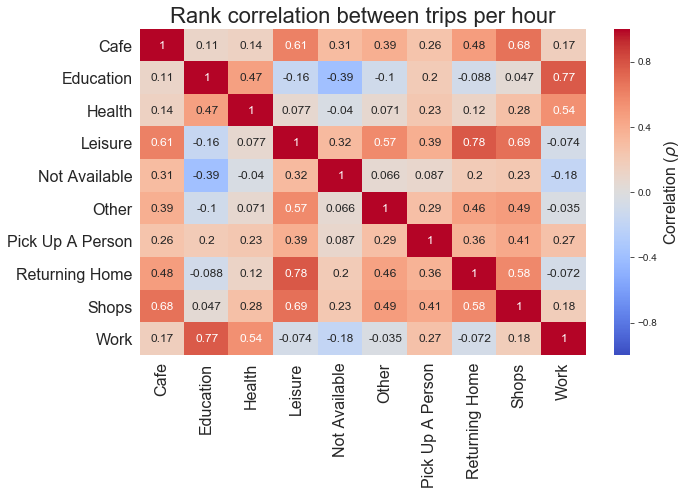
other KstestResult(statistic=0.5, pvalue=9.82488208886262e-164)

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

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

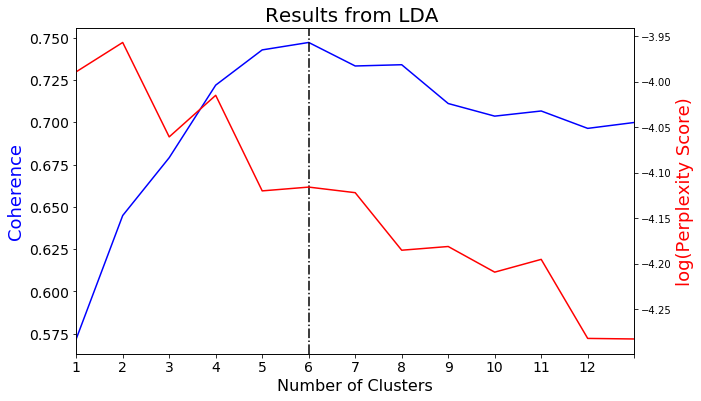
shops KstestResult(statistic=0.6061450521877981, pvalue=3.586226223248267e-249)

work KstestResult(statistic=0.749204542272784, pvalue=0.0)

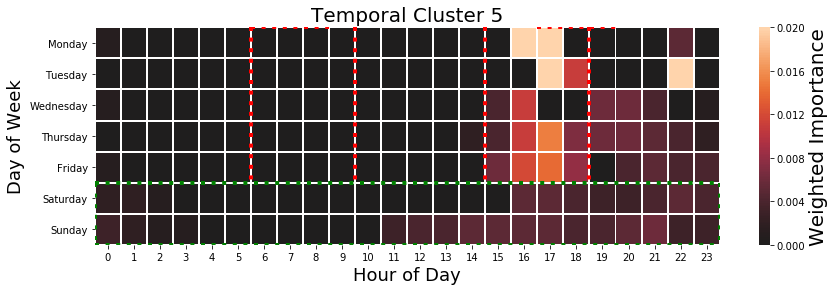
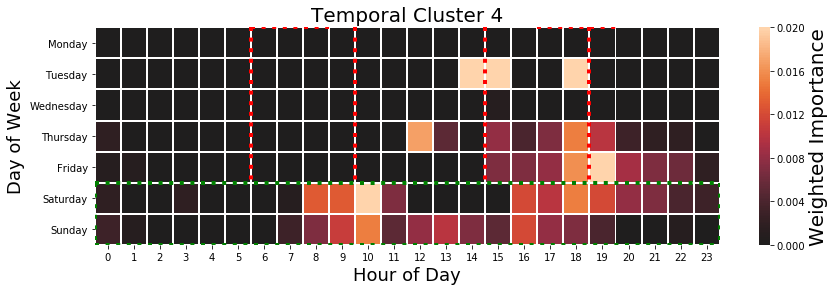
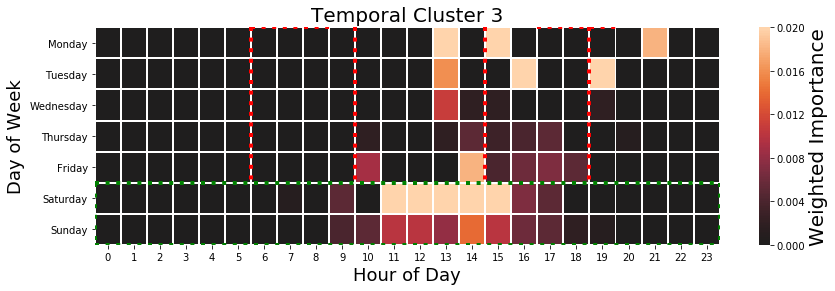
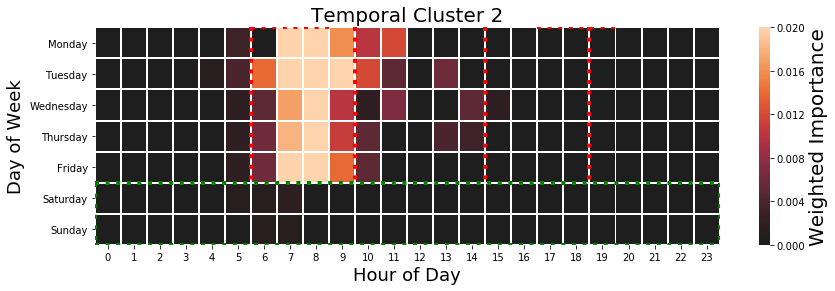
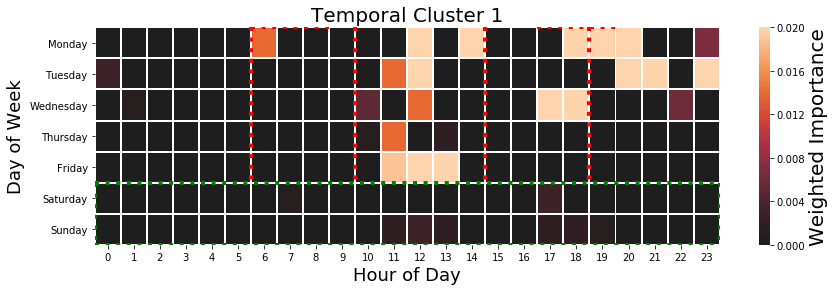


**Figure 4.X**

LDA (temporal clustering)



**Figure 4.X**



**Figure 4.X**

0: [['cafe', 0.153], ['educ', 0.122]],

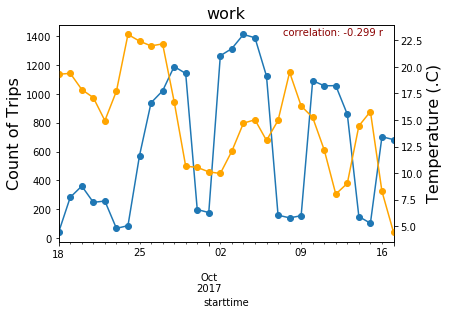
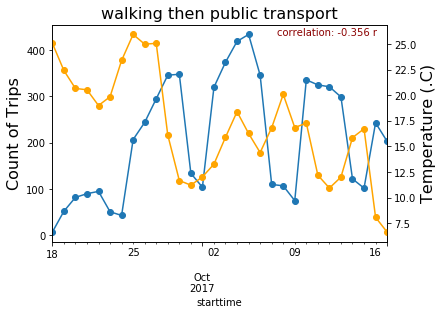
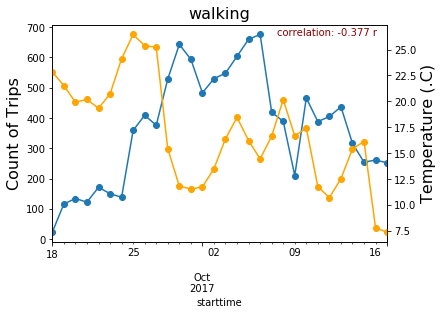
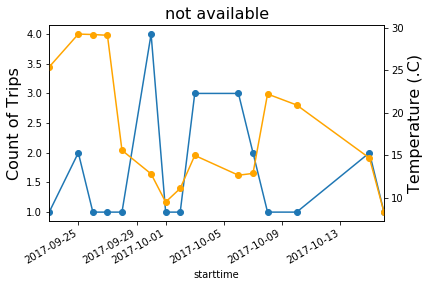
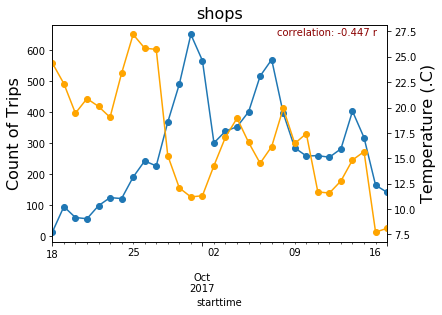
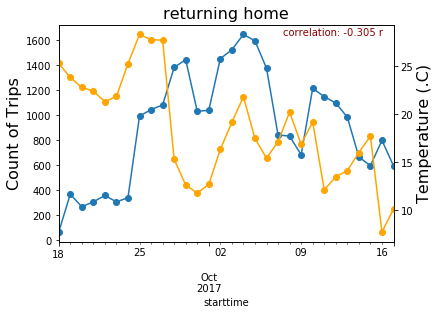
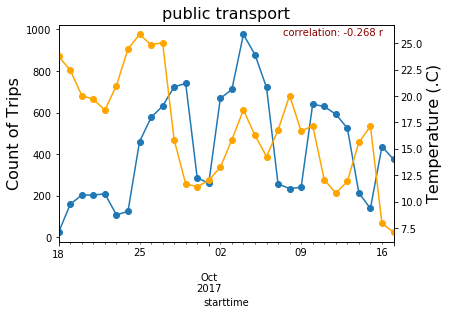
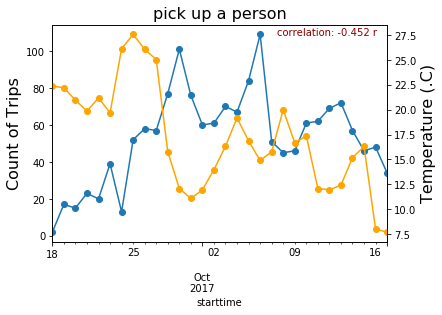
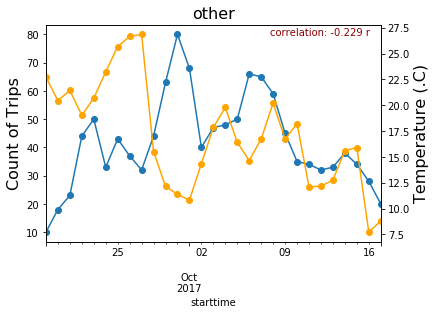
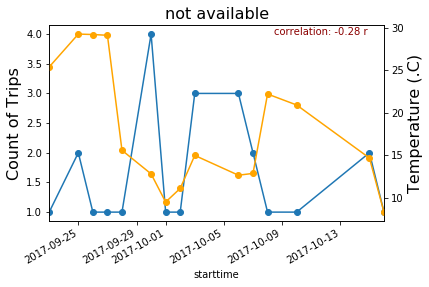
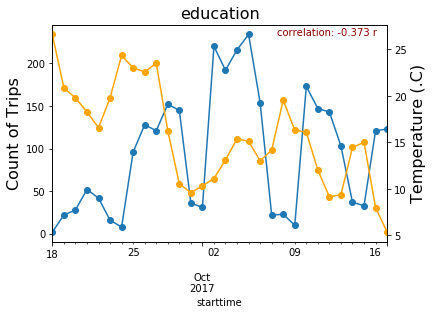
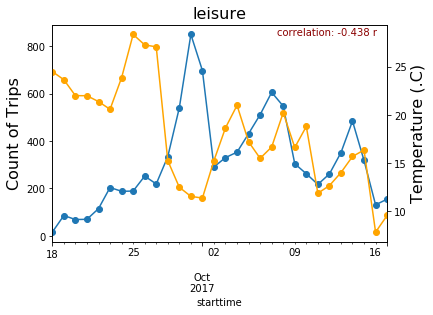
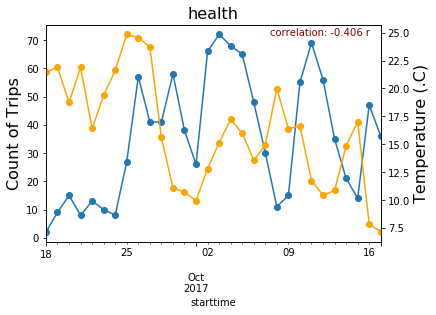
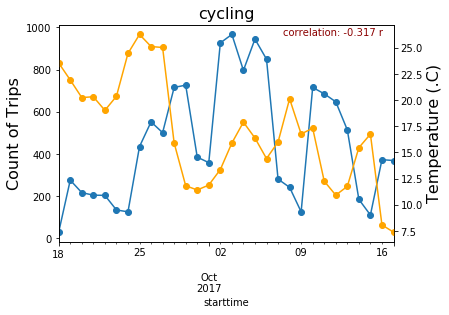
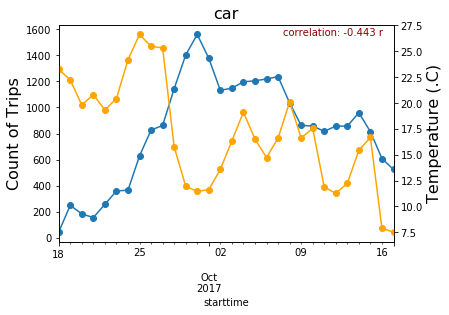
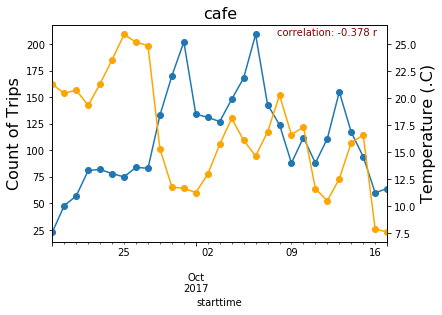
1: [['work', 0.513]],

2: [['shop', 0.35], ['health', 0.048], ['not\_avail', 0.001]],

3: [['leisur', 0.446], ['pick\_up\_a\_person', 0.073]],

4: [['returning\_hom', 0.6], ['other', 0.022]]}

Comparison with weather



**Table 4.X** 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

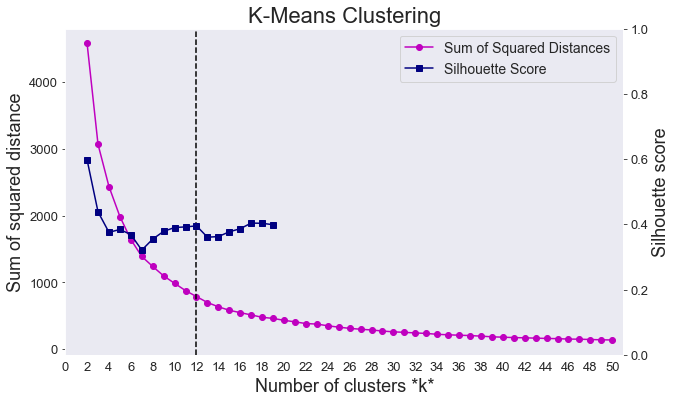
education 0.5872240653154563 0.0 0.001

other 0.5518371624439252 0.0 0.001

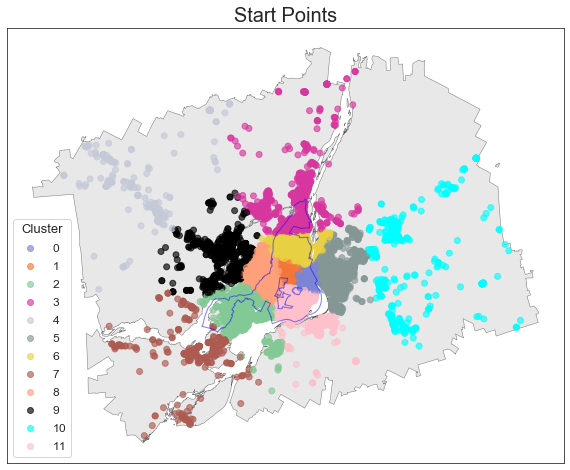
cafe 0.5733876521027782 0.0 0.001

not\_available 0.4356105321751141 0.0 0.001

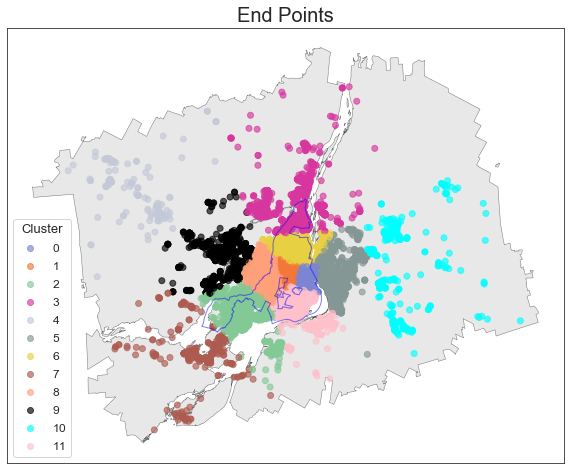
KMeans



**Figure 4.X**



**Figure 4.X**



**Figure 4.X**

Cluster Counts:

5 19069

0 18564

1 10295

10 9725

2 7082

9 1963

4 1837

3 1606

8 843

6 385

11 243

7 189

Name: startclust, dtype: int64

5 19750

0 18380

1 9861

10 9712

2 7076

9 2132

4 1783

3 1654

8 808

6 347

11 176

7 122

Count of purpose per mtl trajet KS tests:

cafe KstestResult(statistic=0.5, pvalue=0.0)

education KstestResult(statistic=0.5, pvalue=0.0)

health KstestResult(statistic=0.5, pvalue=0.0)

leisure KstestResult(statistic=0.5, pvalue=0.0)

not\_available KstestResult(statistic=0.5, pvalue=0.0)

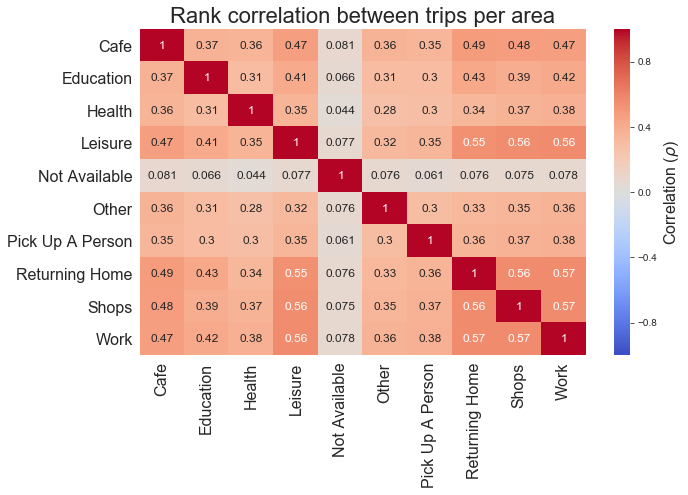
other KstestResult(statistic=0.5, pvalue=0.0)

pick\_up\_a\_person KstestResult(statistic=0.5, pvalue=0.0)

returning\_home KstestResult(statistic=0.5567048189947053, pvalue=0.0)

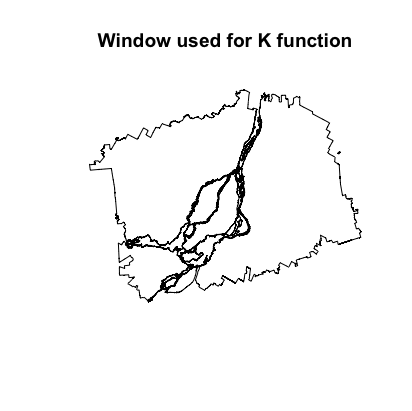
shops KstestResult(statistic=0.5, pvalue=0.0)

work KstestResult(statistic=0.5, pvalue=0.0)

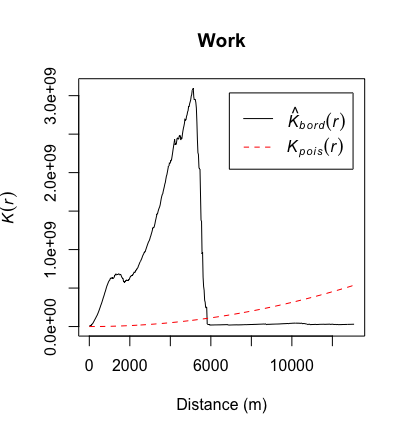
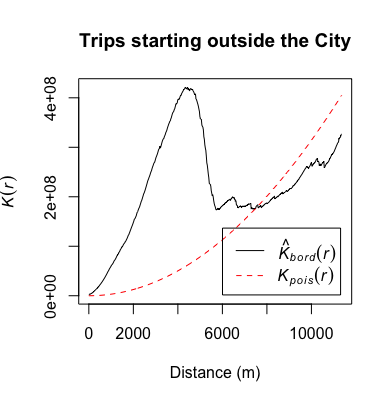
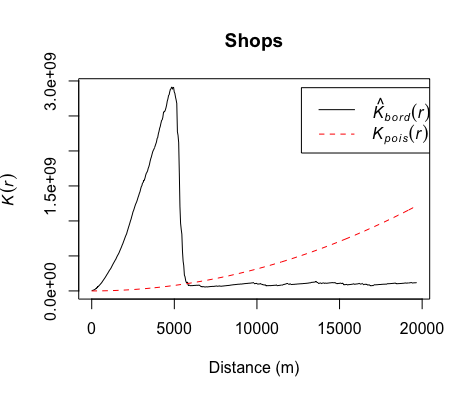
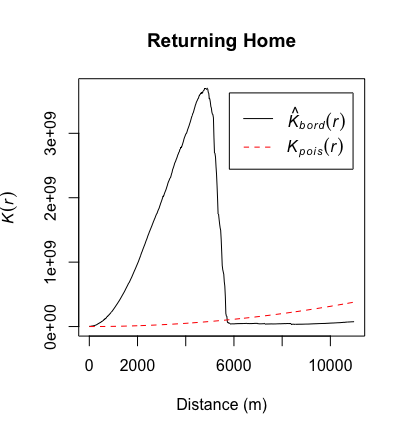
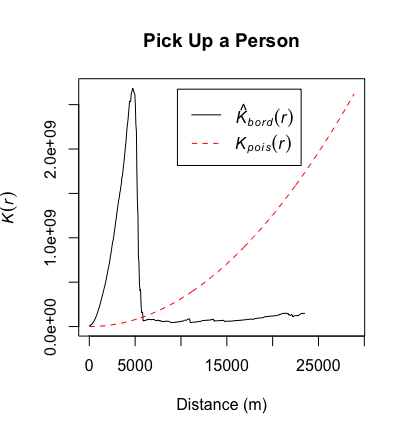
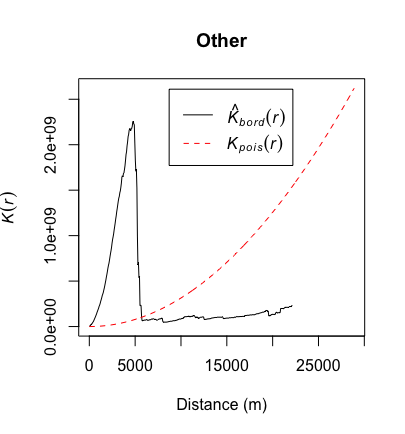
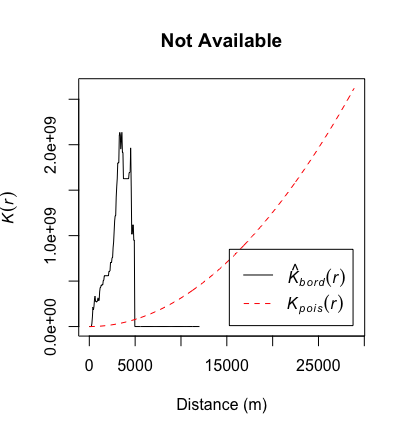
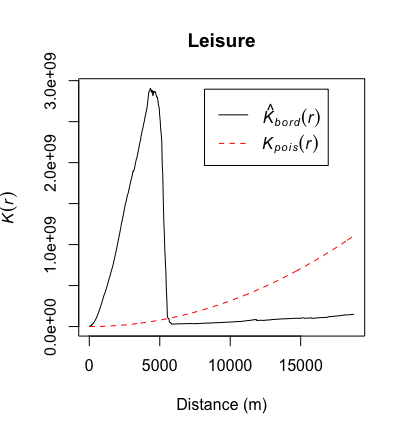
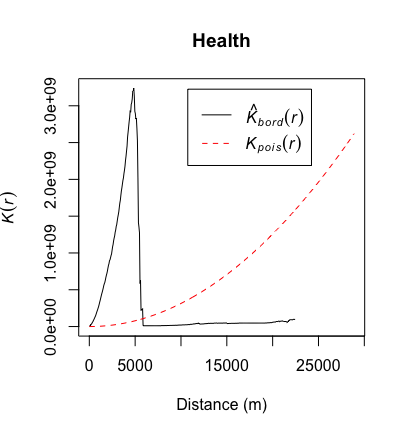
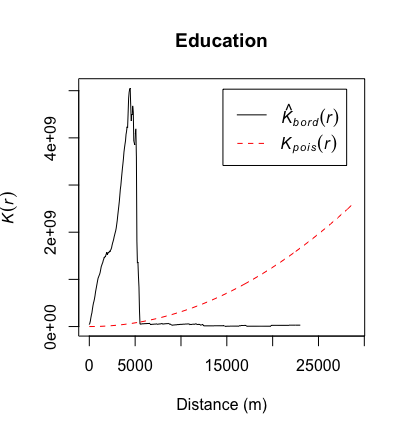
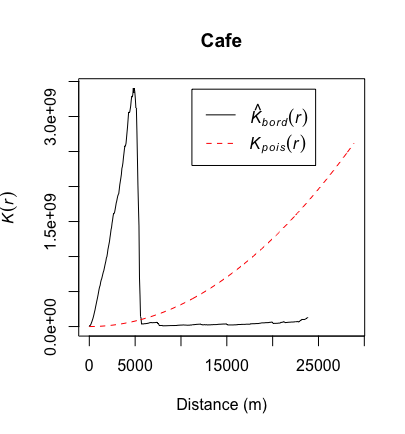


**Figure 4.X**

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



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

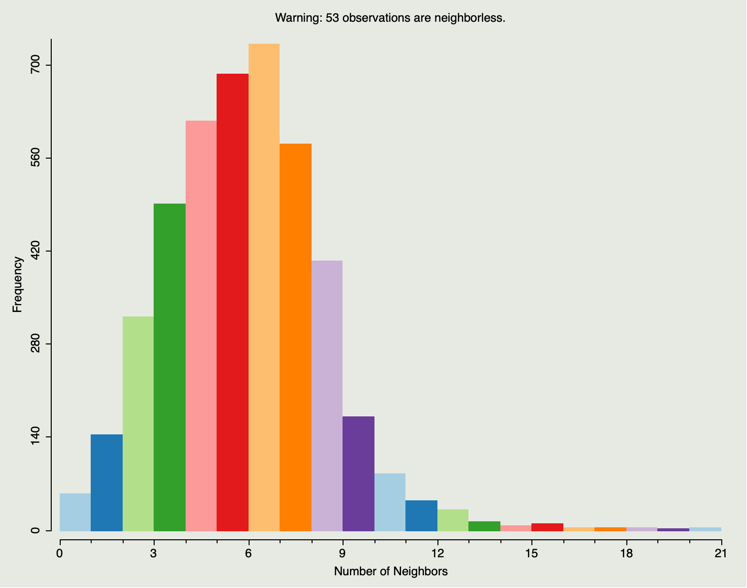


“Using this window, a points pattern object is created from the Treasure Hunt location coordinates and estimations for Ripley’s K is calculated. This estimation of K is plotted in Figure 5 against a Poisson distribution (which indicates Complete Spatial Randomness of Treasure Hunt locations). As shown, 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.”

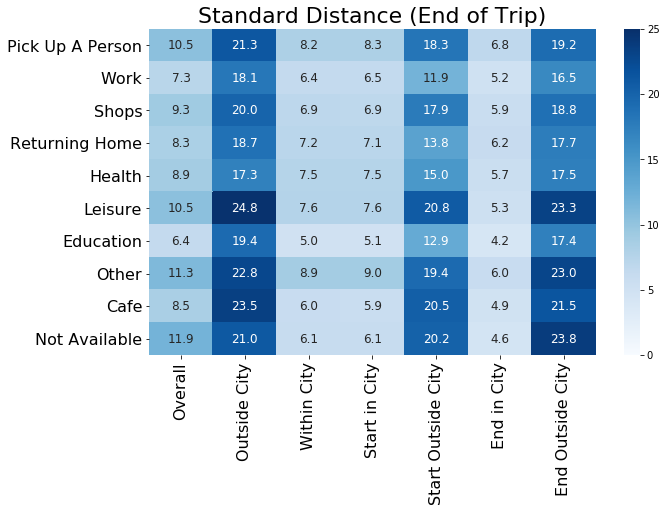
Start in the city cluster end

Start outside the city cluster end

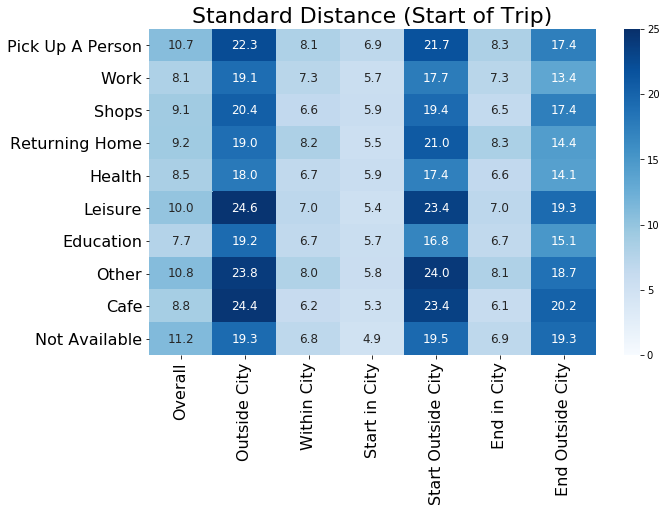
Spatial Regression



**Figure 4.X**



**Figure 4.X**



**Figure 4.X**

Spatial-Temporal:

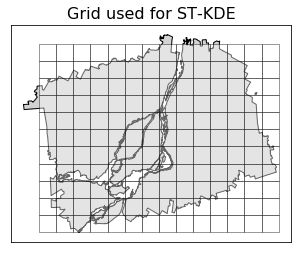
ST-K

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

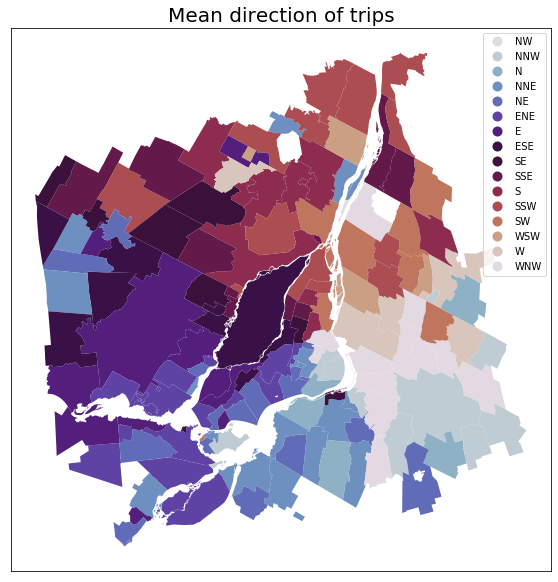


**Figure 4.X**

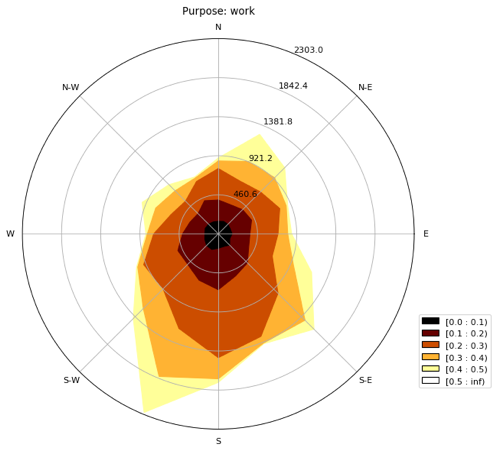
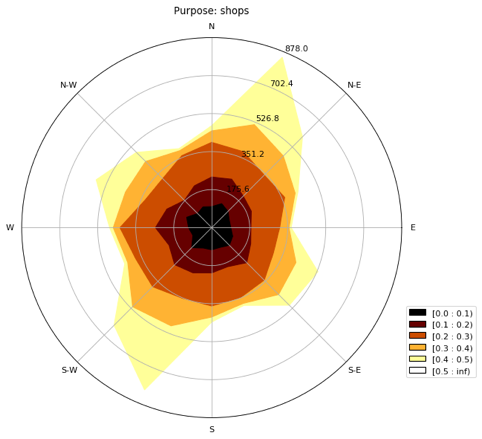
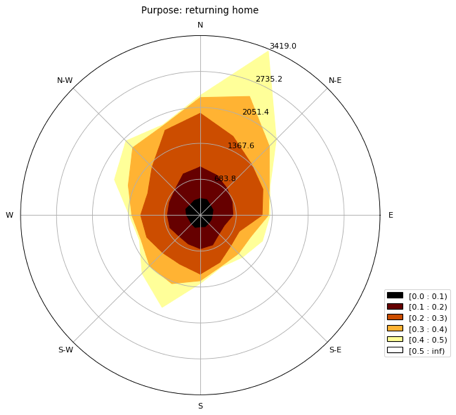
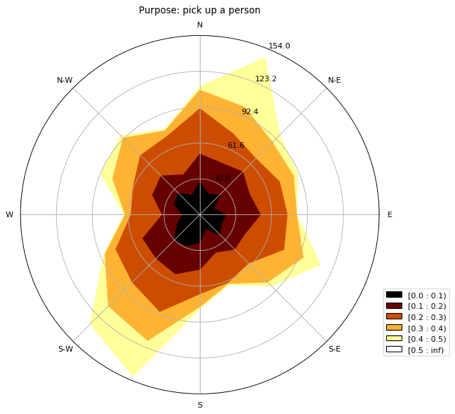
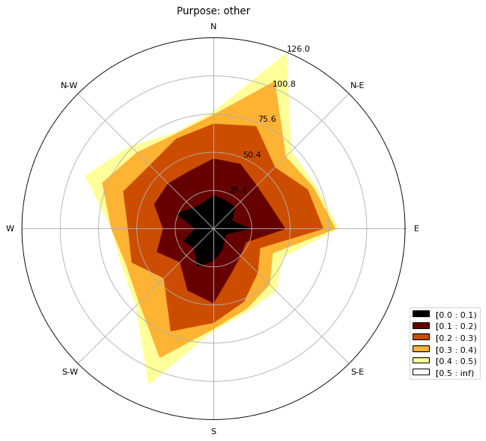
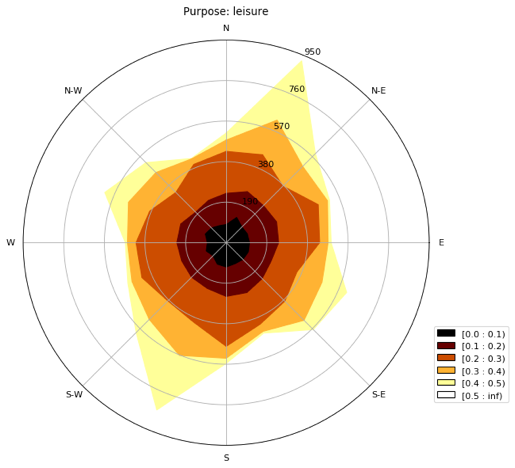
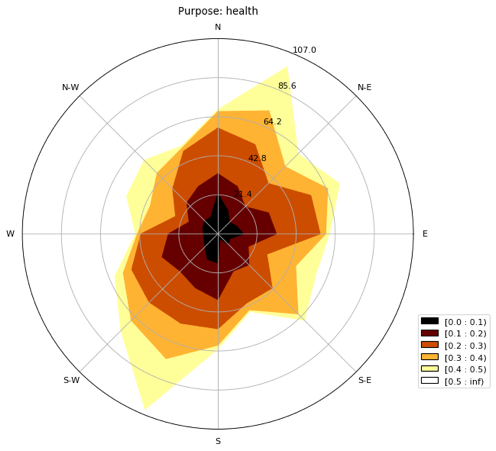
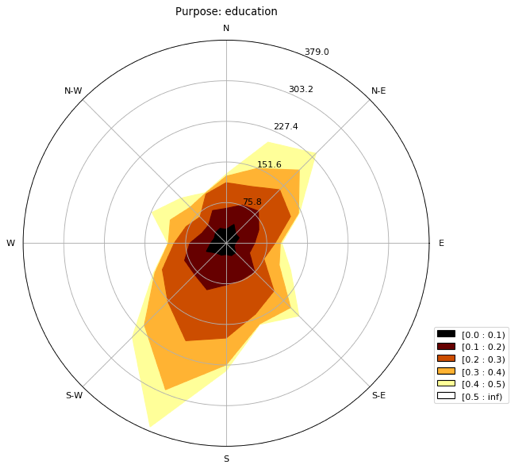
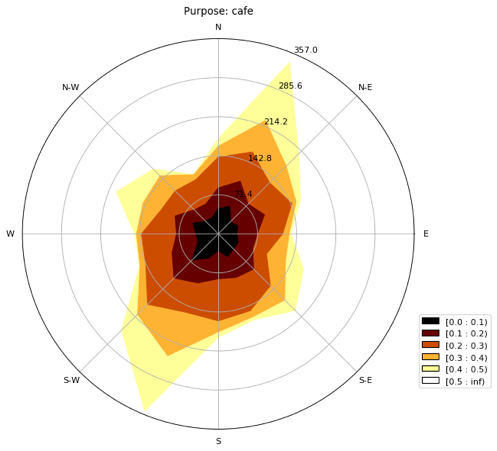
[Some sort of plot for KDE]

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



**Figure 4.X**



**Figure 4.X**

Clustering:

* STSS?

Initially from clustering -> similarity of purposes

*4.3 Modelling:*

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

- “omitted-variable bias (OVB) occurs when a statistical model leaves out one or more relevant variables” (i.e. purpose?)

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

*Classification of purpose of travel:*

Take 1st week -> predict then Take 2nd week -> predict then … compare accuracy scores

Sizes of data:

len(purp\_gdf), len(purp\_city), len(purp\_noncity), len(purp\_rush), len(purp\_nonrush), (71801, 66029, 5772, 39695, 32106)

CANT DO LSTM or RNN as time is not regular

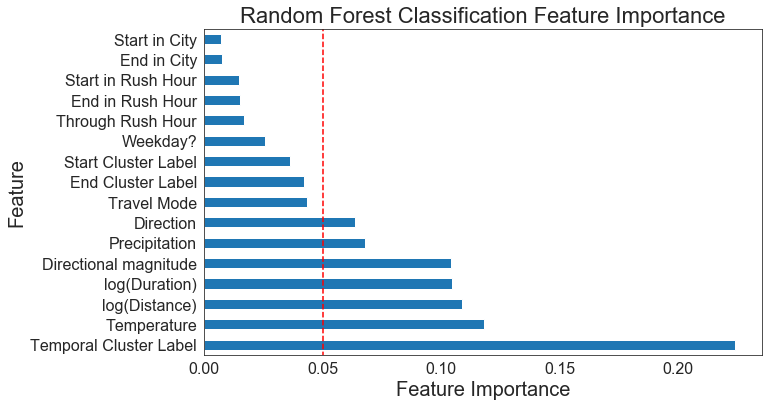
Can do CNN because of grid

*Random Forest:*

Random Forest:

ADD CONFUSION MATRIX OF PREDICTIONS FOR EACH CLASS WITH PROPER NAMES

* Feature importance
* Plot residuals



**Figure 4.X**

|  |  |  |  |
| --- | --- | --- | --- |
|  | RF |  | n |
| All | 0.8116480270099177 |  |  |
| City | 0.8131174162118273 |  |  |
| Non-City | 0.8086500655307994 |  |  |
| Rush | 0.8489446405768798 |  |  |
| Non-Rush | 0.7686380273550376 |  |  |

RF normal:

'purp\_gdf': array([0.80827185, 0.81544217, 0.8151112 , 0.81278586, 0.81742566]),

'purp\_city': array([0.81263883, 0.8150411 , 0.82103977, 0.80835463, 0.8143127 ]),

'purp\_noncity': array([0.80515298, 0.81320451, 0.80322581, 0.82200647, 0.82524272]),

'purp\_rush': array([0.84744822, 0.84464583, 0.84739179, 0.84680851, 0.84749213]),

'purp\_nonrush': array([0.76879298, 0.76944972, 0.76826376, 0.77100142, 0.77171334])

Cross-validated multi-class

{'purp\_gdf': array([0.73622947, 0.73838478, 0.73453903, 0.74275023, 0.74586841]),

'purp\_city': array([0.73694735, 0.74294601, 0.74016885, 0.72891901, 0.73780691]),

'purp\_noncity': array([0.70645161, 0.7516129 , 0.6983871 , 0.69789984, 0.73505654]),

'purp\_rush': array([0.79670736, 0.78838328, 0.7891232 , 0.80281169, 0.79278446]),

'purp\_nonrush': array([0.67220114, 0.66816888, 0.66500593, 0.67852906, 0.66476868])}

Predictions RF:

1.0 21108

0.0 2419

2.0 167

3.0 1

*Comparison of Models:*

Dimensions:

'purp\_gdf': (71801, 22),

'purp\_city': (67177, 22),

'purp\_noncity': (4624, 22),

'purp\_rush': (40342, 22),

'purp\_nonrush': (31459, 62)}

*SVC:*

cv\_scores\_rf

{'purp\_gdf': 0.7364844903988184,

'purp\_city': 0.7402228336866796,

'purp\_noncity': 0.7070773263433814,

'purp\_rush': 0.7959137684969578,

'purp\_nonrush': 0.6675014448083221}

cv\_scores\_svc

{'purp\_gdf': 0.6609411268200042,

'purp\_city': 0.6556903784564031,

'purp\_noncity': 0.7450851900393185,

'purp\_rush': 0.7543003079696537,

'purp\_nonrush': 0.7545752263533038}

cv\_scores\_nn

{'purp\_gdf': 0.7864950411479215,

'purp\_city': 0.7898867788353106,

'purp\_noncity': 0.7234600262123198,

'purp\_rush': 0.8369263126267558,

'purp\_nonrush': 0.7356963976112503}

num\_dims

{'purp\_gdf': (71801, 23),

'purp\_city': (67177, 23),

'purp\_noncity': (4624, 23),

'purp\_rush': (40342, 7),

'purp\_nonrush': (31459, 63)}

Hyper Parameter for SVC:

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

Cs = [0.1, 1, 10]

gammas = [0.01, 0.1, 1]

Best == {'C': 0.1, 'gamma': 0.01}

SVC:

ADD CONFUSION MATRIX OF PREDICTIONS FOR EACH CLASS WITH PROPER NAMES

Results on the test set: RF:

precision recall f1-score support

0.0 0.00 0.00 0.00 0

1.0 0.85 0.89 0.87 15661

2.0 0.00 0.00 0.00 955

3.0 0.00 0.00 0.00 322

4.0 1.00 0.00 0.00 510

5.0 0.76 0.57 0.65 6247

accuracy 0.74 23695

macro avg 0.43 0.24 0.25 23695

weighted avg 0.78 0.74 0.74 23695

Results on the test set: SVC:

precision recall f1-score support

1.0 0.66 1.00 0.80 15661

2.0 0.00 0.00 0.00 955

3.0 0.00 0.00 0.00 322

4.0 0.00 0.00 0.00 510

5.0 0.00 0.00 0.00 6247

accuracy 0.66 23695

macro avg 0.13 0.20 0.16 23695

weighted avg 0.44 0.66 0.53 23695

Results on the test set: NN:

precision recall f1-score support

0.0 0.00 0.00 0.00 0

1.0 0.83 0.93 0.88 15661

2.0 0.00 0.00 0.00 955

3.0 0.00 0.00 0.00 322

4.0 0.00 0.00 0.00 510

5.0 0.73 0.64 0.68 6247

accuracy 0.79 23695

macro avg 0.26 0.26 0.26 23695

weighted avg 0.74 0.79 0.76 23695

NN:

10,000 iterations/epochs for NN

ADD CONFUSION MATRIX OF PREDICTIONS FOR EACH CLASS WITH PROPER NAMES

Best estimator:

ctivation='relu', alpha=0.1, batch\_size='auto', beta\_1=0.9,

beta\_2=0.999, early\_stopping=False, epsilon=1e-08,

hidden\_layer\_sizes=5, learning\_rate='constant',

learning\_rate\_init=0.001, max\_iter=500, momentum=0.9,

n\_iter\_no\_change=10, nesterovs\_momentum=True, power\_t=0.5,

random\_state=3, shuffle=True, solver='lbfgs', tol=0.0001,

validation\_fraction=0.1, verbose=False, warm\_start=False

'alpha': 0.01,

'hidden\_layer\_sizes': (50, 50, 50),

'max\_iter': 500,

'solver': 'lbfgs'

Predictions NN:

1.0 23136

0.0 559

CNN:

Error Terms & Scores

len(all\_true), len(no\_true), len(only\_svc), len(only\_rf), len(only\_nn), len(all\_but\_svc),len(all\_but\_rf),len(all\_but\_nn)

(13773, 3801, 857, 218, 588, 3427, 848, 183)

[Example results writing] “Most of the misclassified trip were …]