# Data Preparation

The following datasets were used for our cluster analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Dataset** | **Description** | **Source** |
| 1 | Public Transport stops | The Geographical Location of all Bus, Tube and Railway stop in the UK | ASK PHILIP |
| 2 | Car Ownership | Number of households owning 0, 1, 2, 3 and 4 or more cars in each MSOA | Nomis official labour market statistics (nomis 2013), from 2011 census |
| 3 | Commuter Flow Data | The number of people commuting between all MSOA pairs, disaggregated by mode of travel | 2011 census (nomis 2011) |
| 4 | Travel Time Data | Travel time between all MSOA pairs using bus, rail, and car | Quant Project, CASA (Batty and Milton 2019) |

The 1st dataset was a shapefile with geographic points. We conducted a points in polygon analysis to determine the number of stops in each MSOA

The 4th dataset was used to compare [1] the relative accessibility and [2] actual commuting patterns of the different MSOAs.

1. For each MSOA, we calculated the average travel time to all other MSOAs by mode. For example:
2. We join the commuter flow data with the travel time data to get the actual average travel time by mode of the trips originating at each MSOA

We end up with the following variables, all at the MSOA level

|  |  |
| --- | --- |
| Variable (MSOA level) | Description |
| Bus\_stops | No. of Bus stops |
| Train\_stations | No. of train stations |
| Metro\_station | No. of tube stations |
| HH\_owning\_cars\_perc | The % of households owning at least 1 car |
| work\_from\_home\_perc | % of MSOA residents who work from home |
| underground\_metro\_perc | % of MSOA residents who use each of these modes for their commute (the mode assigned to a person is the one that makes up the largest portion of the trip) |
| car\_perc |
| train\_perc |
| bus\_perc |
| taxi\_perc |
| motorcycle\_perc |
| bicycle\_perc |
| on\_foot\_perc |
| other\_perc |
| avg\_time\_from\_origin\_car\_UNWEIGHTED | Calculated using Equation (X) |
| avg\_time\_from\_origin\_bus\_UNWEIGHTED |
| avg\_time\_from\_origin\_rail\_UNWEIGHTED |
| avg\_time\_car | Calculated using Equation (Y) |
| avg\_time\_bus |
| avg\_time\_rail |

# Data Transformation & Standardization

To prepare the data for clustering, we followed the methodology used by the Office for National Statistics to classify output areas (Office for National Statistics 2015). This included [1] transforming the variables and then [2] standardizing them. In line with the ONS, we adopted two methods of transformation and 3 methods of standardization, yielding 6 different datasets.

## Transformation

Initial exploration of the variable distribution showed that many of the variables were skewed (Figure X). We applied log transformation and Yeo Johnson transformation to obtain more normally distributed variables. The ONS used a box-cox transformation, but this does not work when there are zeros in the data, and so we used the Yeo Johnson transformation as it can handle zeros (Yeo and Johnson 2000).

Original Variable Distributions

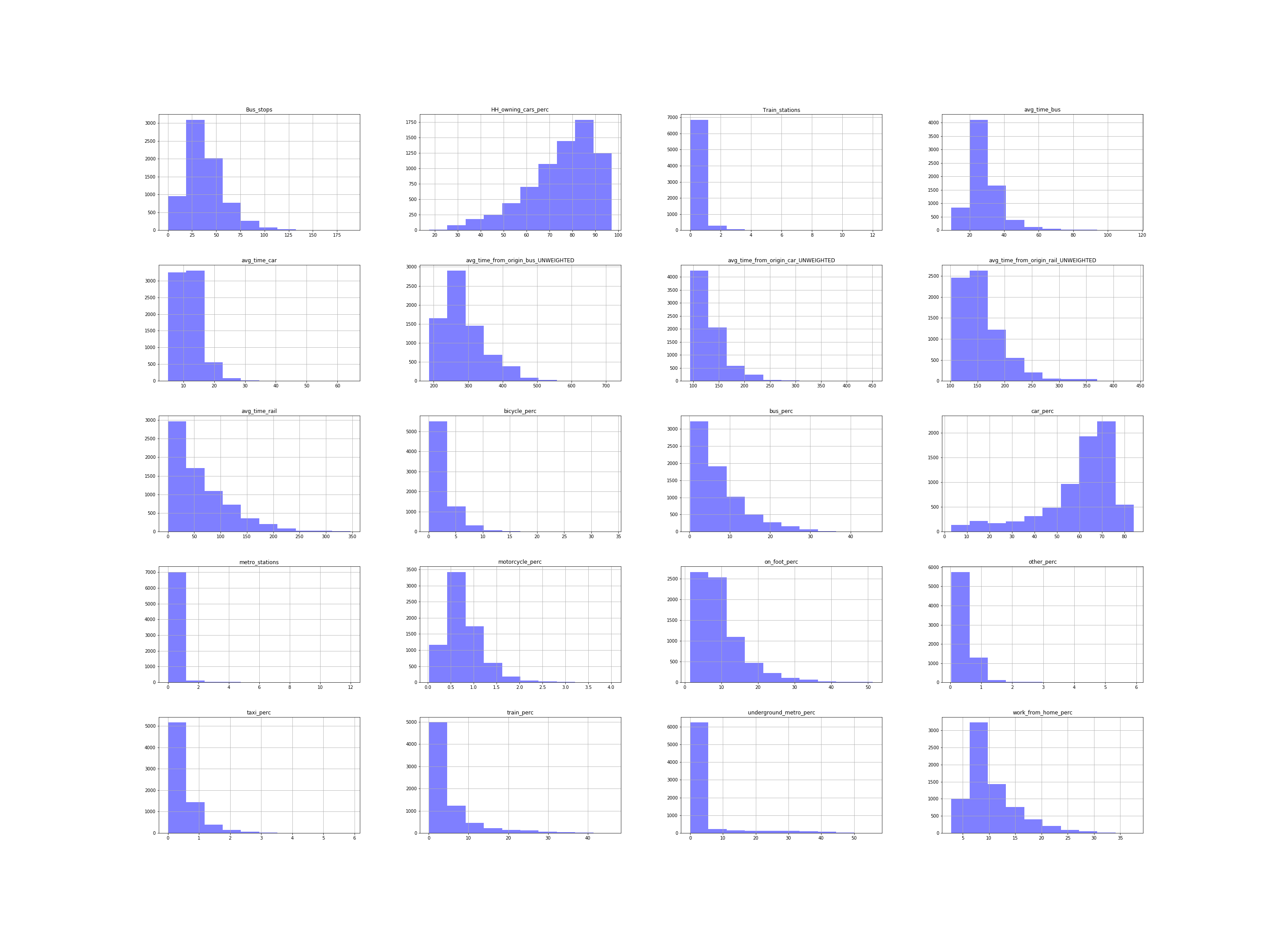


Figure 1: Original variable distributions

Both transformations resulted in more normal distributions for most of the variables, but not all of them. An example is the distributions of unweighted travel times for all bus, rail and car after the Yeo-Johnson transformation (figure 2).

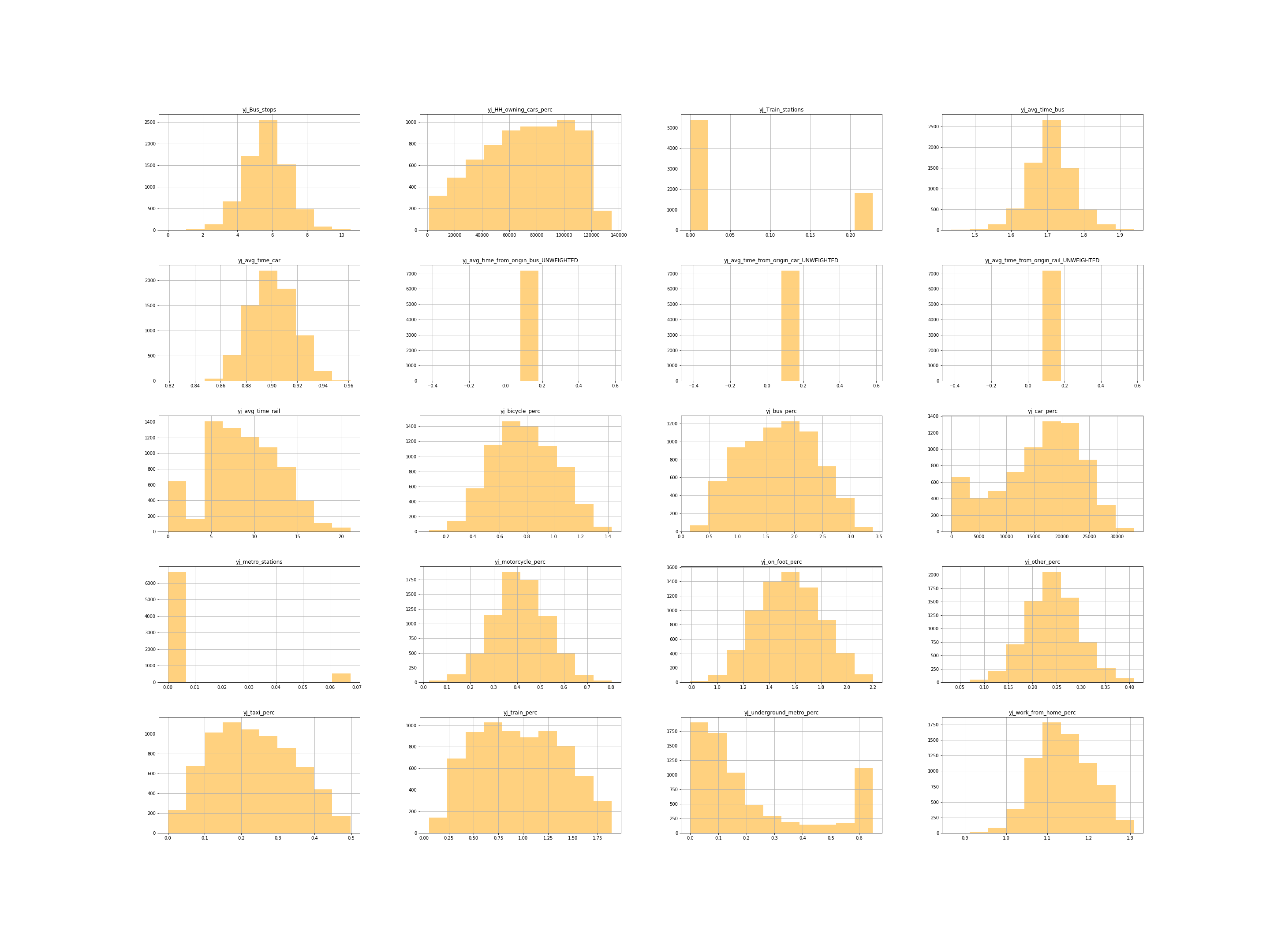


Figure 2: Variable distributions after Yeo-Johnson transformation

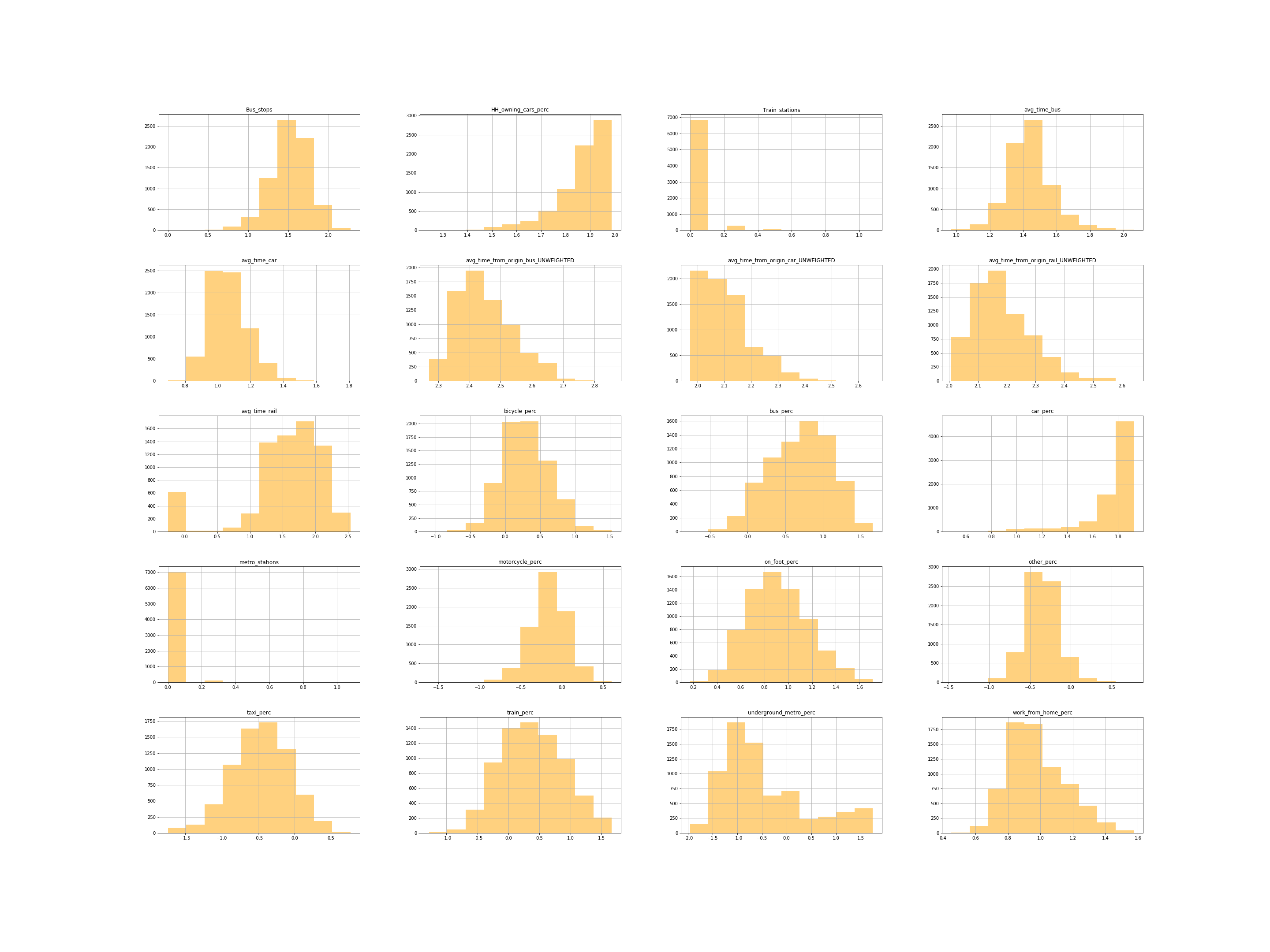


Figure 3: Variable distribution after log transformation

## Standardization

The variables we were using had different units and ranges. This is an issue with clustering since variables with the largest size or variability have the biggest influence on the clustering algorithm, especially k-means (Bin Mohamad and Usman 2013). Data standardization is carried out to adjust the relative weight of the variables to avoid this issue (Milligan and Cooper 1988).

We carried out three different standardization techniques on our transformed data:

### z-score

### Range

All variables are standardized to have values between 0 and 1

### Inter-decile range (IDR)

IDR standardization is more suited to data with extreme outliers than range standardization as it uses the 10th and 90th percentile instead of the maximum and minimum values.

# Data Analysis

## Clustering

We use three different clustering algorithms and compare the results

### k-means

### hierarchical (agglomerative)

### DBSCAN

In total we have 2 transformation techniques, 3 standardization techniques and 3 clustering techniques, leading to 18 combinations. We analyze the results of all 18 combinations and check the histograms of each. Our results showed that DBSCAN consistently failed to assign MSOAs to reasonable clusters, with all results showing only 2 clusters with most MSOAs assigned to one of them. This can be attributed to the ‘curse of dimensionality’, where the distance between different pairs of points decreases as the number of dimensions increases (Steinbach, Ertöz, and Kumar 2004). As DBSCAN works by grouping points within a certain radius of each other, slight variations in that radius has a big effect on the number of points included when dealing with high-dimensional data.

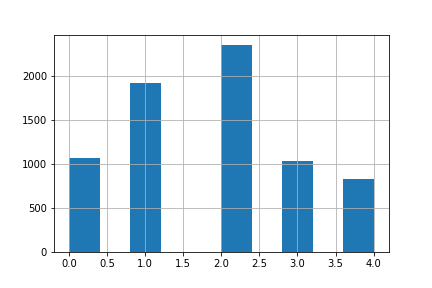
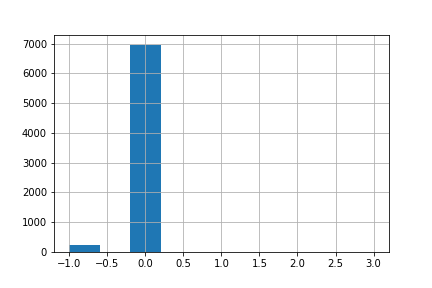


Figure X: Histograms showing Number of MSOAs assigned to each cluster. LEFT (log\_range\_DBSCAN) &) RIGHT (log\_range\_kmeans)

K-means and Hierarchal clustering algorithms provided better results, which were compared by looking at the variable distribution of each cluster and maps of the clustering results to see if they made sense or not.

To improve interpretability of results, we decided to remove some variables and redo the cluster analysis. Variables that were highly correlated or that were deemed unnecessary (i.e. motorcycle\_perc) were removed. This was done twice with only 14 of our initial 20 variables being used to produce the final clustering results.

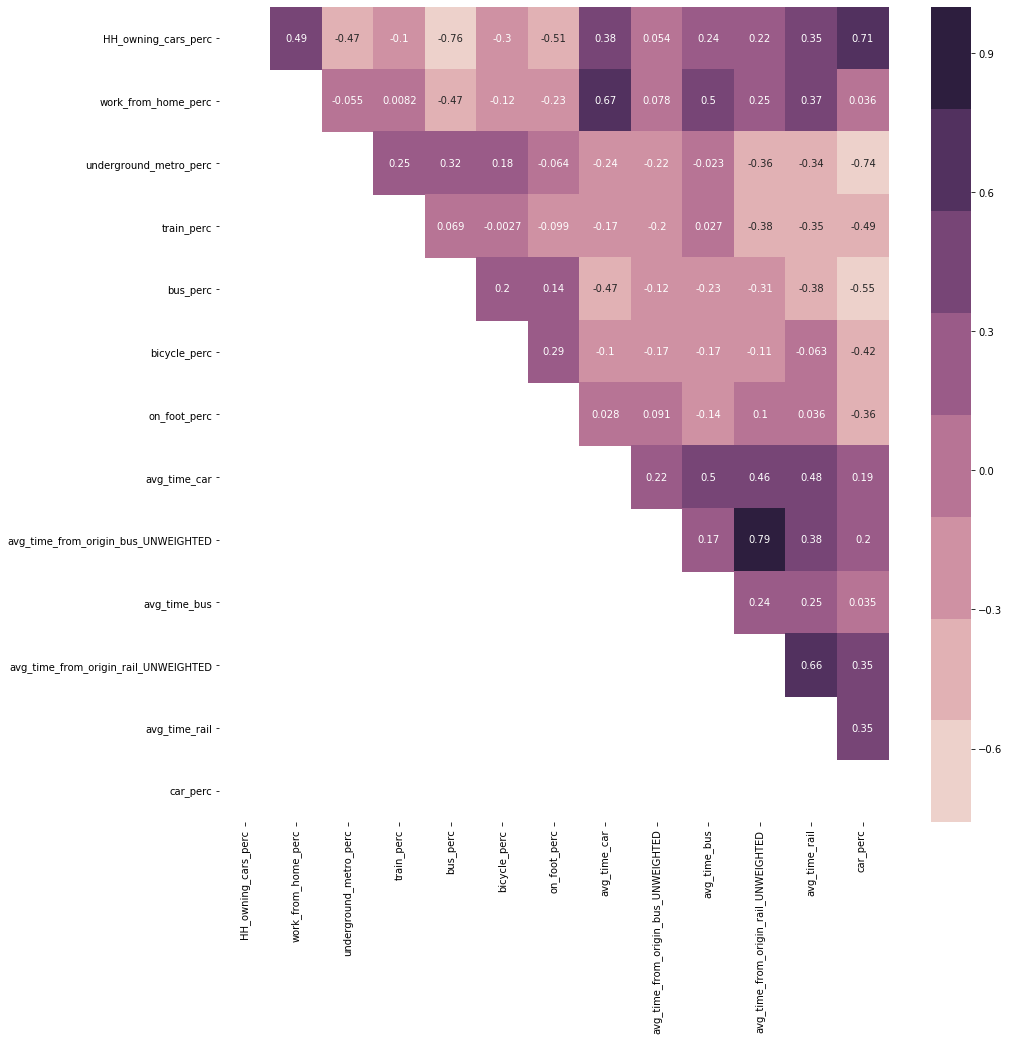


Figure X: Correlation of variables used to produce clustering results

In the end we settled on the results produced by:

log (transformation) -> z-score (standardization -> kmeans (clustering)

The output was 5 clusters with distinct combinations of variable characteristics, as can be seen from the variable averages in each cluster (Figure X)

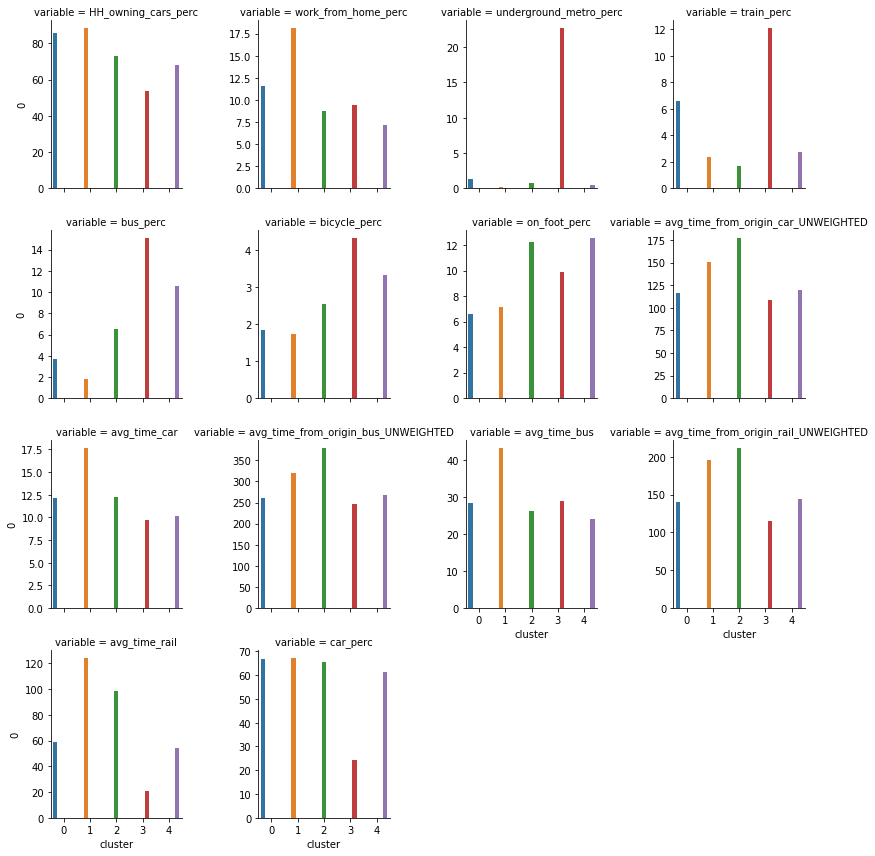


Figure X: Variable averages in each cluster

Table X: Cluster Descriptions

|  |  |  |
| --- | --- | --- |
| ***Cluster*** | ***Plot Colour*** | ***Extended Description*** |
| 0 | Blue | **Good train accessibility but car dependant**: The cluster is composed of rural areas that surround land-locked urban areas. They are mainly in the center of the UK, compared to rural areas in cluster 2 which are on the outskirts. This central location is reflected in their relatively better accessibility scores across all transport modes with the second best accessibility scores for these. The clusters benefit from being on train routes with the second highest train usage but that is the only mode of public transport that they are serviced by. As a result, the cluster is associated with high car ownership and usage, followed by train and walking. |
| 1 | Yellow | **Solely car dependant**: The cluster is made up of rural areas far from the cities. They have no public transport options and people depend on cars to move around. They have poor accessibility even by car, and this could be due to a lack of direct road connections between them and other parts of the country. The cluster is found on the periphery of cluster 2, which is itself made up of coastal cities (like Newcasle and Cornwall) with poor accessibility |
| 2 | Green | **Lack of accessibility across all Transport modes -** This cluster shows the third highest usage of bus, bicycle and walking to work, but has the lowest train usage, working from home and all around accessibility. The most popular modes to travel to work are by car, by walking and bus but the lack of accessibility across all modes and little train usage is the defining feature. This can be found in coastal towns and cities such as Newcastle, Cardiff and Blackpool which might suggest they are at the end of train lines and other transport networks and therefore lack connectivity apart from internal bus usage. |
| 3 | Red | **High public transport and good accessibility** - The cluster is associated with high usage of public transport including the underground/metro/tram, train and bus. It is noted to have very good accessibility to all MSOAs through all transport modes therefore highlighting the ability of people to easily move around. This cluster dominates London, but can also be found in the centre of some MSOAs in big cities like Manchester and Birmingham. The cluster suggests that the transport profile of London is different to the rest of the UK and can only otherwise be found in high accessibility centres of large cities. |
| 4 | Purple | **Car reliant but high public transport** - This cluster has high car usage but is notable for the large number of people who use the bus and walk to work. These MSOAs also have a high degree of accessibility but the overall transport profile is more shifted towards cars than the previous cluster. This is found in large Urban areas across the UK such as Manchester and Birmingham, suggesting that the main differnece between these and London is the degree of usage of public transport with the main difference occuring due to the lack of usage of an underground/metro/tram. |

## Classification

The next step was to run a classification analysis in order to determine whether socioeconomic characteristics can be used to predict our clustering results. In other words, are the variations in transport characteristics part of a larger socioeconomic divides across the UK. After checking for variable correlation, we used the variables in Table X in our classification

Table X: Variables used in Classification

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Description** | **Source** |
| Net annual income (£) | Average net annual income in 2018 | Office for National Statistics (ONS 2020) |
| Pop\_Per\_Hectare | Population density | Office for National Statistics (ONS 2019) |
| percent\_unemployed |  | 2011 Census (nomis 2011) |
| percent\_at\_or\_above\_qual\_  level\_4 | The % of people living in the MSOA that achieved Qualification Level 4 or above |
| perc\_households\_owned | The % of households in the MSOA that are owned |
| avg\_number\_of\_bedrooms |  |
| perc\_bad\_health | The % of residents who suffer from bad or very bad health |
| perc\_employed\_females\_  working\_fulltime | The % of the labor force that is made up of employed females working >35 hours per week |
| mean\_age |  |
| perc\_christian |  |
| perc\_non\_religious |  |

A Random Forest classification with 100 trees and no pruning was done. Oshiro, Perez, and Baranauskas (2012) note that beyond a certain threshold of trees, there is no improvement in model performance, and suggest a value between 64 and 128. Pruning is done in decision trees to avoid over-fitting. This is not necessary in random forest which use random selection of features and so produces different trees that are not correlated with each other (Breiman 2001).

### Results

Table X: Random Forest Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| 0 | 0.83 | 0.52 | 0.64 |
| 1 | 0.59 | 0.07 | 0.12 |
| 2 | 0.77 | 0.76 | 0.76 |
| 3 | 0.60 | 0.81 | 0.69 |
| 4 | 0.62 | 0.81 | 0.70 |
|  |  |  |  |
| accuracy |  |  | 0.65 |
| macro avg | 0.68 | 0.59 | 0.58 |
| weighted avg | 0.66 | 0.65 | 0.60 |

The accuracy score shows that 65% of MSOAs are classified correctly by the model. The precision value shows that the model was prone to false positives, particularly with cluster 1, 3, and 4. The recall score shows that the model was unable to give the correct value for cluster 1, meaning that most MSOAs in cluster 1 were misclassified. The confusion matrix (Figure X) shows that only 19 out of 284 MSOAs in cluster 1 were correctly classified.

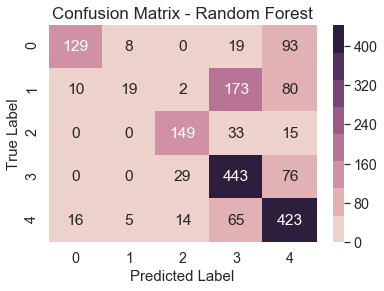


Figure : Confusion Matrix for Random Forest Classification

### Feature Importance

To understand which variables were most predictive of transport characteristics, we use feature importance. We opted for permutation importance (Altmann et al. 2010) as it is less biased in its interpretation of feature importance than the default sklearn feature importance which is based on gini impurity.

The importance is based on calculated the coefficient of determination (R2), randomly reshuffling one variable, then recalculating R2. The decrease in model performance (difference in R2) is a measure of the variable’s importance.

A screenshot of a cell phone

Description automatically generated

We added a random variable to the model to see if any variable performed worse than it, but none did. Population density is the most important feature (Figure X), which is not surprising given that public transport is mostly associated with urban agglomerations. Religious variables, unemployment rate have little predictive power, indicating that they may show uniform distribution across the study area