Data Preparation

The aim of our study was to understand the different transport profiles that are seen across England and Wales. Therefore, in terms of data for the initial clustering algorithms our aim was to find as much transport information we could. This was primarily undertaken at the MSOA level given that this is the lowest UK geographic level for which relevant transport data could be found. Therefore, the datasets collected were such that these would provide us with an idea of how different groups of people commute in England and Wales commute. This includes those shown in table X below, which provides a brief description of each dataset along with the source from which it was obtained.

No.	Dataset	Description	Source
1	Public Transport stops	The Geographical Location of all Bus, Tube and Railway stop in the UK	Data.gov (DfT 2014)
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 is a csv containing latitude and longitude values for all transport stops in the UK. From this our primary concern was for bus stops, train stations and tube/tram/metro stations in England and Wales. To obtain the information in terms of the number of different stops within each MSOA, this was merged with a shapefile of the UK MSOAs and a points-in-polygon analysis was performed in R, as can be seen in figure X. The results from this were then outputted to another csv to latter be merged.

Figure 1 - Showing the points of all a) bus stops, b) tram/metro and underground stations and c) all train stations across England and Wales

The 3rd dataset contains travel to work data from one MSOA to another in the form of Origin – Destination pairs for each travel mode, as shown in Fig X.

```
In [3]: transport
```

Out[3]:

	Area of residence	Area of workplace	All categories: Method of travel to work	Work mainly at or from home	Underground, metro, light rail, tram	Train	Bus, minibus or coach	Taxi	Motorcycle, scooter or moped	Driving a car or van	Passenger in a car or van	Bicycle	On foot	Other method of travel to work
0	E02000001	E02000001	1506	0	73	41	32	9	1	8	1	33	1304	4
1	E02000001	E02000014	2	0	2	0	0	0	0	0	0	0	0	0
2	E02000001	E02000016	3	0	1	0	2	0	0	0	0	0	0	0
3	E02000001	E02000025	1	0	0	1	0	0	0	0	0	0	0	0
4	E02000001	E02000028	1	0	0	0	0	0	0	1	0	0	0	0

Figure 2: Commuter Flow Data from 2011 Census (MSOA level)

To transform the data to get information about each individual MSOA and their transport use, this data was grouped by area of residence and the sum for each transport mode was calculated (fig X). This allows us to get the variation in mode share per MSOA. These sums were then turned into % to allow for comparison between MSOAs.

```
In [2]: #merge by origin MSOA and sum over all transport options
        transport_msoa = transport.groupby("Area of residence", as_index=False).sum()
        transport_msoa
```

Out[2]:

	Area of residence	All categories: Method of travel to work	Work mainly at or from home	Underground, metro, light rail, tram	Train	Bus, minibus or coach	Taxi	Motorcycle, scooter or moped	Driving a car or van	Passenger in a car or van	Bicycle	On foot	Other method of travel to work
0	E02000001	4785	701	880	260	265	43	22	135	11	243	2184	41
1	E02000002	2484	159	389	204	366	20	22	1076	87	33	116	12
2	E02000003	4539	313	580	901	482	22	31	1778	103	40	268	21
3	E02000004	2858	182	290	393	285	19	31	1322	92	38	193	13

Figure 2: Commuter Flow Data Grouped by Area of Residence

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:

$$Accessibility\_Bus_i = \sum_{j=1}^{n-1} Time\_Bus_{ij}$$

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

$$Avg\_Commute\_Bus_i = \sum_{j=1}^{n} Time\_Bus_{ij} * Commuters\_Bus_{ij}$$

$i = Origin\ MSOA$
$j = destination\ MSOA$
$n = total\ number\ of\ MSOAs$
$Commuters\_Bus_{ij} = number\ of\ people\ commuting\ between\ i\ and\ j\ by\ bus - 2011\ census$

As a result of this the following variables, as shown in table X, could be measured for each MSOA. The data was subsequently merged into one dataset (all_transport_data.csv) so that we could conduct our cluster analysis.

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

Before clustering could be performed on the data it is noted the algorithms used, of K-Means, DBSCAN and Hierarchical clustering, are sensitive to the initial inputs to the extent that different units, scales and variations are likely to influence to the outcomes. Therefore, the data must be cleaned, transformed and standardized prior to performing these algorithms. We considered 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 [2] standardizing them. They start off with 167 variable which they reduce to 60 by trying different variable transformation and standardization combinations and eliminating correlated variables. Our analysis was similar, starting with 20 variables, narrowing this down to 14. A variation of our work is we apply 3 different clustering algorithms, whereas the ONS only apply 1 because each method produces slightly different outcomes depending on the data. We therefore compare different (transformation + standardization + clustering) combinations and use visual inspection and variable distribution of clusters to choose the combination that we considered to best represent reality. The steps of our methodology are outlined below.

Transformation

Initial exploration of the variable distribution showed that many of the variables were skewed to some degree (Figure X). Using skewed data in clustering analysis is likely to results in clusters that are not reflective of the true underlying groups of data because extremes and outliers will likely influence the way in which groups are formed, especially for those using distance based metrics such K-Means (Kumar, et al., 2015). To control for this the data is transformed prior to standardization. However, since each variable is not skewed to the same degree, or necessarily in the same direction, then three different transformations were applied to the data for which the outcomes and results could be compared. This includes Log and Yeo Johnson transformations to obtain more normally distributed variables. While the ONS used a box-cox transformation, 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

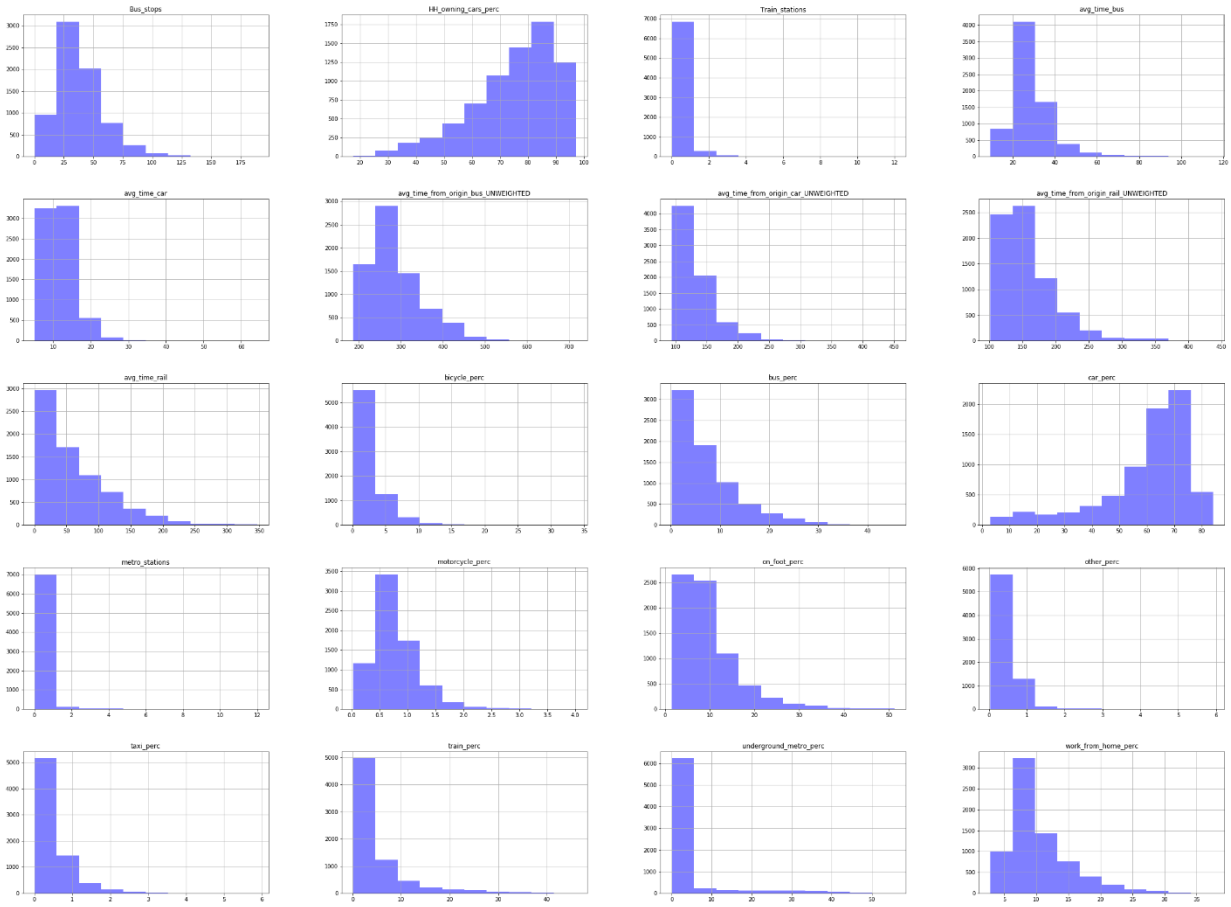


Figure 3: Original variable distributions

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).

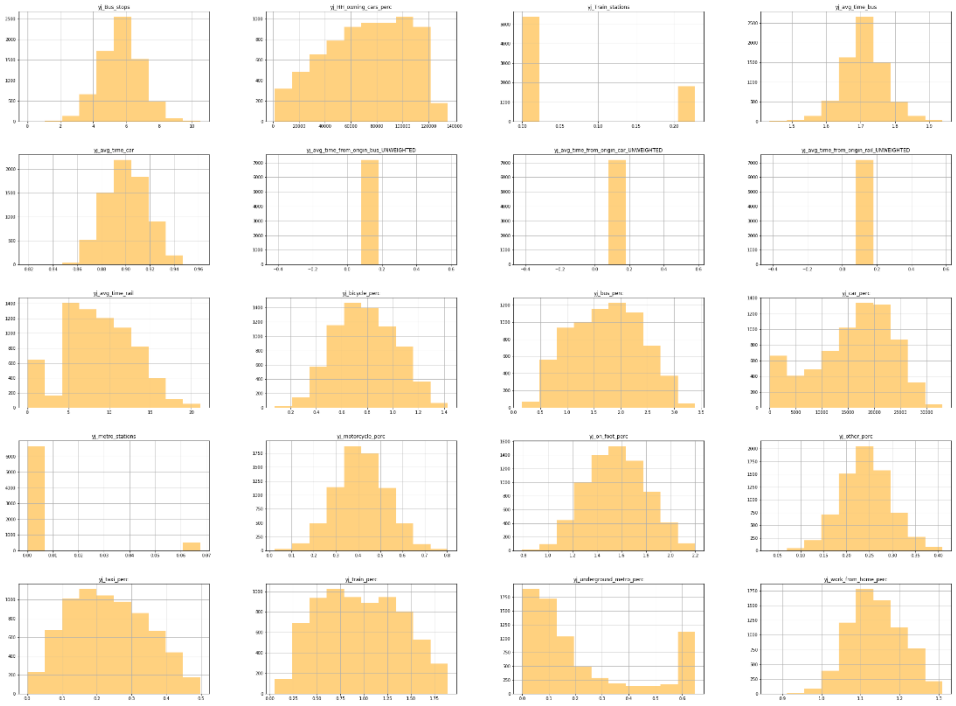


Figure 2: Variable distributions after Yeo-Johnson transformation

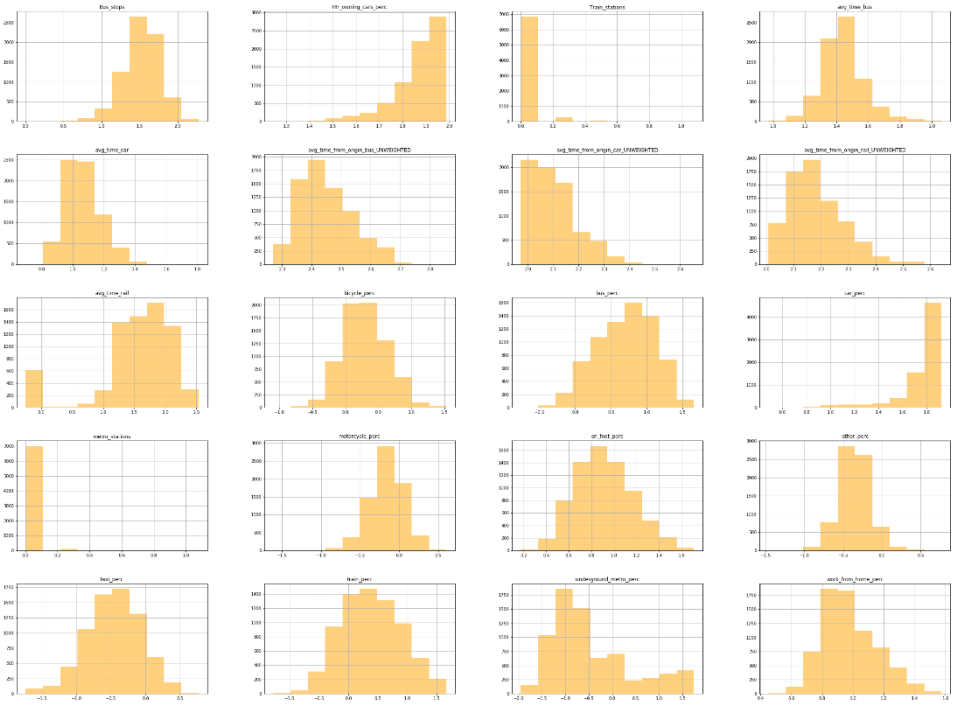


Figure 3: Variable distribution after log transformation

Standardization

We tried to set as many variables as possible in the same units but that was not possible for all values and therefore there were discrepancies in units and ranges for each of the variables. For example, travel time which cannot be changed into a percentage value. 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). This rescales all of the variable so that they are within the same range, thereby preventing large-scaled variables from having a bigger influence on the clustering algorithms

The standardization technique used normally depends on the distribution of the data, but since there was no consistent distribution of variables in either transformation, we carried out three different standardization techniques on our transformed data. This allowed us to compare the cluster outputs resulting from different transformation and standardization combinations. The following standardization techniques were used:

z-score:

$$z = \frac{x - \mu}{\sigma}$$

$x = raw\ score$
$\mu = variable\ mean$
$\sigma = variable\ standard\ deviation$

The resulting values show many standard deviations each value is from the mean. The mean is 0 and negative values are those smaller than the mean.

Range:

$$r = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$x = raw\ score$
$x_{min} = small\ value\ for\ variable$
$x_{max} = largest\ value\ for\ variable$

All variables are standardized to have values between 0 and 1

Inter-decile range (IDR):

$$i = \frac{x - x_{med}}{x_{90} - x_{10}}$$

$x = raw\ score$
$x_{med} = variable\ median$
$x_{90} = variable\ 90^{th}\ percentile$
$x_{10} = lvariable\ 10^{th}\ percentile$

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

These algorithms differ in their underlying assumptions and the way they perform the analysis, and so they produce different results. Both *k-means* and *hierarchical clustering* require specifying the number of clusters, for which the optimal number of clusters was found using elbow plots and silhouette scores after multiple runs of each cluster value. Elbow plots are based on minimizing the Within-Cluster Sum of Squares (WSS). The higher the number of clusters, the lower the WSS, as the variation becomes 0 when the number of clusters is equal to the number of points. The elbow method ensures that we do not overfit to the data by choosing a number of clusters after which the improvement is marginal.

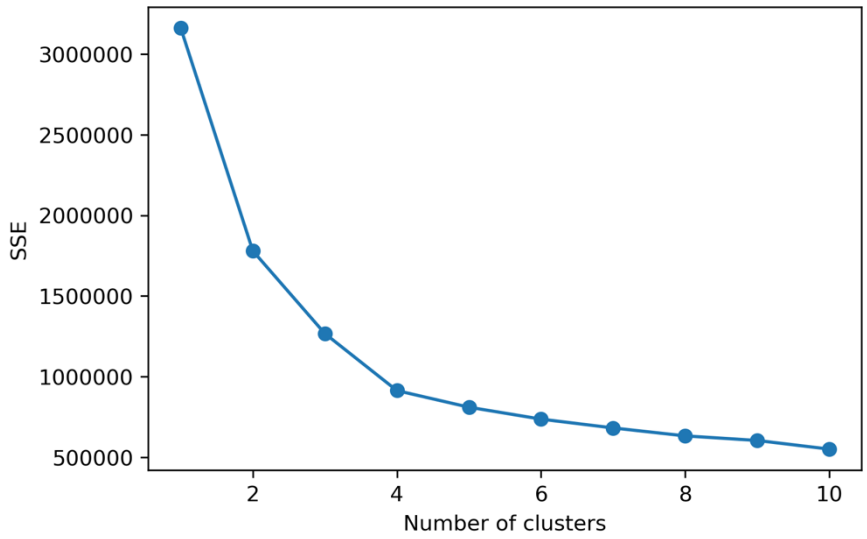


Figure X: elbow plot for log-idr-kmeans

Silhouette scores give us an indication of how compact and separated from each other the clusters are (Chen et al. 2002). The closer points in a cluster are to each and the further away they are from points in the nearest cluster, the better the silhouette score. Comparing silhouette scores allows us to choose the best number of clusters.

The two methods measure different things and so did not always give the same results. We therefore used them both for guidance and selected the number of clusters that provided the best results.

DBSCAN is a density-based clustering algorithm where clusters are formed if a minimum number of points are within a given distance (ε) of each other. Unlike, k-means and hierarchal clustering DBSCAN highlights outliers which it does not add to the clusters. The issue with

this is that it does not perform well on high-dimensional data and when there are clusters of different densities, as ε cannot be calibrated to suit different clusters.

Consequently we have 2 transformation techniques, 3 standardization techniques and 3 clustering techniques, resulting in 18 combinations of results. We analyzed each of the results firstly by checking the histograms to see the distribution of MSOAs to clusters in each of the 18 results.

For example, DBSCAN consistently failed to assign MSOAs to reasonable clusters, with all results showing the majority of MSOAs assigned to 1 cluster and the rest identified as outliers. 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.

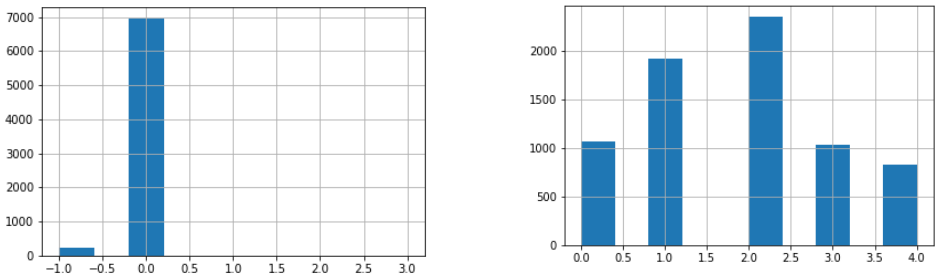


Figure X: Histograms showing Number of MSOAs assigned to each cluster. LEFT (log_range_DBSCAN) &) RIGHT (log_range_kmeans)

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

Variable Selection

To improve interpretability of results, we decided to remove some variables and redo the cluster analysis. Variables that were highly correlated (i.e. motorcycle percentage) were removed.

After further inspection of the clusters we saw that some variables did not appear to influence the clustering results. Therefore, we removed variables in sequence to see what affect, if any, they had on the clustering results. The first step was to remove other and taxi (commuter mode shares). The results appeared to be unaffected and had improved interpretability. Then the number of bus stops, train stations, and metro stations was removed. Even though the last three variables appear intuitively to be of great importance, we hypothesized that they were not having an effect because the accessibility scores by mode were making them redundant.

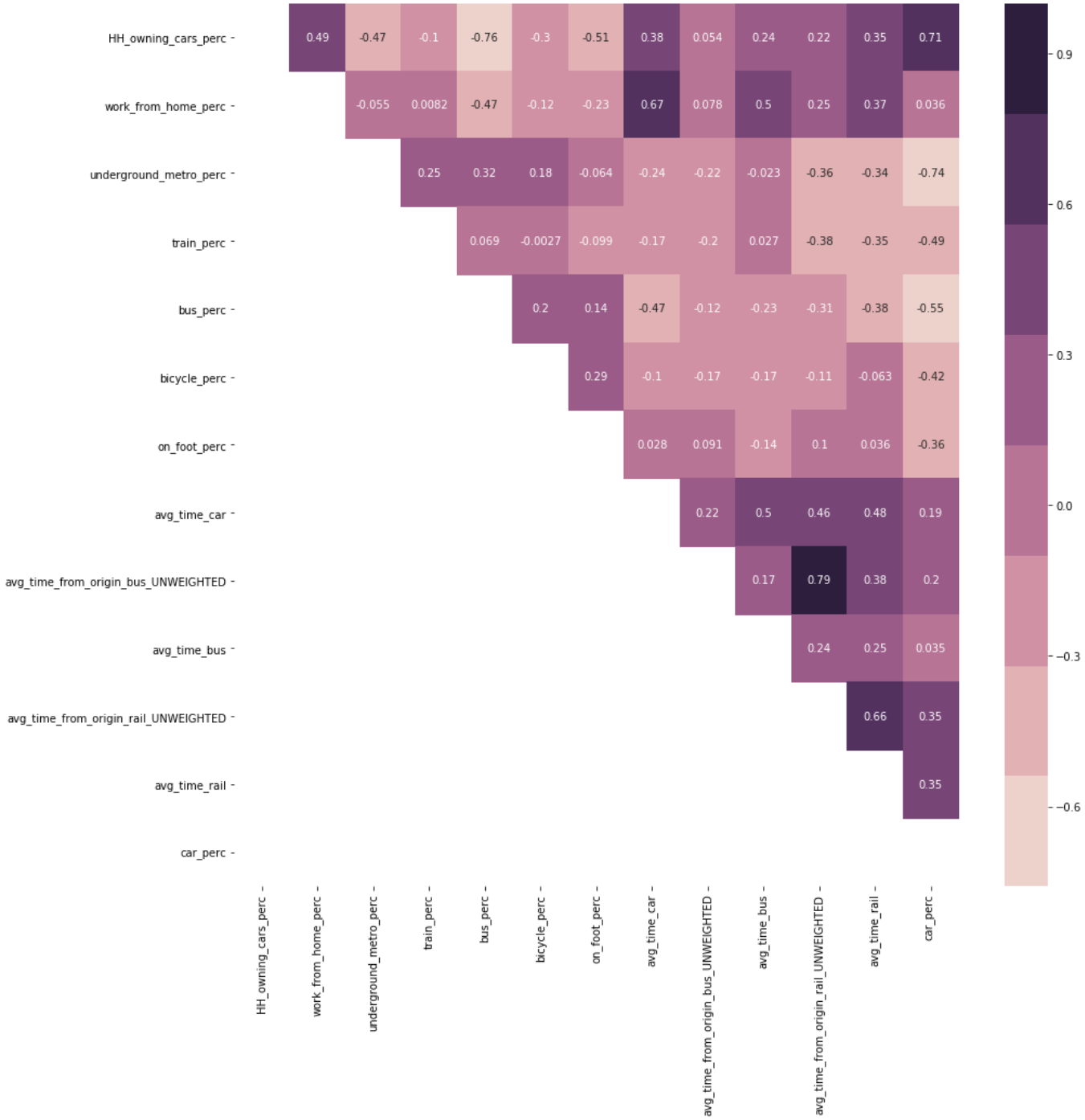


Figure X: Correlation of variables used to produce clustering results

Comparing Results

Evaluation of our initial clustering results enabled us to filter down the initial 20 variables to 14. We then reran our clustering algorithms and ran the previous steps again. Thus, we removed the results which had a skewed distribution in the number of points assigned to each cluster, did not appear to represent geographical reality and did not appear to produce meaningfully

distributed clusters. The result we decided that best represented differentiation in clusters and mapped onto our knowledge of existing geography in the UK was the:

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

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

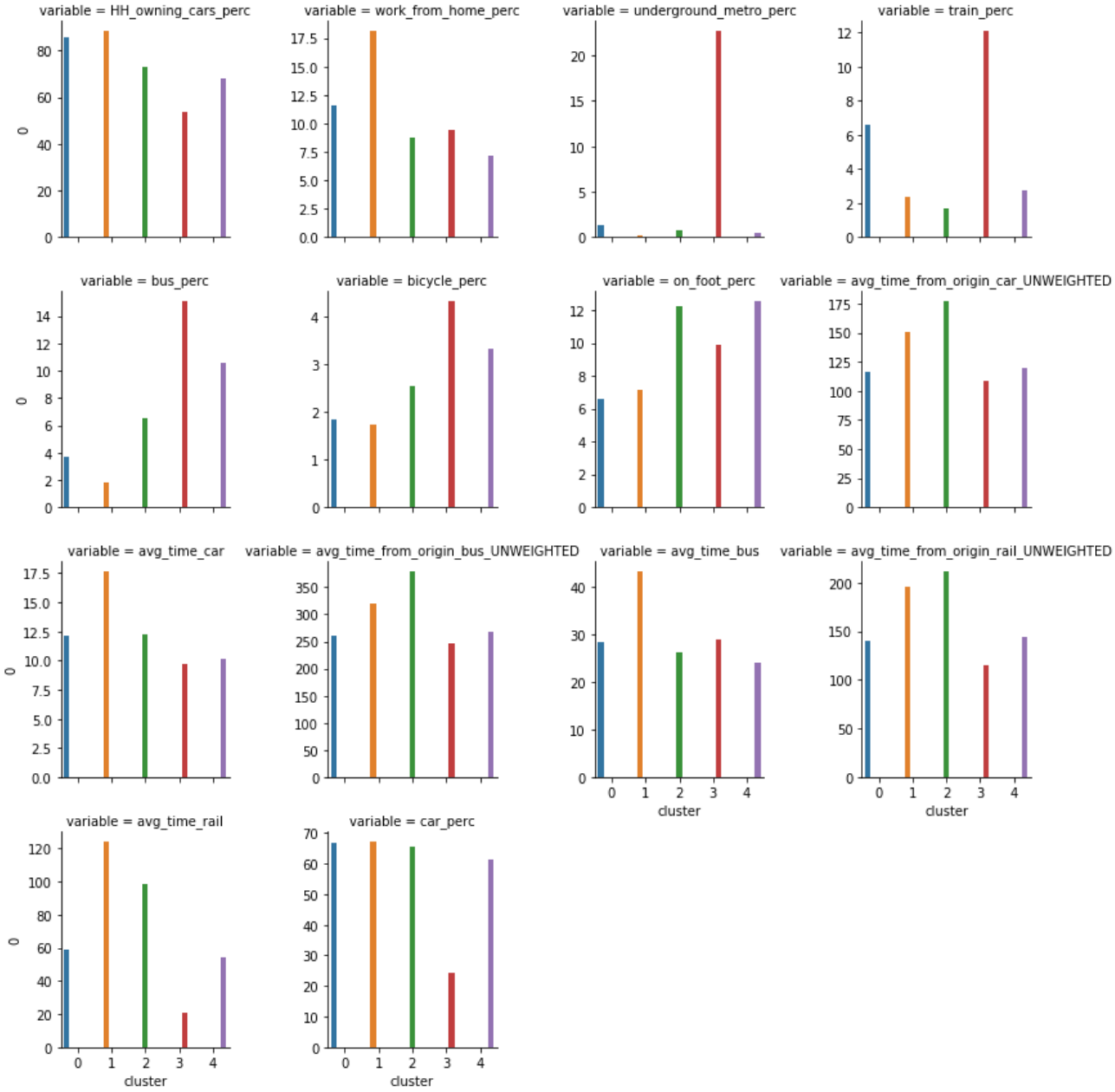


Figure X: Variable averages in each cluster

Table X: Cluster Descriptions

Cluster	***Plot Color***	***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

Given that our aim was to not only create different transport profiles but to understand the demographic factors related to these results, the next step was to run a classification analysis. In other words, are the variations in transport characteristics part of a larger socioeconomic divides across the UK. Previous research i.e. (Titheridge, et al., 2008), (Pinjari, et al., 2007), (Ferdous, et al., 2011) suggests that demographic characteristic such as income, unemployment, education level and sex can influence transport travel mode choice. Therefore, based on available data at the MSOA level, the following variables (Table X) were chosen to understand how they are related to the transport profile classes that we identified

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.

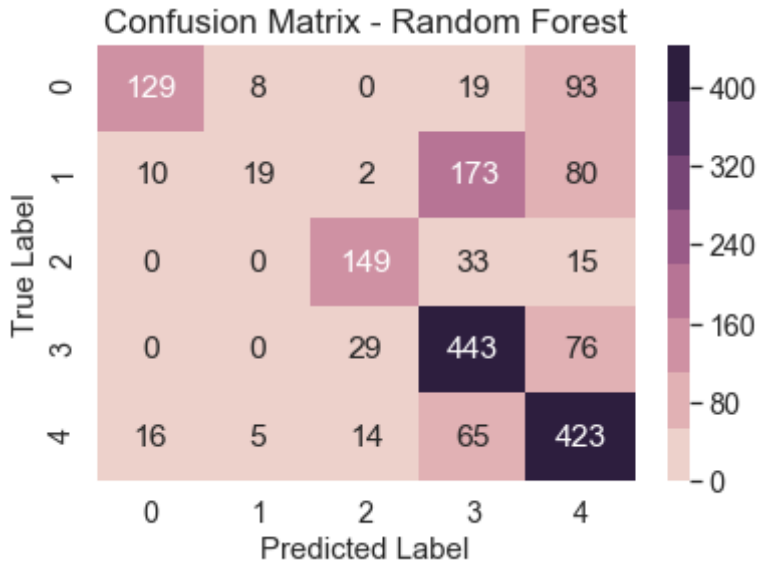
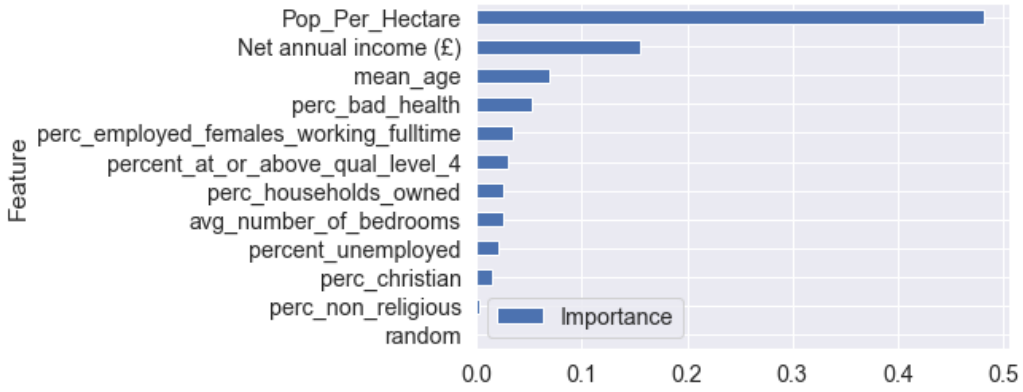


Figure 4: Confusion Matrix for Random Forest Classification

Feature Importance

To understand which variables were most related to transport characteristics, we use feature importance. The default feature importance, which is based on gini impurity, is biased, especially when variables vary in scale; continuous and high cardinality variables (variables with many unique values) tend to rank higher even if they are no more informative than other variables (Strobl et al. 2007). 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.

The importance is based on calculated the coefficient of determination (R^2), randomly reshuffling one variable, then recalculating R^2. The decrease in model performance (difference in R^2) is a measure of the variable's importance.



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

Word count: 3246
Word count – tables (don't count anyway) 2405