Identifying Transport Profiles Of England and Wales: Data Analysis and Web Visualisation

Cheyne Campbell
Alicja Kotarba
Hussein Mahfouz
Nathanael Sheehan
Philip Wilkinson

Group 0

Spatial Data Capture, Storage, and Analysis
University College London
27 May 2020

Word Count: 5,214

Table of Contents

Project Links

Website:
http://dev.spatialdatacapture.org/~ucfncjc/spidercartographers/index.html

GitHub:
https://github.com/natesheehan/spidercartographers

API:
http://dev.spatialdatacapture.org:8717/

Background

The aim of our analysis and website is to be able to understand how transport usage within England and Wales differs and what factors are likely to affect this. A deeper understanding of the geographical differences in transport choice could inform national and local travel policies to move towards more sustainable modes of transport. To this end, we have undertaken a geodemographic approach at the MSOA level to create different groups of transport profiles. Further, we performed classification analysis to understand what demographic factors are related to these profiles. The results, and the subsequent visualisations, show how transport usage profiles vary across England and Wales. This analysis fits within the broader themes of existing transport literature, discussing the importance of transport and how we can influence individuals' transport choice. The analysis also extends the existing geodemographic analysis literature to the domain of transport.

Transport supports our work, education and social interaction among other things (Delbosc, 2012), but our choice of mode has wide ranging repercussions. As noted by Buehler (2011), our transport behaviour affects global warming, environment pollution and our dependence on oil, while Laverty et al., (2013) provide evidence that shifting towards public transport and more active modes of travel are associated with reductions in risks of cardiovascular problems and diabetes. It is therefore not surprising that there is considerable research and literature on understanding how we can change transport usage habits towards more sustainable forms of transport. This includes how attitudes, beliefs and individual characteristics may affect transport mode choice. For example, it is noted that educating individuals that their transport choice can make a difference in terms of climate change can influence their regular transport mode (Collins and Chambers, 2005), while perceptions of existing infrastructure, regardless of the actual provision, can also influence choice (Ferdous et al., 2011). Furthermore, the physical environment can influence transport modes; accessibility, density and land use mix are all associated with transport mode choice, even when controlling for demographic characteristics (Pinjari et al., 2007). The overall conclusion of such research is that, in order to improve the usage rates of sustainable modes of transport, policies need to not only make sustainable modes more attractive through time and cost reductions, but to also make sure that less sustainable modes, such as automobile use, is made less attractive (Buehler, 2011). It is therefore important that we first understand how transport usage differs across England and Wales to then more effectively target resources and policies to areas in which they are most likely to be successful.

The methodology employed to tackle this problem is geodemographic analysis. This dates as far back as Charles Booth’s 1903 Poverty Map of London in which he separated populations into different economic groups based on underlying demographic and economic characteristics, and has since been expanded to cover a wide variety of areas (Harris et al., 2007). Traditionally this has been based on census data that is produced decennially and split into different administrative units, but the improvement in data collection and storage, along with analytical tools such as GIS, has allowed data such as household surveys, property

information and loyalty card data to be utilised in such analyses. These advancements have allowed more recent work across both the public and private sectors to cover topics such as health, policing and education to inform public resource allocation to areas that need them most (Adnan et al., 2010), and in retail analysis to understand consumer behaviour and hence allowing for more efficient targeted marketing (Titheridge et al., 2008). The methods used for this include cluster analysis which groups together populations, at an aggregate level, based on chosen variables in a situation in which there is no prior classification (Shelton et al., 2006), and classification analysis to understand other variables and characteristics that may be associated with different groups identified by the cluster analysis. This allows for a comparison between and across groups to understand the way in which they differ and thus subsequently how resources may best be allocated to influence such groups. In terms of transport, however, there has been little utilisation of such methodologies to understand transport profiles and how this could be used to inform transport resource allocation decisions (Titheridge et al., 2008). Therefore, our analysis seeks to fill this gap, linking transport studies and geodemographic analysis.

Data

Data Preparation

For the clustering analysis our initial aim was to find as much transport data as we could at the MSOA level, since this is the lowest UK geographic level for which relevant transport data could be found. The data collected is shown in table 1 below.

Table 1: Datasets used for clustering.

No.	***Dataset***	***Description***	***Source***
1	Public Transport Stops	The geographical location of all bus, tube and railway stops in the UK	Data.gov (Department for Transport, 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 (Office for National Statistics, 2013), from 2011 census
3	Commuter Flow Data	The number of people commuting between all MSOA pairs, disaggregated by mode of travel	2011 census (Office for National Statistics, 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 contained latitude and longitude values for all transport stops in the UK. To obtain 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 (see figure 1). The results from this were then outputted to another csv.

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 (see figure 2).

```
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).

This data was grouped by area of residence and the sum for each transport mode was calculated (see figure 3). These sums were then turned into percentages of overall travel from MSOA to allow for mode-share comparisons. This data was also uploaded to MySQL for the API to map the flows 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 3: 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:

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

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.

$$Avg\_Commute\_Bus_i = \sum_{j=1}^{n} Time\_Bus_{ij} * Commuters\_Bus_{ij} \quad (2)$$

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

The variables in table 2 were subsequently obtained for each MSOA and merged into one dataset (all_transport_data.csv) so that we could conduct the cluster analysis.

Table 2: Variables used for 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 (1)
avg_time_from_origin_bus_UNWEIGHTED	
avg_time_from_origin_rail_UNWEIGHTED	
avg_time_car	Calculated using Equation (2)
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 inputs that have different units, scales and variations. Therefore, the data must be cleaned, transformed and standardized. For this, we considered the methodology used by the Office for National Statistics to classify output areas (Office for National Statistics, 2015), which included transforming and standardizing the variables prior to clustering. A variation of our work is that we apply three different clustering algorithms, whereas the ONS only apply one. We do this because each method produces slightly different outcomes depending on the data. We therefore compare different transformation, standardization and clustering combinations using visual inspection and variable distribution to choose the combination that best represents reality.

Transformation

Initial exploration of the variable distribution showed that many of the variables were skewed (see figure 4). Using skewed data in cluster analysis is likely to results in clusters that are not

reflective of the underlying groups of data because extremes and outliers will likely influence cluster formation, especially for algorithms using distance-based metrics (Kumar et al., 2015). Therefore, the data is transformed prior to standardization. Since each variable is not skewed to the same degree, or necessarily in the same direction, two different transformations were applied to the data for which the outcomes and results could be compared. This includes Log and Yeo Johnson transformations, due to their ability to handle zeros (Yeo and Johnson, 2000). The results of these can be seen in figures 5 and 6 below.

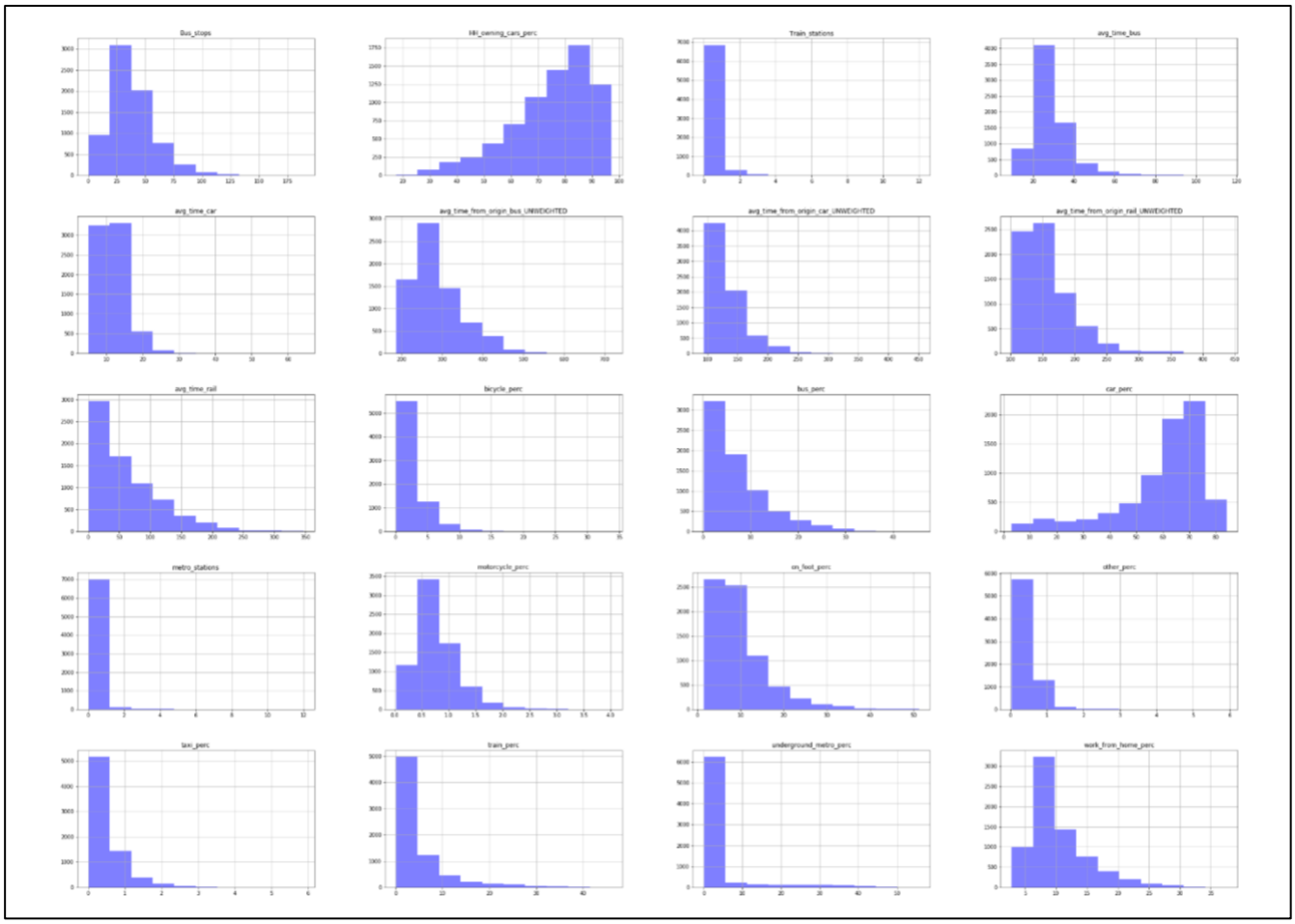

Figure 4: Original variable distributions.

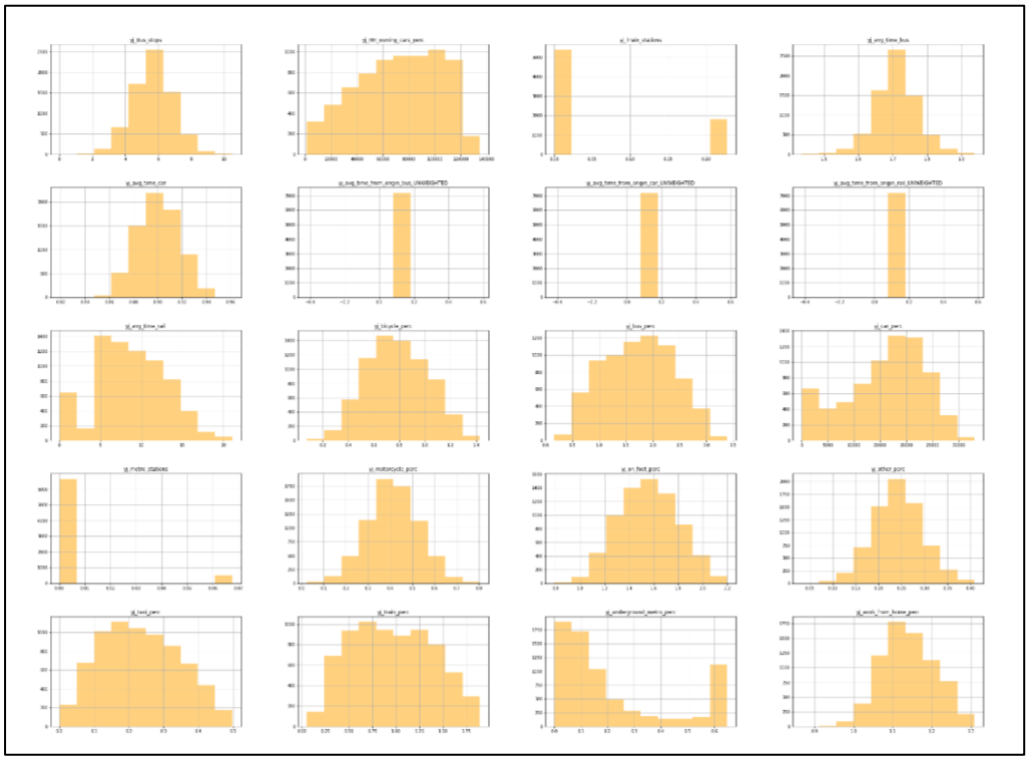

Figure 5: Variable distributions after Yeo-Johnson transformation.

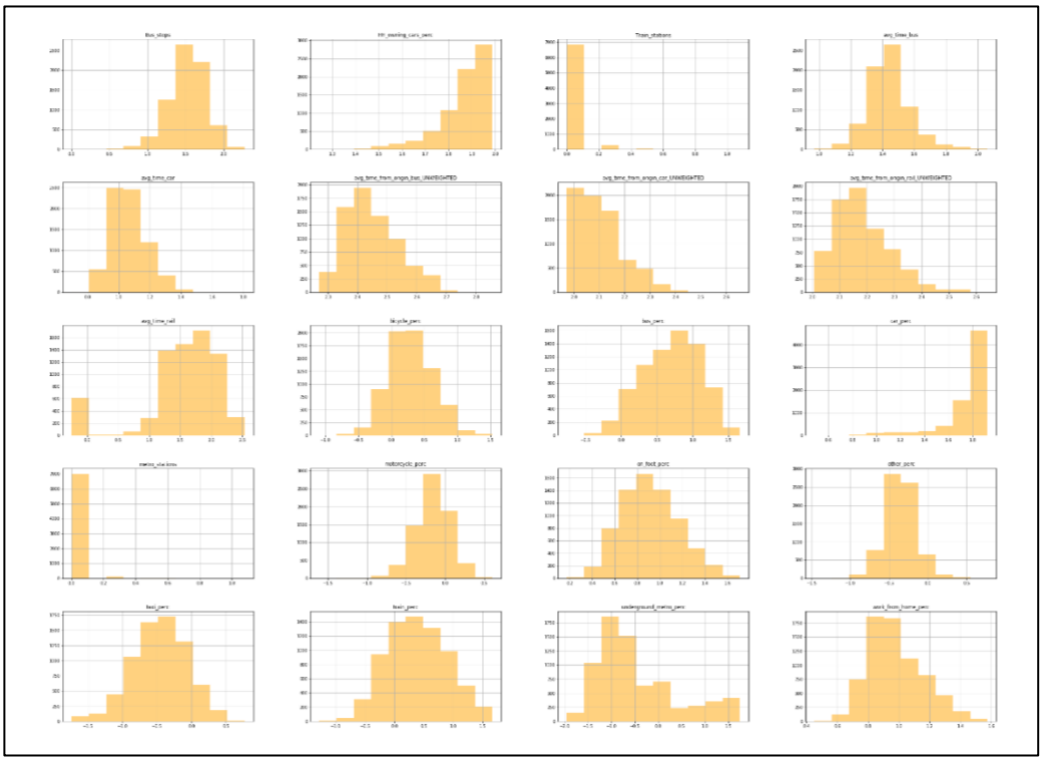

Figure 6: Variable distribution after log transformation.

Standardization

There were differences in units and ranges for each of the variables because it was not possible for all values to be changed to percentages (e.g. travel time). This is an issue with clustering since variables with the largest size or variability have the biggest influence on the clustering algorithm, especially for k-means (Bin Mohamad and Usman, 2013). To avoid this issue, data standardization is carried out to adjust the relative weight of the variables (Milligan and Cooper, 1988).

The standardization technique used 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. 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 how many standard deviations each value is from the mean.

Range:

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

$x = raw\ score$
$x_{min} = smallest\ value\ in\ variable\ distribution$
$x_{max} = largest\ value\ in\ variable\ distribution$

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, producing slightly different results. Both *k-means* and *hierarchical clustering* require the number of clusters to be specified. Therefore, the optimal number of clusters was found using elbow plots (figure 7) 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.

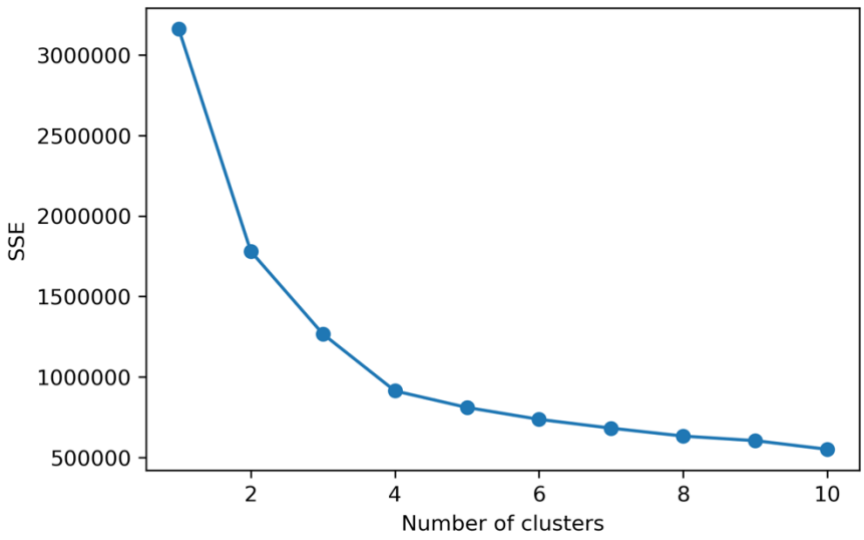

Figure 7: Elbow plot for log-idr-kmeans indicating that 4 clusters should be used.

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. As elbow plots and silhouette scores measure different things, they 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 used 2 transformation techniques, 3 standardization techniques and 3 clustering techniques, resulting in 18 combinations of results. We analysed 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 created multiple clusters but assigned most MSOAs to one cluster and identified the rest as outliers (see figure 8, left). This can be attributed to the 'curse of dimensionality', where the distance between different pairs of points decreases as the number of dimensions increases, as well as to differences in the density of points (Steinbach, Ertöz, and Kumar, 2004).

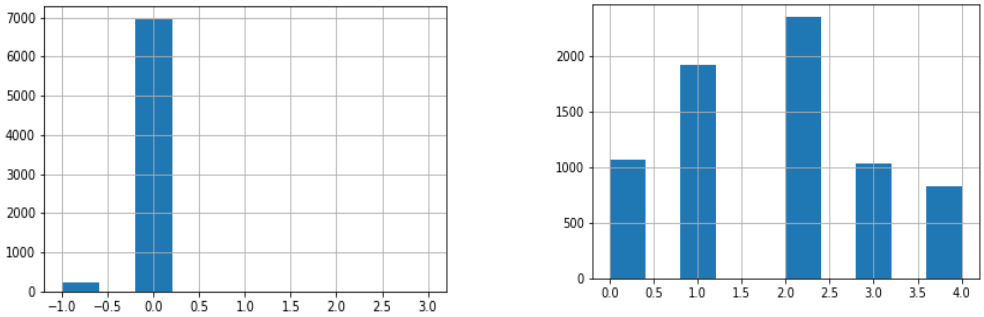

Figure 8: Histograms showing Number of MSOAs assigned to each cluster. LEFT (log_range_DBSCAN) &) RIGHT (log_range_kmeans).

K-means and Hierarchal clustering algorithms on the other hand provided clearer and more consistent results, which were compared by looking at the variable distribution and maps of the clustering results.

Variable Selection

To improve interpretability and clarity of results, variables that did not influence the results were removed. This included variables that were highly correlated (i.e. motorcycle percentage). 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 were further subsequently removed as even though they appear intuitively to be of importance they could not be clearly interpreted and did not influence the clustering results. This could have potentially been due to the difference in geographic size of MSOAs or to their relation to accessibility measures. Consequently, we ended up with 14 variables in the final clustering result

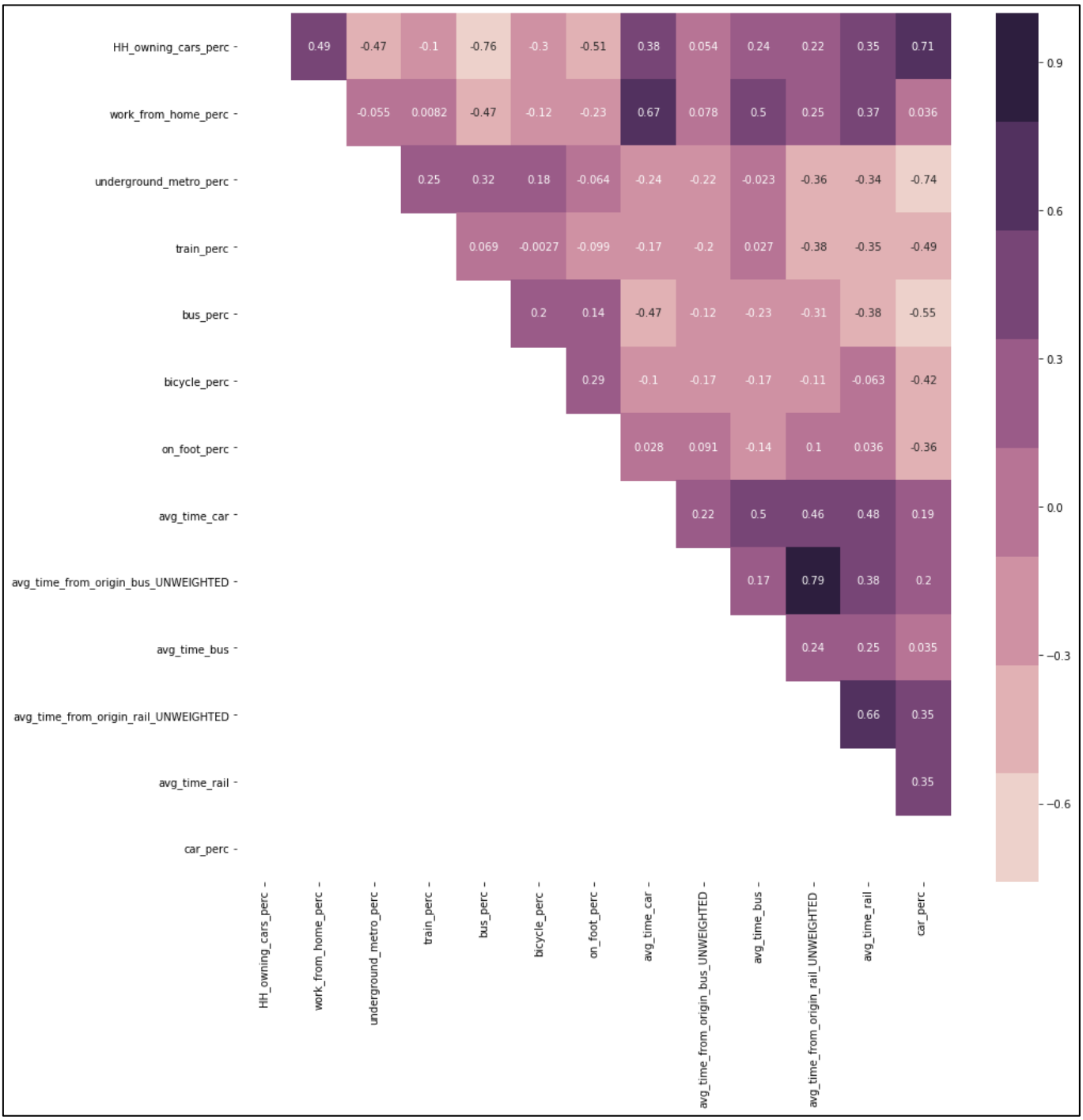

Figure 9: Correlation of variables used to produce clustering results.

Comparing Results

Once 14 variables were chosen, the steps above were repeated. The result we decided that best represented differentiation in clusters and mapped onto our knowledge of existing geography in the UK was the:

The resulting output was 5 clusters with distinct combinations of variable characteristics, as can be seen from the variable averages in each cluster (see figure 10) with descriptions of each cluster given in table 3.

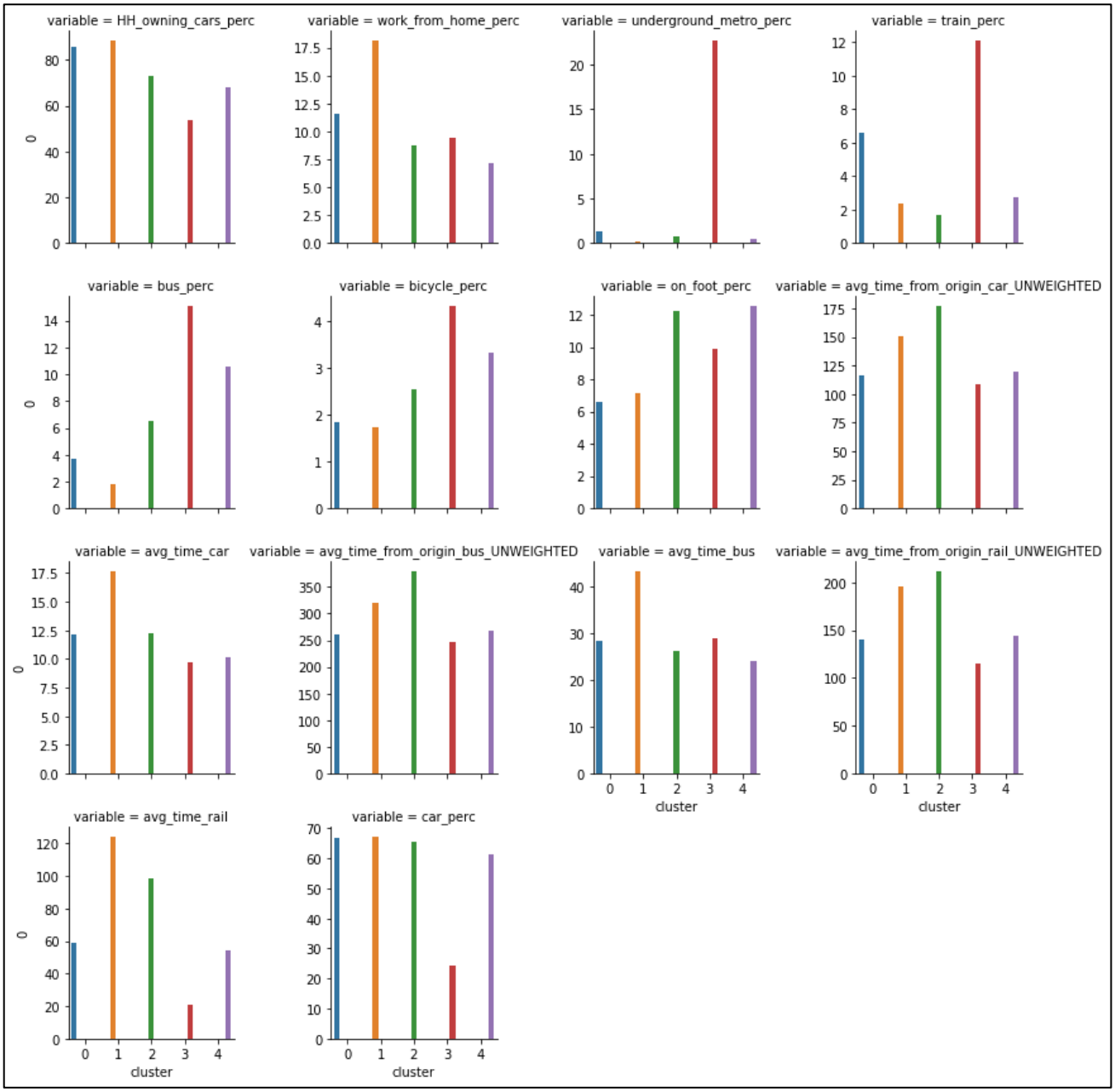

Figure 10: Variable averages in each cluster.

Table 3: Cluster descriptions.

Cluster	***Plot Colour***	***Extended Description***
1	Blue	**Good train accessibility but car dependant:** This cluster is composed of rural areas that surround land-locked urban areas. The MSOAs in this cluster are mainly in the center of England and Wales, compared to the rural areas in profile 2, which are on the outskirts. This cluster has the second best accessibility scores for all measured transport modes due to the central locations. The cluster benefits from being on train routes and has the second highest train usage, but that is the only mode of public transport that the MSOAs in this cluster are serviced by. As a result, the cluster is associated with high car ownership and usage, followed by train and walking.
2	Yellow	**Solely car dependant:** This cluster is made up of rural areas far from the cities. The MSOAs have few 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 and other connections between them and other parts of the country. The cluster is found on the periphery of profile 3, which is itself made up of coastal urban areas with poor accessibility.
3	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 the MSOAs are at the end of train lines and other transport networks and therefore lack external connectivity.
4	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. 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.
5	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 difference between these and London is the degree of usage of public transport with the main difference occurring due to the lack of usage of an underground/metro/tram.

Classification

Our aim was to not only to create different transport profiles but to understand the demographic factors related to these results, therefore the next step was to run a classification analysis. For example, 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 (see table 4) were chosen to understand how they are related to the transport profile groups.

Table 4: Variables used in classification.

Dataset	***Description***	***Source***
Net annual income (£)	Average net annual income in 2018	Office for National Statistics (Office for National Statistics, 2020)
Pop_Per_Hectare	Population density	Office for National Statistics (Office for National Statistics, 2019)
percent_unemployed	-	2011 Census (Office for National Statistics, 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	-	

A Random Forest classification with 100 trees, and max depth set to 3 to aid interpretability of results, was ran. 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.

Results

The results produce an overall score as result of many trees.

Table 5. Random Forest results.

cluster	precision	recall	f1-score
1	0.62	0.81	0.70
2	0.82	0.55	0.66
3	0.52	0.08	0.15
4	0.76	0.77	0.76
5	0.63	0.80	0.70
accuracy			0.65
macro avg	0.67	0.60	0.60
weighted avg	0.65	0.65	0.62

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 5. The recall score shows that the model was unable to give the correct value for cluster 3, meaning that most MSOAs in cluster 3 were misclassified. The confusion matrix (see figure 11) shows that only 24 out of 284 MSOAs in cluster 1 were correctly classified.

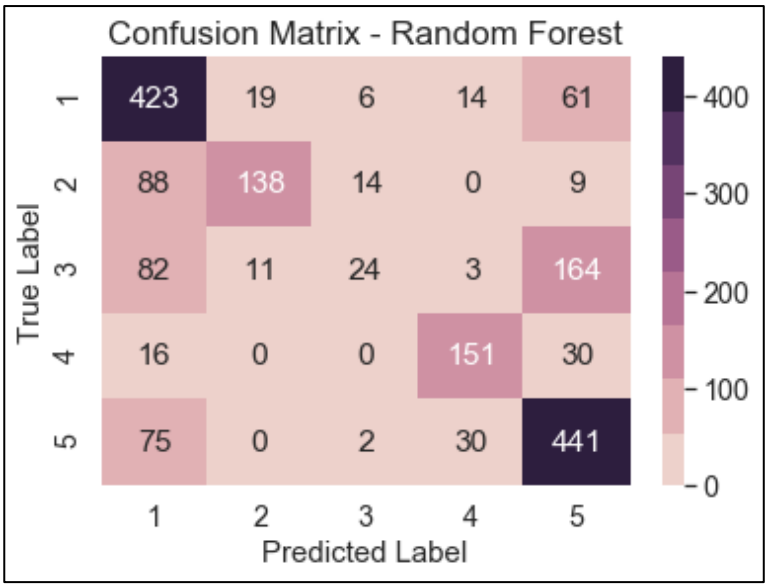

Figure 11: Confusion Matrix for Random Forest Classification.

Feature Importance

To understand which variables were most related to transport characteristics, we use feature importance. The default feature importance, 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). Therefore, permutation importance was used (Altmann et al., 2010), as it is less biased in its interpretation of feature importance.

The importance is based on calculating 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.

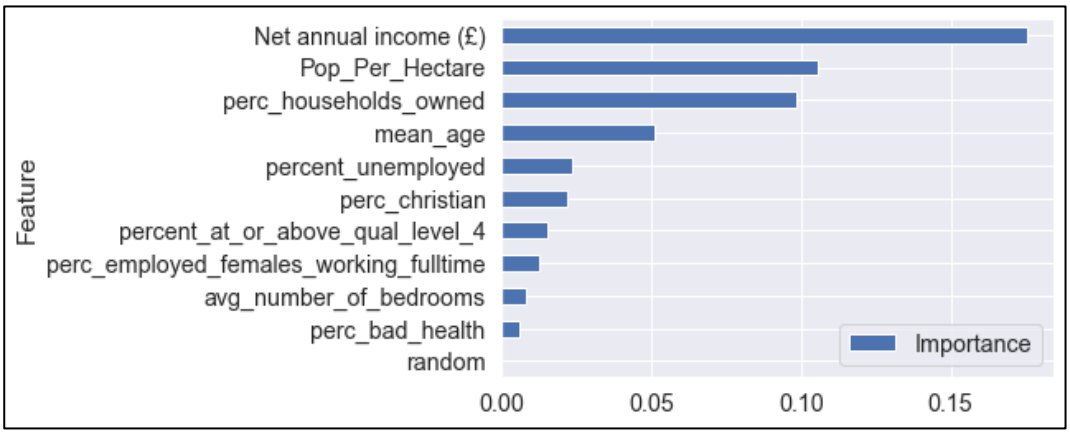

Figure 12: Permutation Importance for variables used in classification

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

Splitting

We can attempt to understand how the model has been trained by looking at individual trees (see figure 13). Here, cluster 4 '***high public transport and good accessibility***' is suggested to have an income above £33,850 and the percentage of households owned is greater than 59.25%. While in contrast, cluster 1 '***good train accessibility but still car dependent'*** is associated with an income below £33,850 and a mean age less than 38.25. Furthermore, looking at cluster 2 and 5, although they are both associated with low income, and a high mean age, cluster 5 is suggested to have a higher population density than cluster 2. From these we can begin to understand how factors may influence transport profile, although it must be acknowledged that the tree shown below is based on a subsample of all data collected (3,449 MSOAs). Therefore, this could be advanced if we had more time by exploring more of the resulting decisions trees and gaining a better understanding of the factors that may influence transport behaviour in the UK.

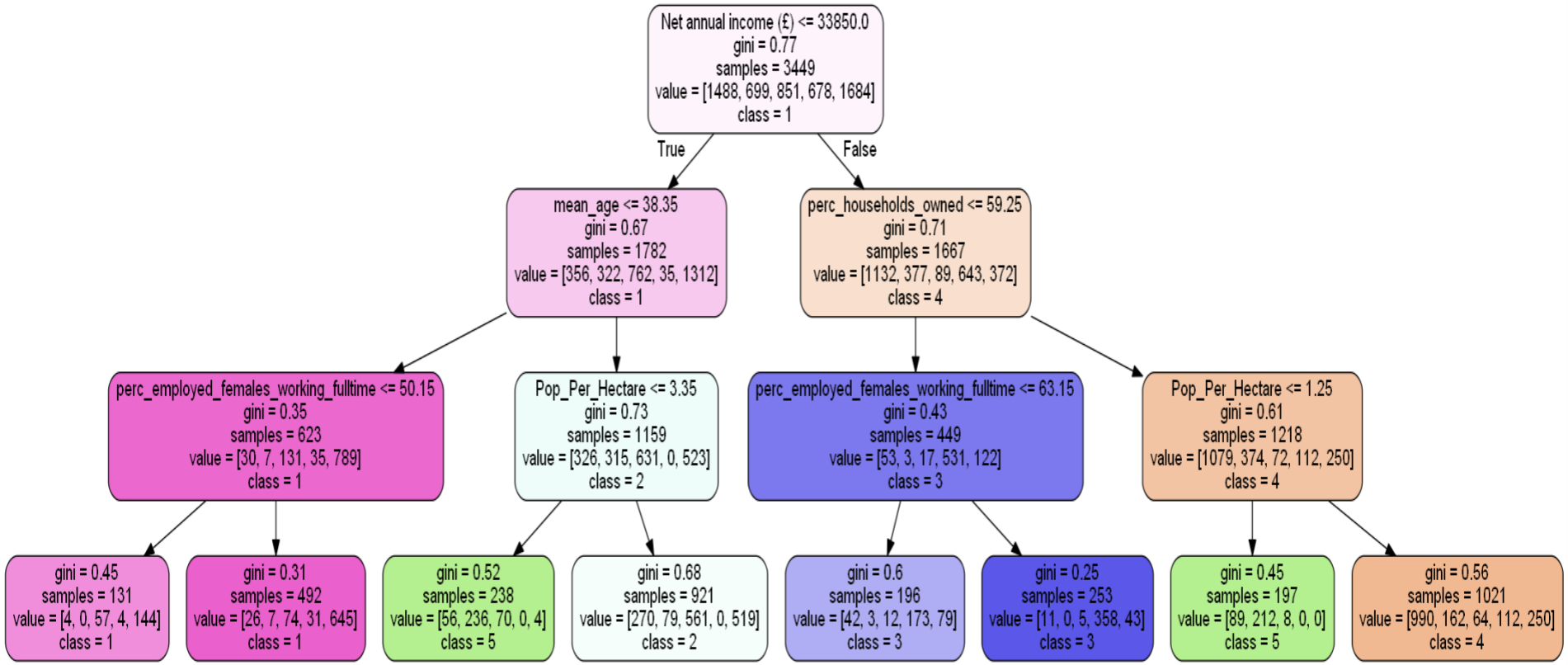

Figure 13: Sample decision tree from random forest.

Further Work

Given additional time and resources, our analysis could be extended if we were to create a hierarchy of clusters, as done by (Jahanshahi and Jin, 2020). This would split up the existing clusters into sub-clusters, giving us a better understanding of variations in areas such as London. Furthermore, we would also seek to extend our classification analysis. This would be achieved by exploring the random forest algorithm further and utilising different variables and methods of classification to improve predictions.

Web Visualisation

Introduction

The above analysis is presented as a website, which includes a story page that guides the reader through the analysis process and findings, and a map page where the user can further interact with the data and the results. As one of our objectives was to inspire a dialogue about transport mode accessibility and choice in England and Wales, the website is targeted towards a general audience with an interest in the subject.

The following sections will discuss the development process, explain how our choice of audience informed decisions regarding the content and design of the website, and provide a detailed explanation of how the website functions. We also suggest some further work that could be done to improve the website.

Development Process and Agile Principles

Our website development process utilized agile methodologies to ensure that all team members had a clear understanding of our mission (Williams, 2010). While some tenets of the methodology were not possible, such as 'collocation' (Dingsøyr *et al.*, 2012, p. 1214), other techniques were critical to establishing a productive development process. Two agile principles were particularly important: 'individuals and interactions over processes and tools' and 'responding to change over following a plan' (Williams, 2010, pp. 4-5). The data analysis process and the web development process were undertaken simultaneously, so it was important that we were able to foster adaptability through sustained cooperation between all team members.

In addition to using agile principles, our team used collaborative tools such as GitHub, Trello, Miro, and Pinterest to manage tasks, create wireframes, and plan the appearance of the site. Each of these platforms helped to establish a shared vision of the final outcome, which was iteratively revised as needed. A segment of an early-stage wireframe can be seen in figure 14.

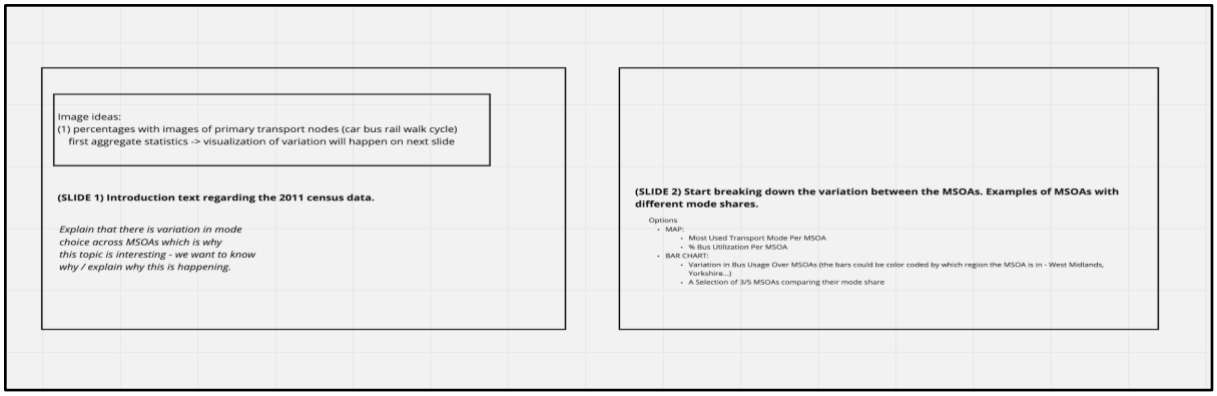

Figure 14: Segment of early-stage Miro wireframe for story page.

Inspiration from Existing Web Visualisations

In order to gain an understanding of the techniques and conventions used by existing web visualisations, we looked at several examples, mostly pertaining to commuting patterns or migration. The methods used in these websites inspired choices that were made for our site, allowing it to fit cohesively with similar precedents while also introducing new features.

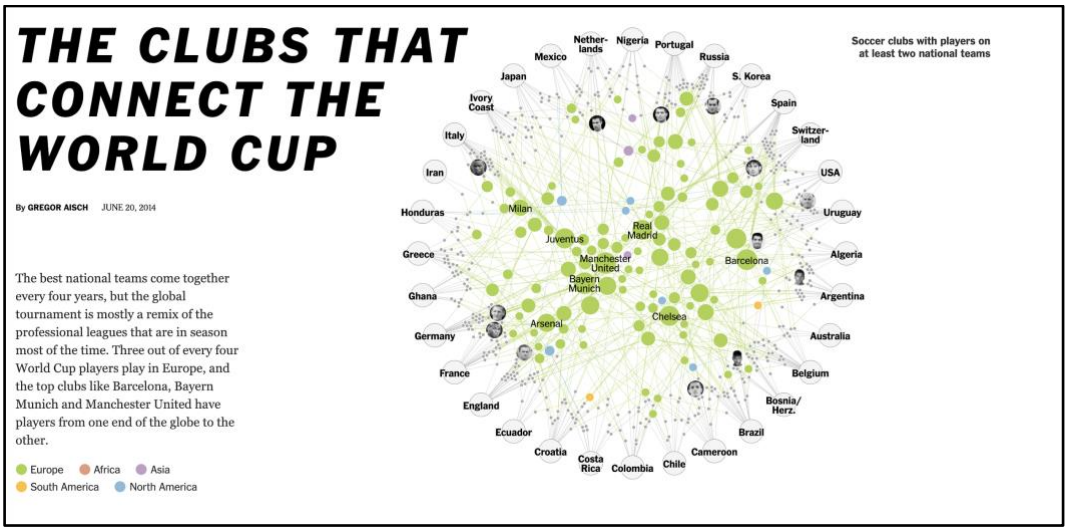

Figure 15: The Clubs That Connect The World Cup (Aisch 2014).

The New York Times (see figure 15) uses 'scrollytelling' techniques to produce compelling, data-driven articles. In doing so, the writers use clear language which summarises the data graphics. We wanted to achieve this level of engagement with the reader. Therefore, we used the same scrolling library for our story page and ensured our language was understandable and concise.

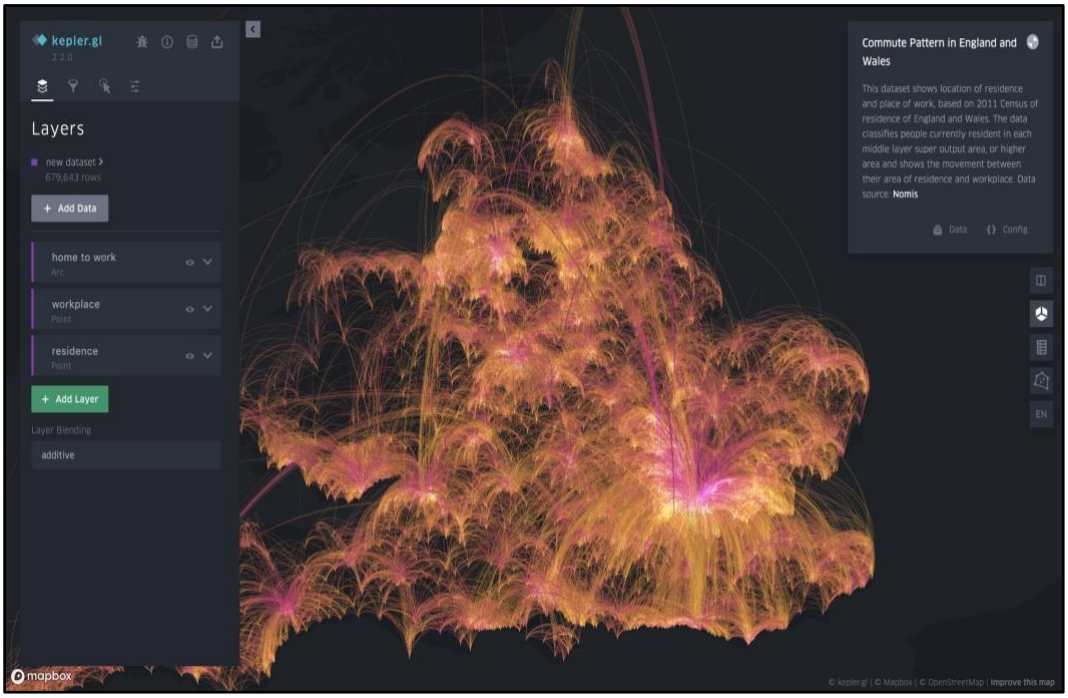

Figure 16: Commute Patterns in England and Wales (Uber 2018).

Commute Patterns in England and Wales (see figure 16) implements kepler.gl to visualize commuting flows as arcs. Our original intention was to display flows as arcs using deck.gl, which was attempted, but time restrictions meant that it was necessary to use another technique. We were also inspired by the 'dashboard' layout of this website, which informed the arrangement of our interactive map page.

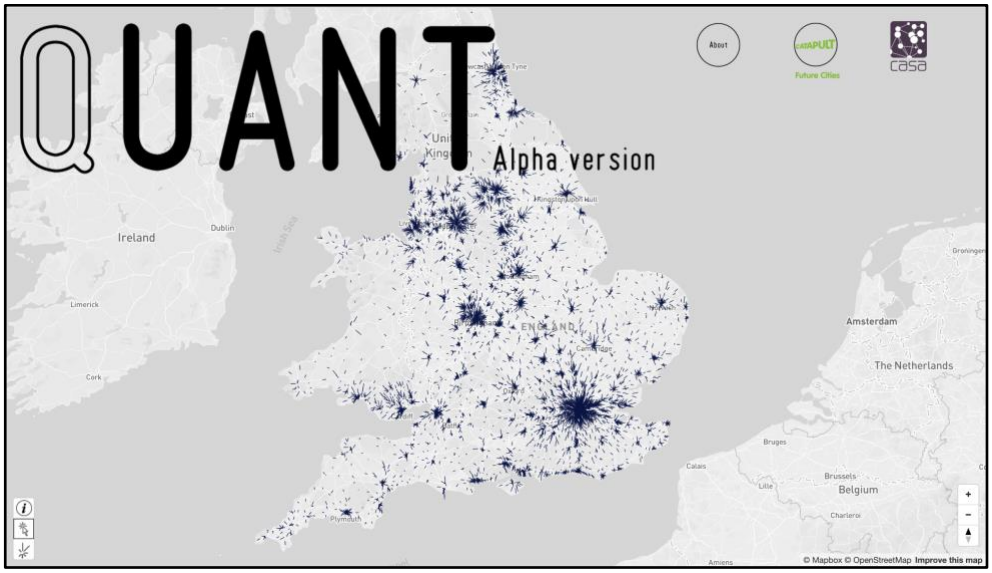

Figure 17: QUANT: Simulating the Impacts of Large Scale Change in England and Wales (Future Cities Catapult and Centre for Advanced Spatial Analysis).

The QUANT website (see figure 17) allows the user to adjust the spatial interaction model to see how commuting flows would react to large-scale changes. Our website does not allow the user to interact with our algorithms, although that is a possible area for further work. QUANT uses both choropleth maps and arrows representing flows, but only on separate maps on the site. Our website took inspiration from this and combined the choropleth and flow elements into one map.

Figure 18. World Migration Map (Galka, 2016).

Our main takeaway from the *World Migration Map* (see figure 18) was that there are many creative and dynamic ways to visualize flows. This website uses moving dots, which better represents movement than lines, arrows, or even arcs. While this method was not implemented, future iterations of the website could explore such techniques.

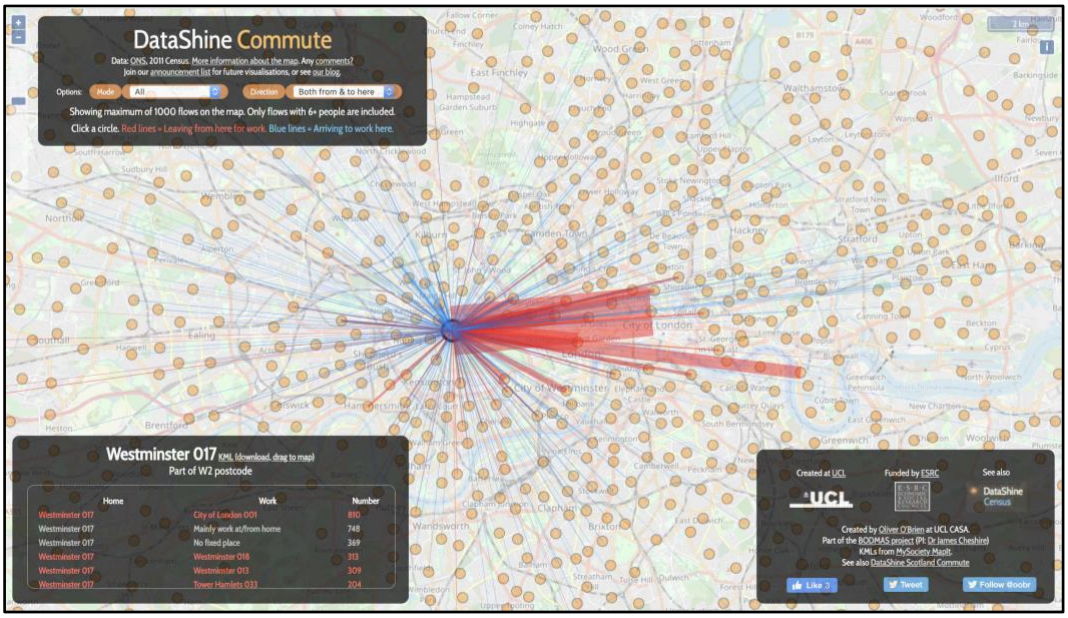

Figure 19: DataShine Commute (O'Brien, 2017).

The last example, *DataShine Commute* (see figure 19) uses lines of varying thicknesses to indicate flow volume and colour to indicate net loss or gain. This visualisation method is the most similar to our final product, although there are differences in how the data is broken down, presented, and combined with other findings.

Most of the examples we looked at, and all of those described above, were single page sites. We elected to include a homepage, separate story and interactive map pages, and an about page. This choice makes our website more of a stand-alone piece of work than the other sites we looked at. Further, the incorporation of complex cluster analysis sets our site apart from those that present the same data.

User Experience and User Interface Design

The website was designed to share the results of our analysis with a general audience. To provide a meaningful, valuable, and pleasant experience, we applied user-centered design, which is oriented towards understanding the users, their needs and their interactions (Norman, 1988). Both user experience (UX) and user interface (UI) design played a vital role in shaping our website.

The goal of UI design is to visually guide the user through a website's interface while creating an intuitive, enjoyable experience (Galitz, 2007). As a website's quality is often measured by user experience and usability (Sivaji and Tzuaan, 2012), we incorporated

Nielsen's 'Ten Usability Heuristics for User Interface Design' (1994b). Although Nielsen released his heuristics in 1994, they are still the most widely used principles for user interface design (Liyanage, 2016; Jain, 2015).

Nielsen's second heuristic suggests that there should be a 'match between the system and the real world' (1994b). We primarily implemented this heuristic in choosing the language for the site, ensuring that the content would be understandable and not overly technical for a layperson. This is evident across the story page and in phrases on the interactive map such as 'view where people work' rather than something like 'view commuting flows'.

Nielsen's third heuristic, 'user control and freedom' (1994b) is incorporated throughout our website's UX. User centric design can be seen on our homepage, where the user is in control of what page they navigate to first. Further, the burger menu, visible throughout the website's pages, allows the user to freely move between pages at any point.

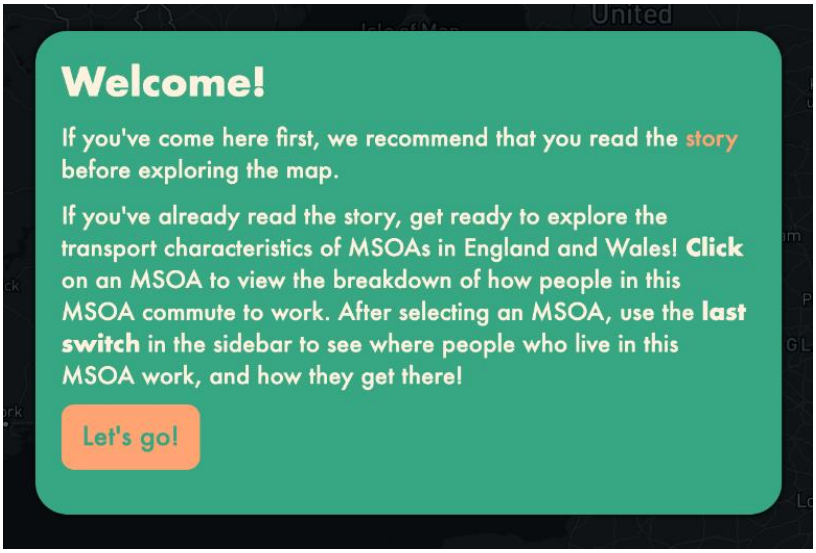

Figure 20: Transitional point on the map page; pop up box that recommends reading the story page first.

However, a continuous journey is suggested (see figure 21) through the use of prompts at transitional points (see figure 20). Another example is on the story page, where the user can choose to scroll or click the 'next page' button to move through the content.

Additionally, Nielsen notes that the user will often make mistakes, and will need simple ways to undo or redo their actions (1994b). An example of how we included such functionality can be seen on the map page, where the user can turn the cluster switches on and off.

In considering Nielsen's fourth heuristic, 'consistency and standards' (1994b), the website was designed to be intuitive and coherent. For example, we used a single colour palette, a consistent placement of the burger menu, and the same hover styling of links.

In line with Nielsen's fifth heuristic, 'error prevention', and ninth heuristic, 'help users recognize, diagnose, and recover from errors' (1994b), we've ensured the user is always informed if anything goes wrong with a familiar system prompt that precisely indicates the problem (see figure 22).

Figure 21. Recommended user's journey.

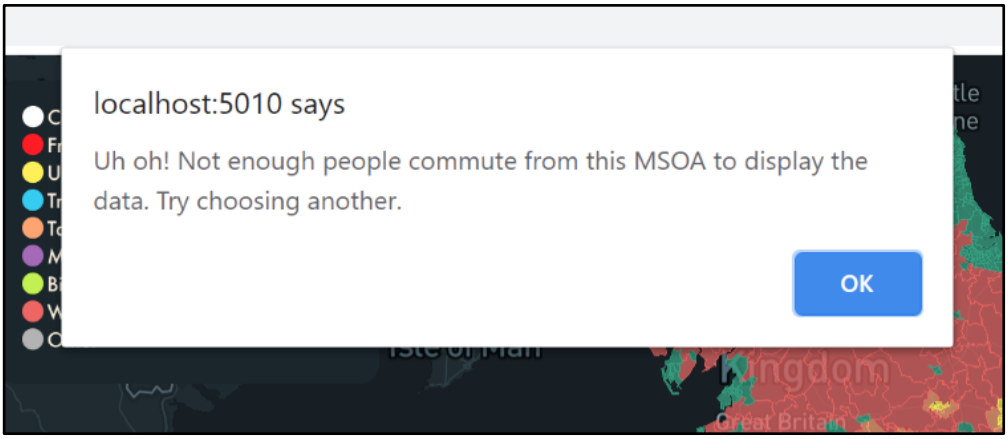

Figure 22: Pop up box informing the user the action can't be completed.

We deployed the eighth heuristic, 'aesthetic and minimalist design' (1994b), by only including necessary information to avoid overwhelming the reader. The website aesthetic was aimed at engaging the reader through the use of bright colours and dynamic icons. Icons were created to represent different transport modes, based upon work by RADIO, Lee, norbertsobolewski, adekvat, Young, and Marchukov. They were made using Adobe Illustrator, converted to SVGs, and combined with JavaScript and CSS animation functions. On the homepage, these illustrations were put together to create a dynamic cityscape and applied throughout the website to offer a sense of cohesion.

As per Neilsen's tenth heuristic, 'help and documentation' (1994b), informational pop ups on the interactive map ensure that the user has access to additional information (see figure 23).

Figure 23: Clickable information pop ups on the map page.

The seventh principle, 'flexibility and efficiency of use' (Nielsen 1994b), could be furthered by enabling the user to save their preferences for the map view, or by preserving frequently selected MSOAs or address searches. These are potential areas for future work.

Technical Integration Between Web Elements

This section is split in two parts. The first discusses the symbiotic relationship between technologies, programming languages, and external libraries that enable our website to function. The second reviews the continuous integration and continuous development (CICD) pipeline.

Programming Languages

The website's aesthetic and structure are developed using HTML5 and CSS3. These are two programming languages which create the foundation for the wide variety of web applications on the internet (Mavrody, 2010). Keeping to web development tradition (Flanagan, 2017), JavaScript is used throughout the website to visualize data from the analysis and to create elements of interactivity. For the most part, all internal CSS and JavaScript is stored in separate files to the HTML and is called in the header, body or footer where necessary.

External Libraries

Both the map and story page required external libraries to visualise data and create a smooth user experience. The only external library unique to the map page was Mapbox. Mapbox's range of web development APIs allowed the geovisualisation of multiple datasets (Cadenas, 2014). Our interactive map utilised Mapbox's vector tiles API to display the clustering and flows, Mapbox studio to create a base map in our colour palette, and the geocoder API to provide a reliable search bar.

D3, a powerful JavaScript library, makes use of scalable vector graphics to create dynamic and interactive data visualisations (Bostock *et al.*, 2011). D3 was simultaneously used on the map and story page. On the map page, D3 version 4 produced the bar chart in the user dashboard. On the story page, D3 version 3 created interactive pie charts in the first section of the story. Meanwhile, D3 version 4 generated boxplots, bar charts and circular bar plots in the second section of the story.

JQuery was also used on the map and story page. JQuery is an open source library which handles the complicated elements of data-driven web development, such as ajax, animation and event handling (Nixon, 2014). Within the map page, JQuery queued the welcome pop up and displayed the map asynchronously. On the story page, JQuery was used in tandem with another external library based on JQuery event handling called ScrollMagic. Embedded in the project repository, ScrollMagic is a JavaScript library which creates 'magical scroll interactions' (Paepke, 2020). ScrollMagic produced a 'scrollytelling' effect by applying a blur transition between sections of the story (Aisch and Almukhtar, 2015; Aish, 2014).

ChartJS, an open source JavaScript library was used to graphically represent interactive bar plots on the story page. Although the result was also achievable using D3, ChartJS was used to display diversity and ability within the web development team.

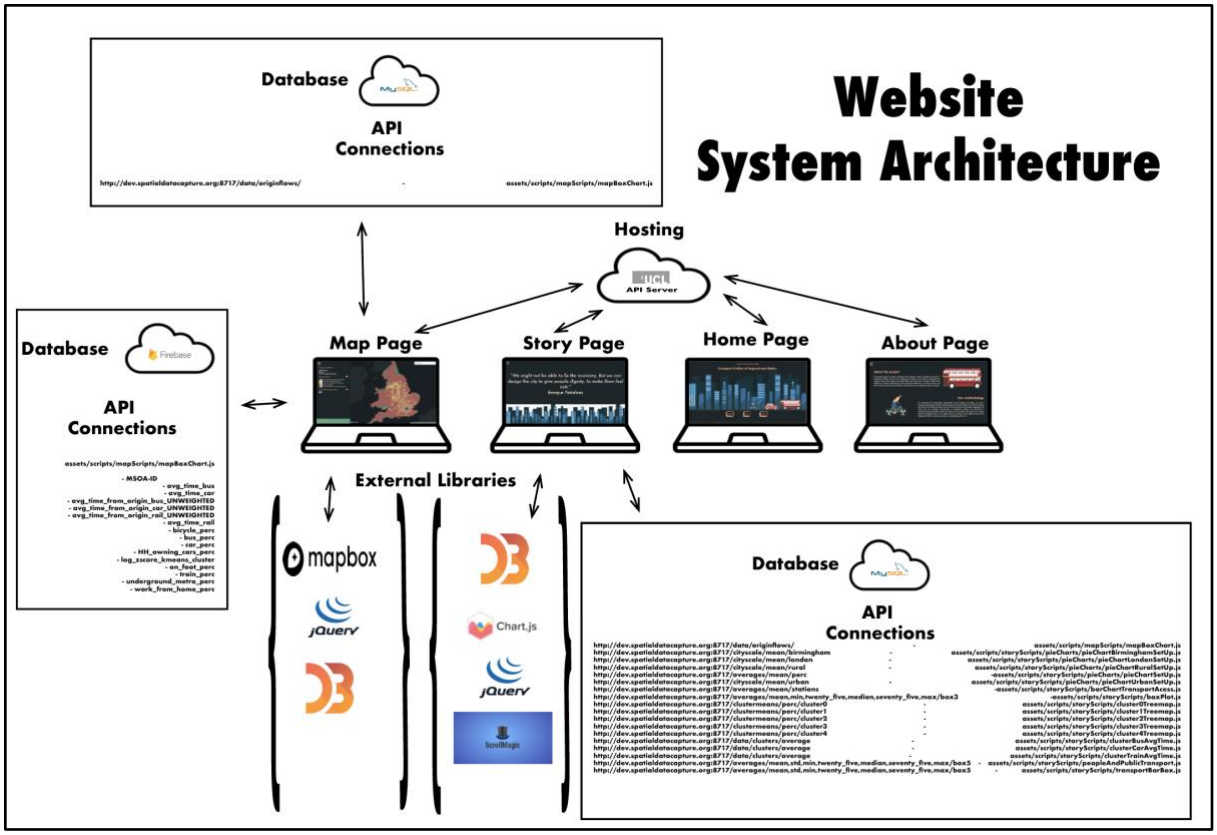

Figure 24: Website system architecture.

System Architecture

The website takes form with a classic system architecture approach (Sikos, 2014). The website repository is hosted on a HTTP server provided by UCL; the website was originally hosted using Firebase Hosting, but the combination of the HTTP server from UCL and the HTTPS server from Firebase caused SSL conflicts. The UCL server also hosts a SQL database containing the results from the analysis. A REST API was developed using this database, allowing the website to retrieve JSON data for visualisation purposes. Another database, hosted in a data centre in the United States, was configured using Firebase. Firebase is a Google platform providing a range of web and app functionality (Firebase, 2020). The Firebase database contains MSOA data from the analysis, which is visualised on the map page. Figure 24 presents a detailed overview of how external libraries, databases and hosting are configured in our project.

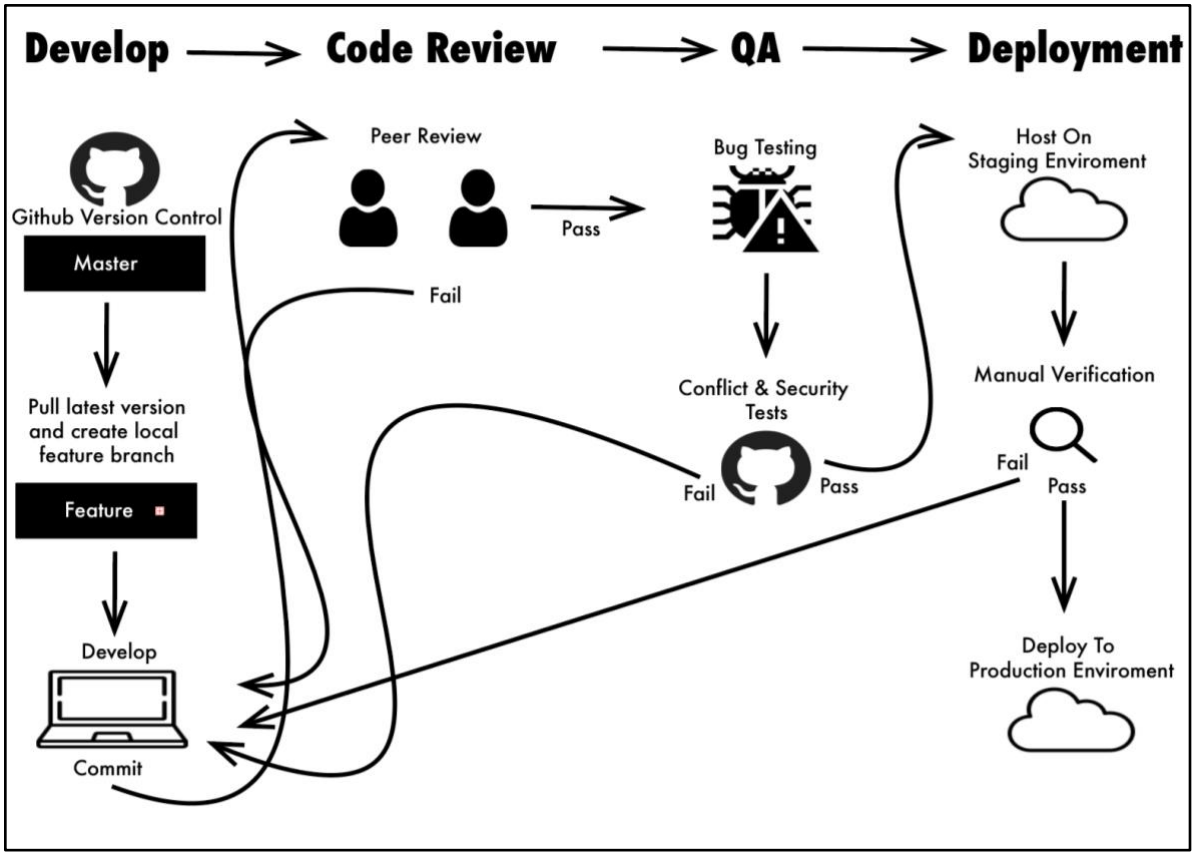

Figure 25: CICD pipeline.

Continuous Integration and Continuous Development (CICD)

To foster a smooth and organised development process among three web developers in three different locations, a robust CICD pipeline was constructed. Adhering to agile software development processes (Germain and Robillard, 2005), the pipeline details the full development process from start to finish (see figure 25). By adhering to this pipeline, merge conflicts, bugs and errors were mitigated.

Conclusion

Our website aims to offer an interactive platform where users can learn and engage with the analysis we conducted. As transportation and commuting are ubiquitous in nearly everyone's lives, we wanted the website to be approachable to a variety of end users, without sacrificing accuracy and depth. To do this, we implemented specific UX, UI, and data visualisation techniques.

Recommendations for Future Work

Our team aimed to be quite thorough in the development of the website. Nonetheless, we identified several areas of improvement that were not undertaken in the current site due to time constraints, but should be addressed in future versions:

- The website was developed for a desktop experience. While the story page is responsive, the entire site should incorporate mobile responsiveness using CSS media queries so the content can be viewed on devices of different sizes.
- The website should be thoroughly reviewed for accessibility, ensuring that the colour palette is suited for colour blind people and the pages can be explored using a keyboard (Firth, 2019).
- We could also expand on the level of interactivity offered to the user. For example, the interactive map page could allow the user to toggle which flows they would like to see based on transport mode. Additionally, we could allow the user to interact with the analysis process as is done on the QUANT website. For example, the user could choose between clustering algorithms or socio-demographic variables for the classification analysis.
- As previously mentioned, our original intention was to utilize deck.gl to visualize the flow data on the interactive map. As deck.gl is a novel and complicated framework, we were not able to do this. A future version should continue to explore this option.
- Our website does not store or record sensitive data. However, in accordance with best practice, the website and data should be moved onto a secure server, as opposed to the current HTTP configuration that UCL provides (Stuttard and Pinto, 2011, p. 51).
- The UX of our website could be improved by adding a spinner when the data on the map page is loading, in line with Nielsen's first heuristic 'visibility of system status' (1994b).
- Our website was built upon the best UX practices, but tests with real users should be conducted (Bruun et al., 2009; Reeves, 2019). By performing a usability assessment, we could identify more areas of improvement that we may not have considered.

Bibliography

adekvat, 2015. Modern and Vintage Trains. Vector. *Depositphotos*, New York. Accessed 21 May 2020. <https://depositphotos.com/90320756/stock-illustration-modern-and-vintage-trains.html>.

Adnan, M., Longley, P., Singleton, A., and Brundson, C., 2010. Towards Real-Time Geodemographics: Clustering Algorithm PErformance for Large Multidimensional Spatial Databases. *Transactions in GIS,* 14(3), pp. 283-297.

Aisch, G., 2014. The Clubs That Connect The World Cup. *The New York Times*. <https://www.nytimes.com/interactive/2014/06/20/sports/worldcup/how-world-cup-players-are-connected.html?mtrref=www.google.com&gwh=BA5F2A697EA5ABA1582C6C6D2640233D&gwt=pay&assetType=REGIWALL>.

Aisch, G., and Almukhtar, S., 2015. Seeking a Fair Distribution of Migrants in Europe. *The New York Times*. <https://www.nytimes.com/interactive/2015/09/04/world/europe/europe-refugee-distribution.html?mtrref=www.google.com&gwh=3BDD0ED2A5ABC2E2E519A623E7705CC8&gwt=pay&assetType=REGIWALL>.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T., 2010. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics*, 26, pp. 1340–1347.

Batty, M., and Milton, R., 2019. QUANT: A National Small Area Land Use Transport Model to Evaluate the Impact of Large Rail Infrastructure Projects.

Bin Mohamad, I., and Usman, D., 2013. Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6.

Bostock, M., Ogievetsky, V., and Heer, J., 2011. D^3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17, pp. 2301–2309. <https://doi.org/10.1109/TVCG.2011.185>.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp. 5–32.

Bruun, A., Gull, P., Hofmeister, L., and Stage, J., 2009. Let Your Users Do the Testing: A Comparison of Three Remote Asynchronous Usability Testing Methods. *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1619-1628.

Buehler, R., 2011. Determinants of Transport Mode Choice: A Comparison of Germany and the USA. *Journal of Transport Geography,* 19, pp. 644-657.

Cadenas, C., 2014. Geovisualization: Integration and Visualization of Multiple Datasets Using Mapbox. *Cal Poly.* Accessed 24 May 2020. <https://digitalcommons.calpoly.edu/cpesp/137/>.

Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S.H., and Zhang, M.Q., 2002. Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data. *Statistica Sinica*, 12, pp. 241–262.

Chng, S., White, M., Abraham, C., and Skippon, S., 2016. Commuting and Wellbeing in London: The Roles of Commute Mode and Local Public Transport Connectivity. *Preventive Medicine,* 88, pp. 182-188.

Collins, C., and Chambers, S., 2005. Psychological and Situational Influences on Commuter-Transport-Mode Choice. *Environment and Behaviour,* 37, pp. 640-661.

Delbosc, A., 2012. The Role of Well-Being in Transport Policy. *Transport Policy,* 23, pp. 25-33.

Dingsøyr, T., Nerur, S., Balijepally, V., and Moe, N.B., 2012. A Decade of Agile Methodologies: Towards Explaining Agile Software Development. *Journal of Systems and Software*, 85(6), pp. 1213-1221.

Ferdous, N., Pendyala, R., Bhat, C., and Konduri, K., 2011. Modelling the Influence of Family, Social Context, and Spatial Proximity on Use of Nonmotorized Transport Modes. *Journal of the Transport Research Board,* 2230(1), pp. 111-120.

Firebase, n.d. *Firebase.* Accessed 23 May 2020. <https://firebase.google.com/>.

Firth, A., 2019. *Practical Web Inclusion and Accessibility*. Apress, London.

Flanagan, D., 2006. *JavaScript: The Definitive Guide*, 5th ed. O'Reilly Media, Inc., Sebastopol.

Galitz, W.O., 2007. *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques*. John Wiley & Sons, Indianapolis.

Galka, M., 2016. World Migration Map. *Metrocosm*. Accessed 21 May 2020. <http://metrocosm.com/global-migration-map.html>.

Germain, É., and Robillard, P.N., 2005. Engineering-Based Processes and Agile Methodologies for Software Development: A Comparative Case Study. *Journal of Systems and Software*, 75(1-2), pp. 17–27.

Harris, R., Johnston, R., and Burgess, S., 2007. Neighborhoods, Ethnicity and School Choice: Developing a Statistical Framework for Geodemographic Analysis. *Population Research Policy Review*, 26, pp. 553-579.

Jahanshahi, K., and Jin, Y., 2020. Identification and Mapping of Spatial Variations in Travel Choices through Combining Structural Equation Modelling and Latent Class Analysis: Findings for Great Britain. *Transportation*. <https://doi.org/10.1007/s11116-020-10098-9>

Jain, A., 2015. 10 Heuristic Principles – Jakob Nielsen's (Usability Heuristics). *UXness*. Accessed 24 May 2020. <http://www.uxness.in/2015/02/10-heuristic-principles-jakob-nielsens.html>.

Kumar, Ch.N.S., Rao, K.N., Govardhan, A., and Sandhya, N., 2015. Subset K-Means Approach for Handling Imbalanced-Distributed Data, in: Satapathy, S.C., Govardhan, A., Raju, K.S., and Mandal, J.K. (Eds.), Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 497–508. <https://doi.org/10.1007/978-3-319-13731-5_54>

Laverty, A., Mindell, J., Webb, E., and Millett, C., 2013. Active Travel to Work and Cardiovascular Risk Factors in the United Kingdom. *American Journal of Preventive Medicine*, 45(3), pp. 282-288.

Lee, A., 2017. Cycle Boiz. GIF. *Dribble*, Austin. Accessed 21 May 2020. <https://dribbble.com/shots/3397120-Cycle-Boiz>.

Liyanage, E., 2016. 10 Usability Heuristics Explained. *Medium.* Accessed 24 May 2020. <https://medium.com/@erangatl/10-usability-heuristics-explained-caa5903faba2>.

Marchukov, V., 2016. C'mon Ladies! GIF. *Dribble*, Austin. <https://dribbble.com/shots/2532208-C-mon-Ladies>.

Mavrody, S., 2010. *Sergey's HTML5 & CSS3 Quick Reference*, 1st ed. Belwin Mills.

Milligan, G.W., and Cooper, M.C., 1988. A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification*, 5, pp. 181–204.

Nielsen, J., 1994a. Enhancing the Explanatory Power of Usability Heuristics. *Proc. ACM CHI'94 Conf*, Boston, April 24-28, pp. 152-158.

Nielsen, J. 1994b. 10 Heuristics for User Interface Design. *Nielsen Norman Group.* Accessed 24 May 2020. <https://www.nngroup.com/articles/ten-usability-heuristics/>.

Nixon, R., 2014. *Learning PHP, MySQL & JavaScript: With JQuery, CSS & HTML5*, 4th ed. O'Reilly Media, Inc, Sebastopol.

norbertsobolewski. Cartoon Red Double Decker Bus. Vector. *123RF,* Malaysia. Accessed 21 May 2020. <https://www.123rf.com/stockphoto/double_decker_bus.html?sti=m0sv0smqdvcw8b3ngj|&mediapopup=92757553>.

Norman, D.A., 1988. *Designing Everyday Things*, Currency-Doubleday, New York.

O'Brien, O., 2017. DataShine Commute. *University College London.* Accessed 21 May 2020. <https://commute.datashine.org.uk/>.

Office for National Statistics, 2011a. Origin Destination - 2011 Census - Nomis - Official Labour Market Statistics. <https://www.nomisweb.co.uk/census/2011/bulk/rOD1>

Office for National Statistics, 2011b. Quick Statistics - Census 2011 - home - Nomis - Official Labour Market Statistics <https://www.nomisweb.co.uk/census/2011/quick_statistics> (accessed 5.7.20).

Office for National Statistics, 2013. Car or van availability - Nomis - Official Labour Market Statistics <https://www.nomisweb.co.uk/census/2011/qs416ew> (accessed 5.25.20).

Office for National Statistics, 2015. Methodology Note for the 2011 Area Classification for Output Areas.

Office for National Statistics, 2019. Middle Super Output Area population estimates (supporting information) - Office for National Statistics <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/middlesuperoutputareamidyearpopulationestimates> (accessed 5.7.20).

Office for National Statistics, 2020. Income estimates for small areas, England and Wales - Office for National Statistics <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales> (accessed 5.7.20).

Oshiro, T.M., Perez, P.S., and Baranauskas, J.A., 2012. How Many Trees in a Random Forest?, in: Perner, P. (Ed.), Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 154–168. <https://doi.org/10.1007/978-3-642-31537-4_13>.

Paepke, J., n.d. ScrollMagic. Accessed 23 May 2020. <http://scrollmagic.io/>.

Pinjari, A., Pendyala, R., Bhat, C., and Waddell, P., 2007. Modeling Residential Sorting Effects to Understand the Impact of the Built Environment on Commute Mode Choice. *Transportation,* 34, pp. 557-573.

QUANT: Simulating the Impacts of Large Scale Change in the UK. *Future Cities Catapult* and *Center for Advanced Spatial Analysis*. Accessed 21 May 2020. <http://quant.casa.ucl.ac.uk/>.

R A D I O, 2014. Cycling. GIF. *Dribble,* Austin. Accessed 21 May 2020. <https://dribbble.com/shots/1749434-Cycling?1412255632#shot-description>.

Reeves, S., 2019. How UX Practitioners Produce Findings in Usability Testing. *ACM Transactions on Computer-Human Interaction,* 26(3), pp. 1-38.

Shelton, N., Birkin, M., and Dorling, D., 2006. Where Not to Live: A Geo-Demographic Classification of Mortality for England and Wales, 1981-2000. *Health & Place,* 12, pp. 557-569.

Sikos, L., 2014. *Web Standards: Mastering HTML5, CSS3, and XML*, 2nd ed. Apress.

Sivaji, A., and Tzuaan, S.S., 2012. Website User Experience (UX) Testing Tool Development Using Open Source Software (OSS). *Southeast Asian Network of Ergonomics Societies Conference (SEANES),* Langkawi, Kedah, pp. 1-6.

Steinbach, M., Ertöz, L., and Kumar, V., 2004. The Challenges of Clustering High Dimensional Data, in: Wille, L.T. (Ed.), *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*. Springer, Berlin, Heidelberg, pp. 273–309. <https://doi.org/10.1007/978-3-662-08968-2_16>.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T., 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8, p. 25. <https://doi.org/10.1186/1471-2105-8-25>.

Stuttard, D., and Pinto, M., 2011. *The Web Application Hacker's Handbook Finding and Exploiting Security Flaws,* 2nd ed. Wiley, Indianapolis.

Titheridge, H., Potter, S., Enoch, M., and Bannister, D., 2008. *Using Geo-Demographic Analysis to Calculate Patronage Figures for Rural Buses. Final report*, London: UCL and the Open University.

Department for Transport, 2014. National Public Transport Access Nodes (NaPTAN) <https://data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251/national-public-transport-access-nodes-naptan> (accessed 5.13.20).

Uber, 2018. Commute Patterns in England and Wales. *Uber.* Accessed 21 May 2020. <https://kepler.gl/demo/ukcommute>.

Williams, L., 2010. Agile Software Development Methodologies and Practices. *Advances in Computers*, 80, pp. 1-44.

Yeo, I.-K., and Johnson, R.A., 2000. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, 87, pp. 954–959.

Young, K., 2018. GIF Porco Rosso Animated GIF on Gifer by Conjuris. GIF. *LowGif.* <http://www.lowgif.com/30eda5679770f2f2.html>.

Appendix

Table A.1: Team structure (for more details, please see individual reflections).

Team Division	***Team Member Names***	***Key Tasks***
Data Team	Hussein Mahfouz	- Data analysis methodology - Data preparation - Clustering - Classification
	Philip Wilkinson	- Clustering template - Classification base template - API construction - Lit review
Web Team	Cheyne Campbell	- Design of homepage - Interactive map page - Visualisation of flows using Mapbox - Report writing and editing
	Alicja Kotarba	- Design of about page - Interactive map page - Web visualisation using D3 - Connection to firebase
	Nathanael Sheehan	- Design of Story page - Web visualisation using d3, chartJS and ScrollMagic - CICD pipeline creator - DevOps