This study uses an unsupervised machine learning algorithm (K-Means cluster analysis) and spatial data on venues from Foursquare.com to investigate the reasons why rough-sleeper density is particularly high in the Westminster and Camden boroughs of London.

Rough Sleeping in Inner London:

a Cluster Analysis using the Foursquare.com data

Banu Simmons-Süer

Rough Sleeping in Inner London: a Cluster Analysis using the Foursquare.com data

1. Introduction:

Rough sleeping or homelessness is a growing social problem in central London. The problem that we want to focus here, is why certain London boroughs are where homeless people congregate most. In London, the borough Westminster has the highest number of rough sleeping people and the borough is considered to be the homeless capital of the UK.

As the number of rough sleeping people in London is rising, policy makers face the problem of dealing with this delicate social problem. A more radical solution to the problem, i.e. reducing the total number of people sleeping rough on the streets of London would require government policies that would devote resources to things like cheaper housing, and support for mental health issues. These are clearly beyond the scope of our study. In this study, we assume that our stakeholders are the Westminster and Camden Boroughs/Councils. We assume that the mission of these councils is to reduce the number of rough sleepers in their boroughs. In order to achieve this objective, one needs to gain insights to why rough sleepers prefer to bed down in these particular boroughs. If the reason is the availability of certain amenities in the vicinity, for example bathroom facilities in nearby parks or the convenience of chain coffee shops where excess, unconsumed food is made freely available, or availability of many underground entrances that offer a shelter, then these factors can be taken into account in local policy making. Setting up amenities such as the food banks for the homeless in the periphery of London, away from the touristic centre of Westminster or Camden would likely to reduce rough-sleepers in these boroughs. It is true that such policies would redistribute the rough sleepers in London, rather than tackling the heart of the problem. However, the displacement effect may be positive especially if moving to the periphery is coupled with counselling for mental health issues, and/or has a disruptive effect in the accessibility to harmful substances.

We will examine the problem in two stages: In the first stage, focusing only on the inner London boroughs, we will describe the relationships between the rough sleep rates (by square mile area and population) and the socio-economic characteristics of the boroughs. A priori, our expectation is that things like unemployment rate or the average annual pay in a borough are not very reliable indicators, given that rough sleepers are often not the original inhabitants of the borough that they bed down. We also think that things like average house price in the borough is not a meaningful variable even if one finds a high correlation between that and the rough sleeper ratio of the borough. Both Westminster and Camden have house prices that are well above the national average. However, it is a well-known fact that, majority of professional Londoners cannot afford accommodation in these expensive boroughs and finding a high correlation between rough sleepers and average house price does not imply any causality.

We will also examine if other borough statistics such as crime rates and ambulance incidents relate to the rough sleeper numbers. This is because the substance abuse/possession incidents and ambulance incidents may move in similar direction as the number of rough sleepers in each borough. According to Greater London Authority's "Chain October-December 2019 Report", only about 20% of all the rough sleep cases don't need alcohol, drug addiction or mental illness-related support. When we refer to the problem of "rough sleepers "or "homeless", we need to be careful in lumping people together and treating them as a homogenous category. People can be homeless for diverse reasons:

mental illness, addiction, family breakdown, and depression are among them. All we want to do in this first stage of the analysis is to let the descriptive data speak for itself.

In the second stage of the problem we will carry out some spatial analysis to support our hypotheses about the boroughs in question. We will use the foursquare.com's data about the venues in the area to help with our cluster analysis. The Westminster borough is a very touristic part of London with the Westminster Cathedral, many hotels, theatres, and restaurants. Rough sleeping people gather where there are services for them, which might be the reason why they hang out in this particular borough. In one of the neighbourhoods of the Westminster borough, the so-called "West End", theatres with their spacious entrances offer good shelter, where numerous homeless bed down on the steps once the audiences have gone home. Shop doorways also offer some shelter and, chain restaurants and cafes tend to donate their excess food and drinks to rough sleepers. Especially in summer, proximity to parks may be a desirable issue for the rough sleeping community. Parks are also attractive venues as they provide bathroom facilities and more privacy during the summer season.

2. Data

Research commissioned by the Greater London Authority found 8,855 people slept rough in the capital between April 2018 and March 2019. However, these numbers are based on the estimates for the year. Furthermore, these numbers are reported by a source (Statista), which is not freely available.

We choose to use the official rough sleeping statistics for 2018, released on the 31st of January 2019 by the UK Ministry of Housing, Communities and Local Government. These statistics are based on counts and estimates carried out by Local Authorities, providing a snapshot figure of the number of people sleeping rough on any one night. As the figures reflect a snapshot, they differ in the scale to model-based estimates of other sources. Furthermore, some people rough sleep (bed down) on a single night and is never seen again (intermittent rough sleepers). Given that the true number is impossible to know, we rely on the official statistics of the UK Ministry of Housing, Communities and Local Government. We use the rough sleeper statistics for 2018 as shown in the pdf file that can be found in the link below:

Table 1:Number of rough sleepers by inner London Boroughs

Table 3: Inner London boroughs by number of rough sleepers, 2018				
Local Authority	**2017**	**2018**	**Change on 2017**	**% change on 2017**
Westminster	217	306	89	41%
Camden	127	141	14	11%
City of London	36	67	31	86%
Lambeth	34	50	16	47%
Southwark	44	47	3	7%
Islington	27	43	16	59%
Wandsworth	13	25	12	92%
Hackney	18	23	5	28%
Kensington and Chelsea	20	20	0	0%
Hammersmith and Fulham	5	12	7	140%
Tower Hamlets	21	10	-11	-52%
Greenwich	8	7	- 1	13%
Lewisham	22	5	-17	-77%
Total (Inner London)	**592**	**756**	**164**	**28%**

https://www.homeless.org.uk/sites/default/files/site-attachments/Homeless%20Link%20-%20analysis%20of%20rough%20sleeping%20statistics%20for%20England%202018.pdf

In order to conduct our analysis on the socio-economic characteristics of the inner London boroughs, we get the data on variables such as population, unemployment, gross annual pay, crime rate, ambulance incidences and transport accessibility for each borough from the following source:

https://data.london.gov.uk/dataset/london-borough-profiles/london-borough-profiles.csv

There are 12 inner London boroughs and 20 outer London boroughs. “City of London” is not strictly considered a borough and therefore removed from the data set. We calculate two rough sleep ratios for the boroughs: In the first ratio we divide the number of rough sleepers in inner London boroughs (as shown in the first link) by the corresponding square mile area of the borough. We scrape the borough areas from https://en.wikipedia.org/wiki/List_of_London_boroughs using the BeautifulSoup library. Furthermore, we use the population of the borough to calculate the second ratio (rough sleepers per 1000 inhabitant in each borough). The motivation behind calculating two ratios is to confirm that the boroughs Westminster and Camden have the highest rate of rough sleepers irrespective of which measure is used.

As a next step, we scrape the table from https://en.wikipedia.org/wiki/List_of_areas_of_London.This table has data on all London boroughs and areas, with the corresponding post codes. After some data cleaning/wrangling, we drop the rows other than the two boroughs we want to focus on: Camden and Westminster. These two boroughs have by far the highest rough sleeper rate among the inner London boroughs. As some locations ambiguously have more than one borough (for example, Brent and Camden), we conduct further checks to decide whether to assign the area to the Camden borough or remove it altogether. In order to make this check, we also downloaded (from https://data.london.gov.uk) the Mapping-template-london-ward-map-2013.csv. Although, we don’t use the ward data directly, (the ward names often differ from the area names listed in Wikipedia), nonetheless, the mapping table is a useful guide in removing certain areas, as the corresponding borough was neither entirely in Camden nor in Westminster. These areas are Cricklewood and Tufnell Park.

The motivation for getting the postcodes is that we can get the latitude and longitude (coordinates) data for the areas of Camden and Westminster using the geolocator/Nominatim. Once that is done, the coordinates data are merged with the other data, and the main areas are marked in the London map.

Foursquare venues consists ten (top-level/first level) categories: 1. Arts and Entertainment, 2. College and University, 3. Event 4. Food, 5. Nightlife Spot, 6. Outdoors and Recreation, 7. Professional and Other Places, 8. Residence, 9. Shop and Service, 10. Travel and Transport. Using the coordinates based on the post codes, location and borough names, we queried the API by searching for the venues in Camden and Westminster boroughs. Our query was for the venues within 300 meter radius of each location with its corresponding latitude and longitudes. The resultant categorical data of detailed venues (106 unique categories) are converted to binary values through ‘one-hot encoding’, to use in the cluster analysis.

3. Methodology:

As we mentioned in Introduction, we analyse the issue in two stages:

3.1.Preliminary Analysis:

In stage 1, we examine the problem of rough sleeping by describing the relationships between the rough-sleeper statics and the socio-economic characteristics of the inner London boroughs. Next, we plot a selected number of socio-economic characteristics of these inner boroughs and the rough-sleeper rates using the seaborn library. The results of the preliminary analysis is shown in Section 4.

3.2.Cluster Analysis using the Foursquare data:

In the second stage of our analysis, we use the coordinates based on the post codes, location and borough names in Camden and Westminster boroughs, to query the API by searching for the venues within 300 meter radius of each location. The resultant categorical data of detailed venues (106 unique categories) are converted to binary values through 'one-hot encoding', to use in the cluster analysis. Then, our data are ready to do the 'K-means' cluster analysis.

Partition based cluster methods such as 'K-means' are implemented by using a distance function in order to find a global minimum using a defined metric, for example Euclidean. By using the Euclidean distance, K-means treats the data space as isotropic, meaning that data points in each cluster are modelled as lying within a sphere around the cluster centroid. Moreover, as clusters are modelled by the position of their centroids, K-means method assumes that all clusters have the same radius. When this restrictive assumption is violated, K-means may behave in a non-intuitive way. Furthermore, the Euclidean space is linear which implies that small changes in the data result in proportionately small changes to the position of the cluster centroid. This makes the K-means method susceptible to outliers, which can distort the results of the K-means. If there are significant differences in cluster density, the results of the K-means method may be difficult to interpret.

We face the problem of determining the optimal number of clusters a priori in using the K-means method. Unfortunately, there is no easy answer to this problem. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named 'elbow' and 'silhouette' methods, respectively. The elbow method is one of the most popular methods to determine this optimal value of k.

2. Statistical testing methods such as 'gap statistic': consists of comparing evidence against the null hypothesis.

Elbow Method:

The Elbow method looks at the 'total within cluster variation' or the 'total within cluster sum of squares' as a function of the number of clusters: One should choose a number of cluster so that total within sum of squares is minimized, to make the clusters as compact as possible. The elbow method minimizes the sum of squared distances between cluster points and their cluster centroids. When the

sum of squared distances of samples to their closest cluster centre is computed, the Elbow method is so called 'inertia' –based. When the average of the squared distances from the cluster centres of the respective clusters using the Euclidean distance metric, then the Elbow method is 'distortion' based. Normally these two calculations should result in a similar optimal k. Typically the calculated error measures (the within cluster dispersion) decreases monotonically as the number of cluster increases, but from some k onwards, the decrease flattens drastically. The location of this kink is called 'elbow' and it is supposed to give the optimal number of the clusters. Note that, the elbow method is a heuristic method and sometimes gives ambiguous results. An alternative is the average silhouette method (Kaufman and Rousseeuw [1990]).

Average silhouette method:

The average silhouette approach measures the quality of a clustering (i.e. how well each object lies within its cluster). The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990). For each k to the chosen kmax, the average silhouette of the data observations is calculated and the cluster location of the maximum is considered as the appropriate number of clusters. The function 'pam' (partition around medoids) is based on the search for k representative objects, called medoids, among the objects of the dataset (Kaufman and Rousseeuw 1987). These medoids are computed such that the total dissimilarity of all objects to their nearest medoid is minimal:

Gap statistic method:

The gap statistic by Tibshirani, et al. (2001) can be applied to any clustering method and it attempts to formalize the heuristic elbow method.

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be the value that maximizes the gap statistic. This means that the clustering structure is far away from the random uniform distribution of points. The algorithm works as follows:

The data is clustered, by varying the number of clusters from k = 1, …, kmax, and the corresponding total within intra-cluster variation is computed. Next, reference data sets with a random uniform distribution are generated. Each one of these reference data sets are clustered with varying number of clusters k = 1, …, kmax, and the corresponding total within intra-cluster variation is calculated. Finally the gap statistic is computed as the deviation of the observed total within intra-cluster variation from its expected value under the null hypothesis (i.e. computed with the reference data set with random uniform distribution.

Below we compute the results of our cluster analysis, using these methods to select the optimal number of clusters.

4. Results:

4.1.Preliminary Analysis:

As described above in (the Data) Section 2, the rate of rough sleepers can be calculated in relation to square mile area or population. Below in Figure 1, we observe that both area and population based Ratio 1 and Ratio 2 move in similar direction.

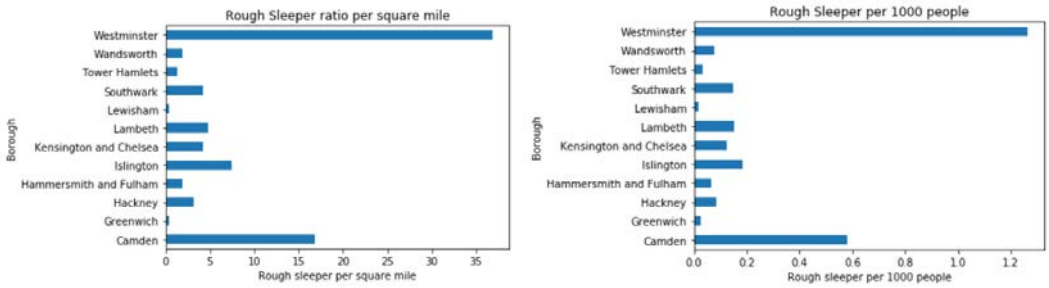


Figure 1 Rough Sleep Rates in Inner Cantons

As shown in the heat map of London below, the borough Westminster has by far the highest ratio of rough-sleepers, followed by Camden. Our analysis in this paper will therefore focus on these two boroughs.

Figure 2 London Heat-map of rough sleep intensity

Fig 3 shows that both rough-sleeper ratios are very similar and the calculated ratios lie on a straight line. Furthermore, it seems that the unemployment rate in a borough does not have a strong relationship with the corresponding rough-sleep ratio in the same borough. This is not surprising as one does not expect the rough sleepers to originate from the borough/location that they bed down. Indeed, there is no negative relationship between rough sleep ratio and the gross annual pay in the borough.

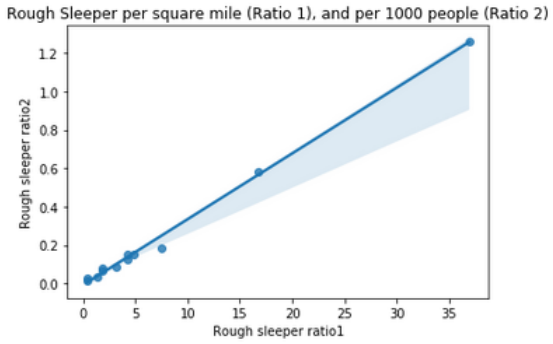


Figure 3 Rough Sleeper ratio 1 (per square mile) vs. ratio 2 (per thousand people)

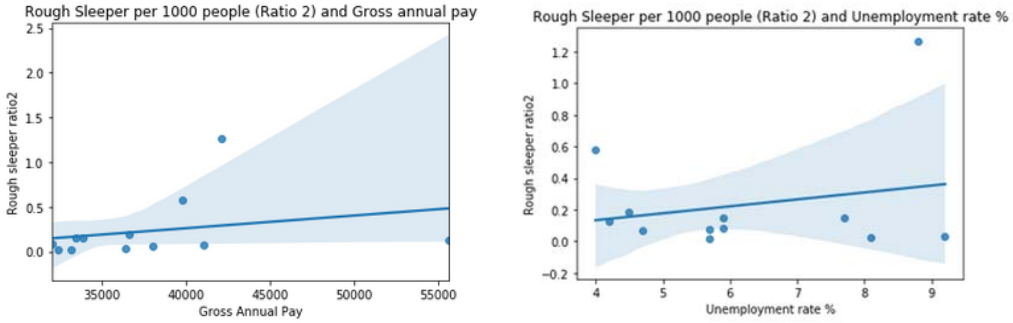


Figure 4 Rough sleep rates and socio-economic characteristics of inner London boroughs

Rough sleep rates seem to positively relate to the criminality rate and the ambulance incident rates in their boroughs, as can be seen in Figure 5. This observation does not necessarily imply causality. Nevertheless, ambulance incidents are usually high when substance abuse is common in a borough.

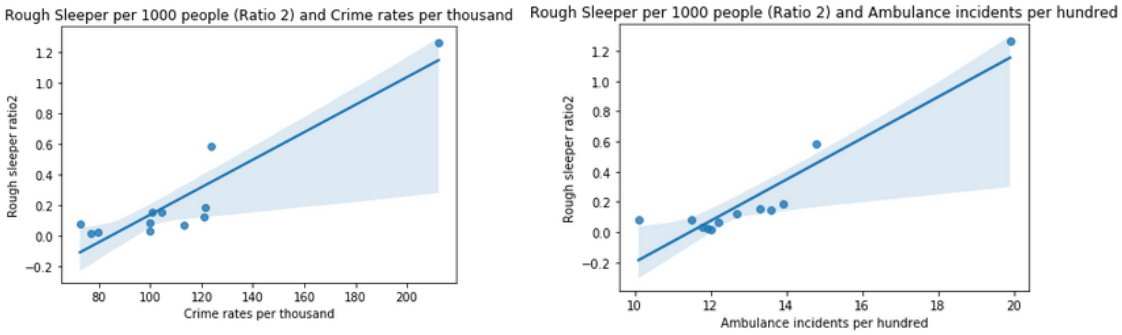


Figure 5 Rough sleep rates, Criminality and Ambulance incidents.

One of the hypothesis we wanted to test was that the rough sleep rates are positively related to the transport accessibility score of a borough. As we expressed in the Introduction section, we believe that rough sleepers prefer areas with underground stations as the entrances of underground stations[1] offer shelter from cold and rain. Thus, the positive relationship between transport accessibility/number of underground stations and the rough-sleeper ratio is not surprising.

[1] To the best of our knowledge, the number of undergrounds was not available in one table, it was manually derived through research from myLondon.com

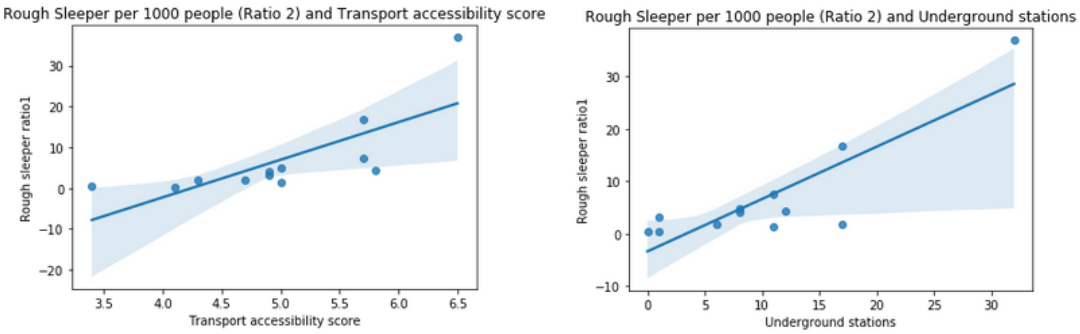


Figure 6 Rough sleeper rates versus transport accessibility score and the number of underground (metro) stations in a borough

We also calculate the correlations between the rough sleep ratio and the variables shown above:

Variables:	Correlation with rough sleeper per 1000:
Unemployment rate %	0.222480
Gross Annual Pay	0.259476
Transport accessibility score	0.714759
Underground stations	0.848301
Crime rates per thousand	0.916834
Ambulance incidents per hundred	0.940492

Table 2: Correlations of rough sleeper ratio with various characteristics of inner London boroughs

The calculated correlations confirm that unemployment rate or gross annual pay in a borough are not really relevant variables for the rough sleep rate in a borough. Camden and Westminster are relatively rich London boroughs. Given that the rough sleepers do not originate from the borough they bed down, the correlation between average pay and the rough sleep rate is unsurprisingly positive.

Crime rates, ambulance incidents are highly correlated with the rough sleeper rate. For the convenience of sheltering, the number of underground stations in a borough is also positively correlated with the rough-sleeper rate. Westminster is number 1 among the London Boroughs with the highest number of underground stations (32), providing support for our 'shelter-effect' hypothesis.

5. Cluster Analysis using the Foursquare data:

At first, we explore the venues in a single location ('Aldwych'). Not surprisingly, theatres is a dominant feature in this location. When the query is extended to all the locations in Camden and Borough, about 106 unique venue categories are obtained. By one hot encoding, we convert these data into binary variables to enable the K-means clustering method.

The hypothesis we want to test using the Foursquare data is that rough-sleeping tends to intensify in Westminster and Camden boroughs as these boroughs tend to have locations (neighbourhoods) that have clusters of theatres ('shelter effect' of large entrances), hotels as hotels in central locations imply affluent tourists. Most begging activities take place when numerous tourists are moving

around the city's main attractions. The presence of hotels guarantees the circulation of tourists all day long, benefitting vagrancy.

We also hypothesize that the Camden and Westminster boroughs would contain clusters of garden/park type of locations with convenient bathroom facilities , offering some quiet and privacy in summer.

Another hypothesis is that, cafes, coffee shops, bakeries especially the chain ones such as Café Nero, Starbucks, and Pret a Mange may be a distinct cluster, given that kind passer-bys are more likely to enter a chain coffee shop to buy a rough sleeper a sandwich than buy lunch at a restaurant. Furthermore, the coffee shop type of venues tend to have daily fresh produce and the excess food is likely to be donated to the rough-sleeping community.

Choosing the Optimal K:

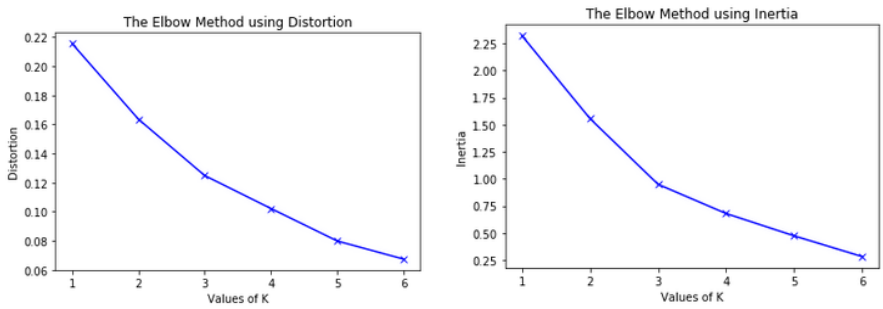


Figure 7 The elbow method in selecting the optimal cluster

At first sight, there appears to be a slight kink at K=3 in the inertia version of the Elbow method when we set the maximum cluster at kmax=7, however, the Elbow method is ambiguous in this case. We also checked the 'silhouette' method below. The score we calculate is plot against the number of clusters. It seems that the score keeps increasing with the number of clusters not quite reaching an optimum with kmax=7. In fact, the silhouette score tends to favour k=8, which yields implausible (too-detailed) clustering.

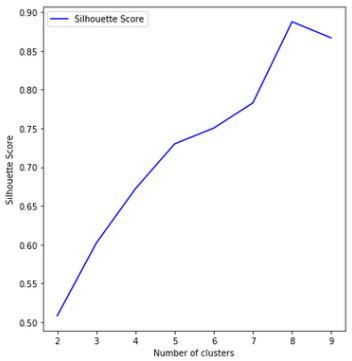
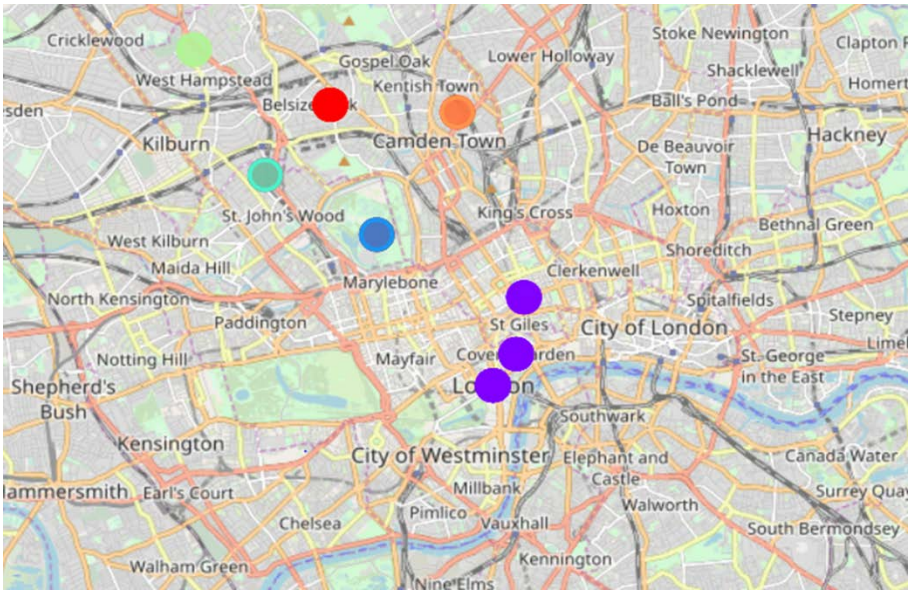


Figure 8 Average silhouette score method

We believe that, intuitively, more than six clusters for a relatively small area is implausible. For the purpose of our analysis, distinguishing coffee shops from bagel shops is not really meaningful and

outliers can distort the results[2]. As k=6 is a compromise mid-number between the elbow and silhouette score , we chose and implemented that as the cluster number.

As the information on rough sleeper rates is at borough level, we have no way of knowing the exact distribution in various locations of Westminster and Camden. Nevertheless, it is well known that the areas near theatres and hotels are the most intensive locations. This area corresponds to Cluster 1. As expected, Parks and Gardens (Cluster 2) is also prominent in the boroughs where rough-sleep rate is high. Below, we show which locations/neighbourhoods in Camden and Westminster corresponds to the cluster labels 0 to 5.



Red: Cluster 0 (Cafes) Purple: Cluster 1 (Theatres and Hotels), Blue: Cluster 2 (Parks & Gardens) Turquoise: Cluster 3 (Café/Yoga studio) Light green: Cluster 4 (Gyms/Fitness centre) Orange: Cluster 5 (Skate park, various stores)

Figure 9: K-means Cluster Labels

The locations in the Westminster and Camden boroughs belong to the Cluster labels shown below:

[2] When we check the 'gap statistic', we get k=6 as the optimal number. However, the code snippet to calculate the statistic is neither a standard part of the python library, nor our original work. Therefore, we leave it out.

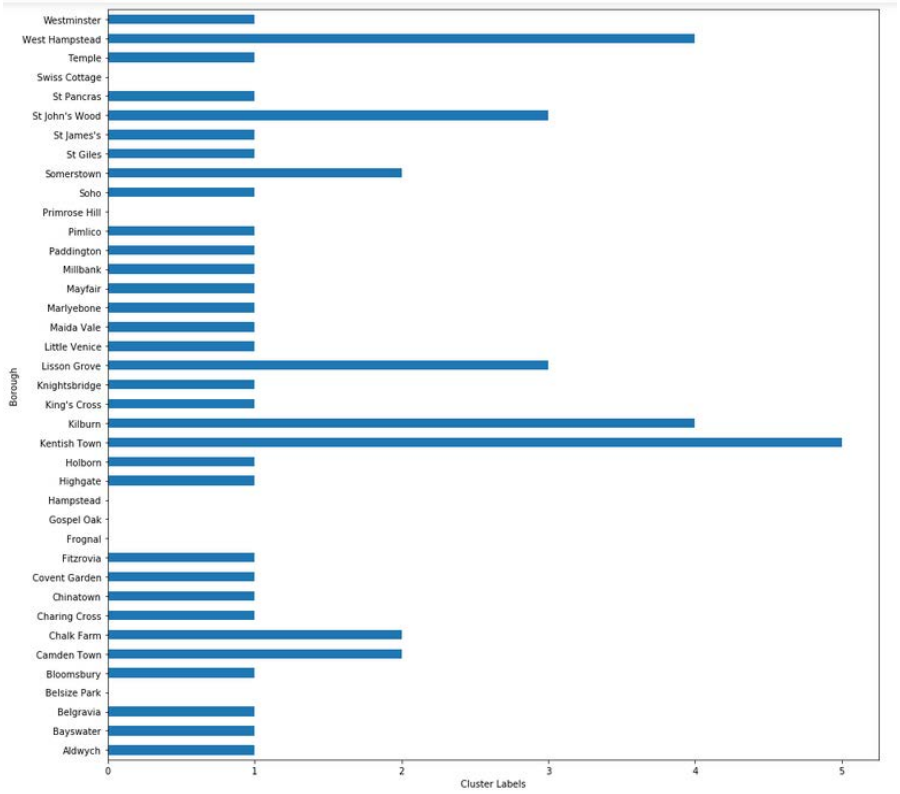


Figure 10 Camden and Westminster Locations with the corresponding cluster labels

Next we superimpose the main four homeless centres in Camden and Westminster boroughs on the cluster map. As we expected, one of the homeless centres/charities is very close to the Cluster 1, where the density of homeless/rough-sleepers is highest in Westminster. Likewise, another homeless centre in Camden is between Cluster 0 and Cluster 5.

Potentially one could address the chicken and egg problem: are the rough sleepers preferring the area because of the venues or because of the homeless centres? The answer is probably both. However, rough sleepers have been bedding down in the theatre district for more than 30 years (Cluster 1 (purple). The homeless charity 'the Connection' (the black and red circle in the purple cluster) was founded in 2003. Thus, this clears the question of what came first: the charity moved here because of the rough-sleeper density.

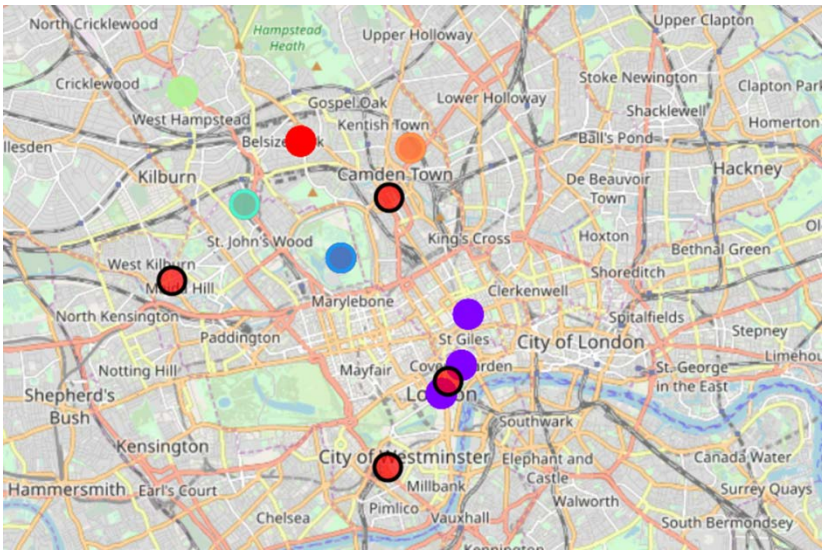


Figure 11 Homeless Centres and Cluster Labels

6. Discussion:

The purpose of this analysis was to gain some insights to the rough-sleeping problem in Westminster and Camden. The rough sleep statistics are only available at borough level (not at location level), therefore, our hypotheses were based on anecdotal evidence (personal observation and journalistic sources). We hypothesized that rough sleepers tend to concentrate in the vicinity of theatres, hotels, parks & gardens and cafes.

Our cluster analysis indeed backs up these hypotheses by identifying these as predominant clusters in the area. The optimal cluster was chosen as K=6. It is possible that our K-means procedure was prone to outliers and considering 6 clusters was an exaggeration. For that reason, the sensitivity of the results to a more coarse clustering (K=3) was also analysed. The K-means procedure identified Parks & Gardens, Theatres & Hotels as distinct clusters even when K was set as 3 (see Appendix). Thus, the key finding of our analysis remains valid when a smaller number of k is selected.

Some of the locations in our data are very close to each other (i.e. possibly overlapping in the 300m radius) and have the same post code. These observations inevitably creates some duplication of the venues. Nevertheless, it is advisable to keep such duplicate cases in the data for the purposes of this study. As our rough sleep statistics are available at borough level anyway, we are more concerned about forming an overall cluster profile of these boroughs. It is unlikely that our results are prone to any duplication-related distortion. Furthermore, we wanted to be able to identify the cluster-labels of all the known locations in these boroughs, without dropping anyone of them.

7. Conclusions:

Our spatial analysis of Westminster and Camden via K-means clustering using the Foursquares.com data revealed that these boroughs contain attractive venues for rough-sleepers.

Our analysis showed that the socio-economic characteristics of these boroughs do not really explain the high ratio of rough sleepers in the area.

When the cluster labels and the homeless help centres are superimposed on the London map, it is clear that these centres are especially close to Cluster 2 (Parks & Gardens) and Cluster 1 (Theatres and Hotels). These results imply that these are the locations where the density of homelessness/ rough sleeping is at its highest.

It was reported in English newspapers such as ‘the Guardian’ that soup kitchens act like a magnet for not just vagrants but also drug dealers. Our preliminary analysis indicated that criminality rates and ambulance incidents are indeed highly correlated with the rough sleeper rates in inner London boroughs.

In order to avoid stigmatizing any group of individuals in discussing this sensitive social problem, we deliberately left out the nationality/immigration profiles of the London inner boroughs from the scope of our analysis.

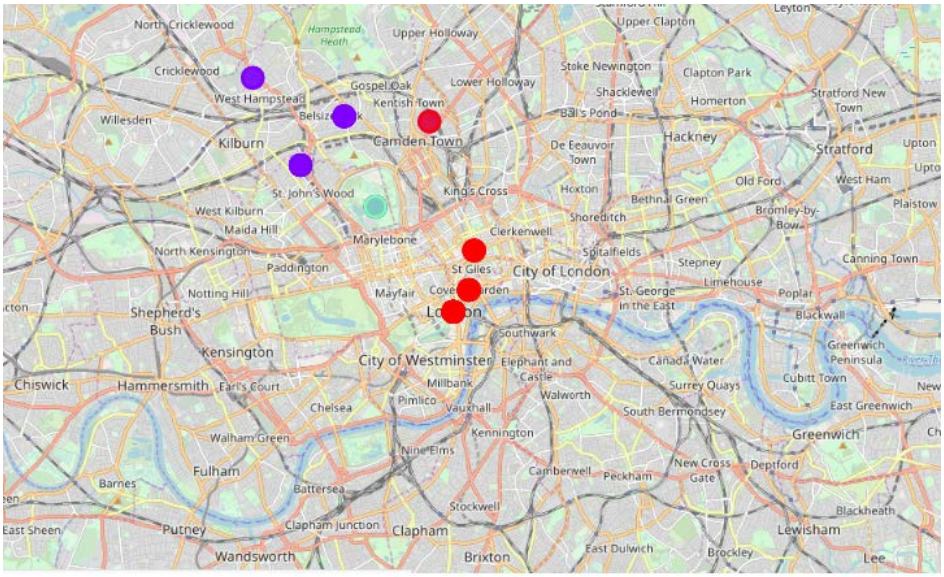
Our analysis concludes that at local council level, there are not many policy alternatives given that ‘not helping’ is not a humanitarian option. As the attractiveness of the venues cannot be changed,

the only option to reduce the number of rough sleepers in these particular boroughs is to move the homeless help centres elsewhere.

At national level, according to a report by BBC on Feb 26th, 2020, the UK government pledged to allocate 236 million GBP to help rough sleepers (https://www.bbc.com/news/uk-politics-51653744). Clearly, a national policy is a more effective way of tackling the problem.

8. Appendix:

Clusters when K means is calculated with K=3



Red: Cluster 0 (Theatres, Hotels), Purple: Cluster 1 (mostly Cafes), Light blue: Cluster 2 (Parks & Gardens)

References:

Kaufman, L. and P.J. Rousseeuw, Finding Groups in Data (John Wiley & Sons, New York, 1990

Tibshirani, R., G. Walther, and T. Hastle, Estimating the number of clusters in a data set via the gap statistic. J.R. Statist. Soc. (2001), 63, pp 411-423