Coursework Report Template for Module INM433 “Visual Analytics”

Baber Abbasi

Abstract—We analyse the stop and search rate for London for the year 2019. We analyse how the stop and search rate varies through London and try to explain that with aggregated demographic factors using spatial binning and geographic weighted regression. However our results are inconclusive although there is some correlation on the aggregate for stop and search rates with demographic factors, we fails to discover a causal link.

1. Problem Statement

Stop and Search (SS) is a legal power that allows police officers to search a suspect for prohibited items such as drugs, weapons or stolen property. Critics have consistently suggested that these powers consistently target minority,— especially black communities— however the police see it as a vital tool in fighting crime.With the UK government considering extending the powers of the police to provide them with greater leeway and freedom with these powers, we believe there needs to be more research on this hot button issue. This report focuses on exploring the spatial-temporal dynamics of all stop and search incidents in London in 2019. It addresses the following questions:

- Is there any temporal pattern to the stop and search incidents and does it vary with crime?
- How do the number of stops vary through London?
- Can we identify demographic factors in boroughs with more stop and searches?
- Are these factors similar throughout London or are there local variations?

The data is suited to answer the above questions as it includes all SS incidents for 2019 described with date and time, an anonymised location (Longitude, Latitude) near the incident, and an officer defined race and census demographic data. It is taken directly from the police department and includes demographic details about the persons searched.

2. State of the Art

Reference [1] analysed the rates by which New Yorkers of different ethnic groups were stopped by the police. They analysed 175,000 stops on a 15 month period from January 1998 to March 1999. The stops were aggregated the stops into 77 police districts. Population estimates were calculated by merging census tracts to their corresponding police districts, and they estimated corresponding day and night population estimates. The variability among the precincts was modelled by partitioning the precincts by proportion of ethnic minority into three levels and estimating through a multi-model model. We also utilise the same aggregation method by aggregating the stops into census tracts however variability between the tracts will be modelled through a geographic weighted regression.

Reference [2] uses a unique aggregation technique to analyse the ethnic breakup of police stops in New York City. They utilises a 2-D space time view to visualise stop and search rate by using a double faced image with a map of New York on the bottom and a time scatter plot extending on top. The years (2006-2019) are on the y axis and the number of stops are represented on the x-axis (if the image is seen in landscape). They populate this scatter plot by stops precise to the minute and street intersection. Their aggregation method visualises the ethic category that was stopped the most at that point in space-time. We also investigate the spatial breakup of ethnicity, however we experiment with hexagonal bins of different sizes to visualise their extent. As our data is restricted to 2019, for time series analysis we use line graphs which allows us fine-tuned analysis of total search rate for individual days. The major benefit of both these methods is that in addition to filtering by ethnic categories we can also investigate other elements such as stop reasons and outcomes.

3. Properties of the Data

The data is collected by the UK government[3] who regularly release all stop and search data in quarterly instalments. The data for 2019 includes 235,814 incidents along with an approximate anonymised location, the date and time, the object of the search, gender, the age range of the suspect and an office described ethnicity. The suspects were overwhelmingly male with approximately 6% female records and only 100 records with other type. The ethnic breakup is approximately 40% Black, 47% White, 17% Asian and 4% Other. We will be utilising the officer defined ethnicity rather than the self defined one as this is more inline with census data. Moreover reported crime statistics were also download ed from the same location which had approximately the same columns except the date is restricted to the month the crime was reported.

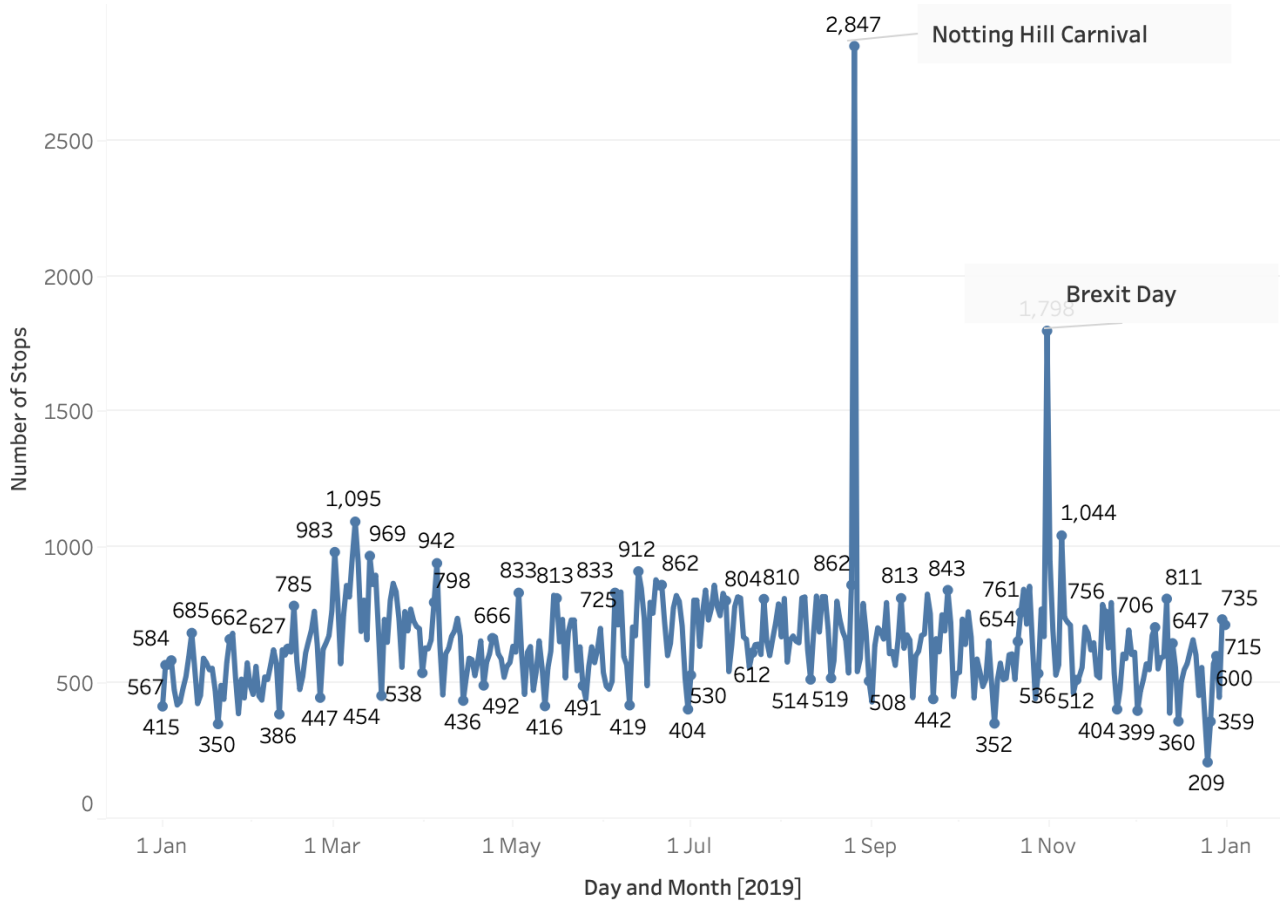

Figure 1. Absolute count of searches

There are 82,331 records without any geo-positional data which we will omit in our spatial analysis however utilise for the temporal analysis. There are also 4212 records where the searches were carried out only on vehicles which we will omit as well as we are concerned with demographic data of suspects searched.

The geo-spatial data for the searches include the latitude and longitude values, however the co-ordinates are anonymised to an approximate public location near the incident before distributing. This might be a centre point of a street, a public or commercial place or a large postcode catchment area. Hence we will be careful not to aggregate the co-ordinates into a too fine granular census area. Census area shape-files were downloaded from [4] which included polygon geometry co-ordinates. The stops were aggregated into the census data through a spatial join.

Outliers

Figure 1. plots the number of SS events through time in granular day format. We can see the trend is pretty flat except for two days on 26 August and 31 October. Further investigation revealed that the anomalous events on the two dates were not representative of the SS rates. The former was a bank holiday Carnival event in Notting Hill where more than a million people attended and the latter was a Brexit protest due to the request for a delay of ratification. Both dates were omitted from our spatial analysis.

4. Analysis

4.1. Approach

Our Approach is the following:

Initial Analysis:

Step 1: Clean up

We use graphs and summary statistics to explore the data. Missing values are removed, outliers are dealt with and data is cleaned up.

Step 2: Temporal Analysis

For the Temporal analysis we will make use of line graph and tables to explore the data. We have one row for each incident and the computational task will be to aggregate and filter the counts of data. The visual approach will mainly be dealt with line graphs where we will choose the appropriate scale or visualisation

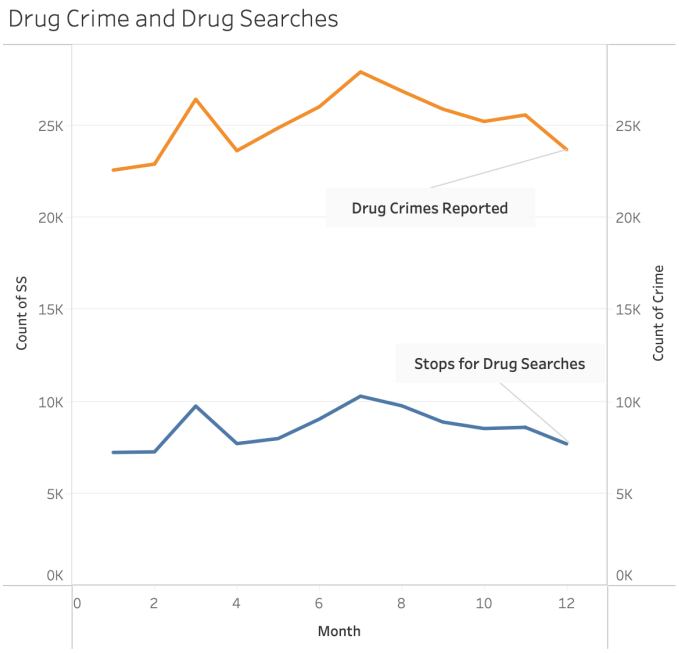

Figure. 2(a)

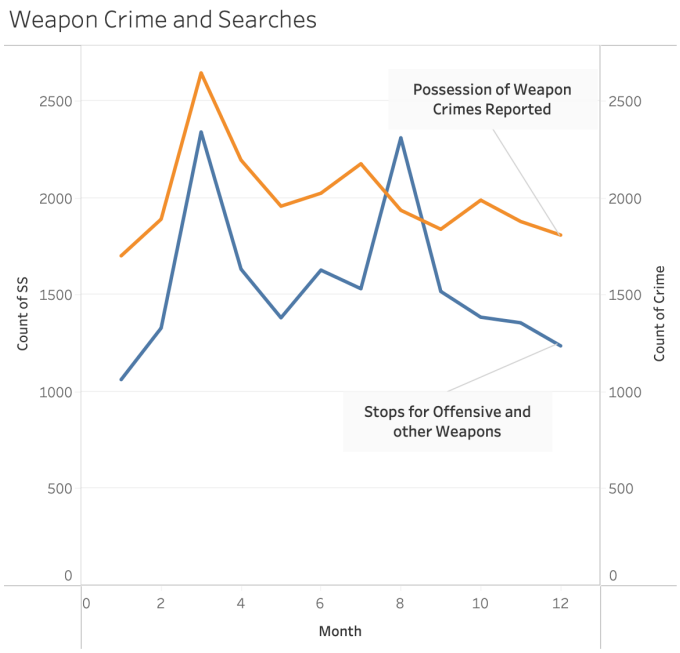

Figure. 2(b)

Step 3: Spatial

For the temporal approach I will be using the geo coordinates of the stops and search incidents. My task is mainly to find the appropriate metric to find out some pattern from the data. I will start by plotting the points on the map, however with the large amount of points the map is expected to get crowded. I will then try some aggregation approaches to find a global pattern from the data: this will include kernel heat maps and hex aggregation. The computational task will be to aggregate the points to the appropriate metric and my task will be to choose the appropriate scale for analysis.

Step 4: Spatial modelling

Our next task will be to merge the points to an appropriate census track to link the census data to the aggregate of points. The last step will be instrumental in this process as there are multiple size of tracks which we can choose. We will start from a rough idea from the last step and aggregate upwards. The visual task will be to illustrate through choropleth maps and visually identify how the variation among the census tracts. We will further be making use of scatter plots and correlation heat maps to model the rate of stop and searches to demographic data. This will allow us to make a parsimonious model which describes the attributes in the census tracts which lead to great search rates. The model(s) will be validated through regression and QQplots.

4.2. Process

We started by visualising the number of crimes and the number of stop and searches by means of a line graph. As the crimes data was aggravated by month we could not go into the granular details and could only analyse the monthly trends. The absolute values of crimes are more than the number of stop and searches by a factor of 25 on monthly average, so we decided to filter by subsets of crime. We decided to restrict our analysis to subsets: drugs as they form the majority of the reason for searches; and knife and weapons as abating those crimes is the major justification given for SS.

Figure 2(a) and Figure 2(b) illustrate the counts of reported crimes and the respective stop and search reasons. The trend for drugs almost mimics that of crimes reported. Similarly the trend for weapon crimes follows those of reported crimes during the first half of the year and is mostly similar on the second half as well except for a jump in August during the carnival festival. The causal relationship is inconclusive however as we are comparing searches to *reported* crimes and one would naturally have an effect to another.

Spatial Analysis: Exploratory Analysis

We next turn to spatial analysis. As we have a great number of points we look for a suitable aggregation metric to represent our data properly. The points on their own were too crowded and we experimented with kernel density estimation however we failed to visualise any local patterns: We noticed that the centre of London was more dense then the outskirts when zoomed out of the map, however zooming failed to provide us with any local patterns as the map was still too crowded. Although we will be using the predefined census boundaries for our analysis later this step was necessary to explore if there was any local variation.

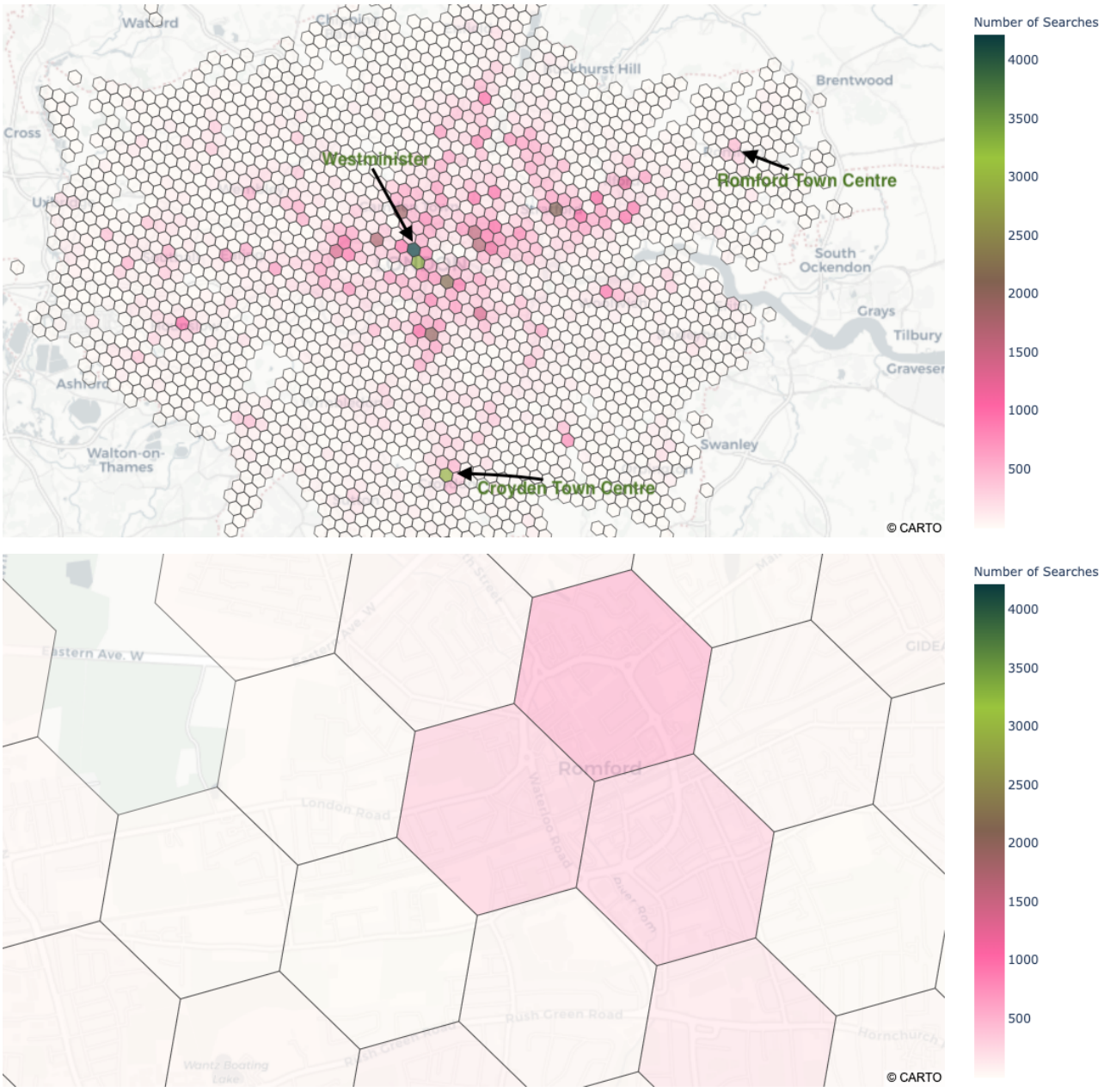

Figure 3(a) Top, Figure 3(b) Below: Hex Bins

We aggregated the points into equally sized hexagonal bins. As the co-ordinates of our data is anonymised to predefined public places, the challenge was was to find the appropriate aggregate size so as not to bias the aggregation. We experimented with different sized bins but finally settled on 0.73km squared bins. As the police has not released the locations of the anonymised locations we had to use human judgement to specify the hexagonal bin size. We used a base map with panning and zooming functionality and by visually inspecting the locations of the bins — for example checking if there is more coverage in popular areas and inspecting multiple bins for the coverage of streets—we settled on the appropriate scale which is shown on Figure 3(a).

The figure shows that there is great variability in the absolute count of SS. Although the counts are mainly consolidated in the centre of the city, there are a number of local clusters emerging throughout. The map is interactive which allowed us to investigate further (as illustrated in Figure. 3(b)) and identify the locations of interest. For example there are local clusters at Croydon and Romford town centres which are pretty popular destinations, as well as areas of with significant

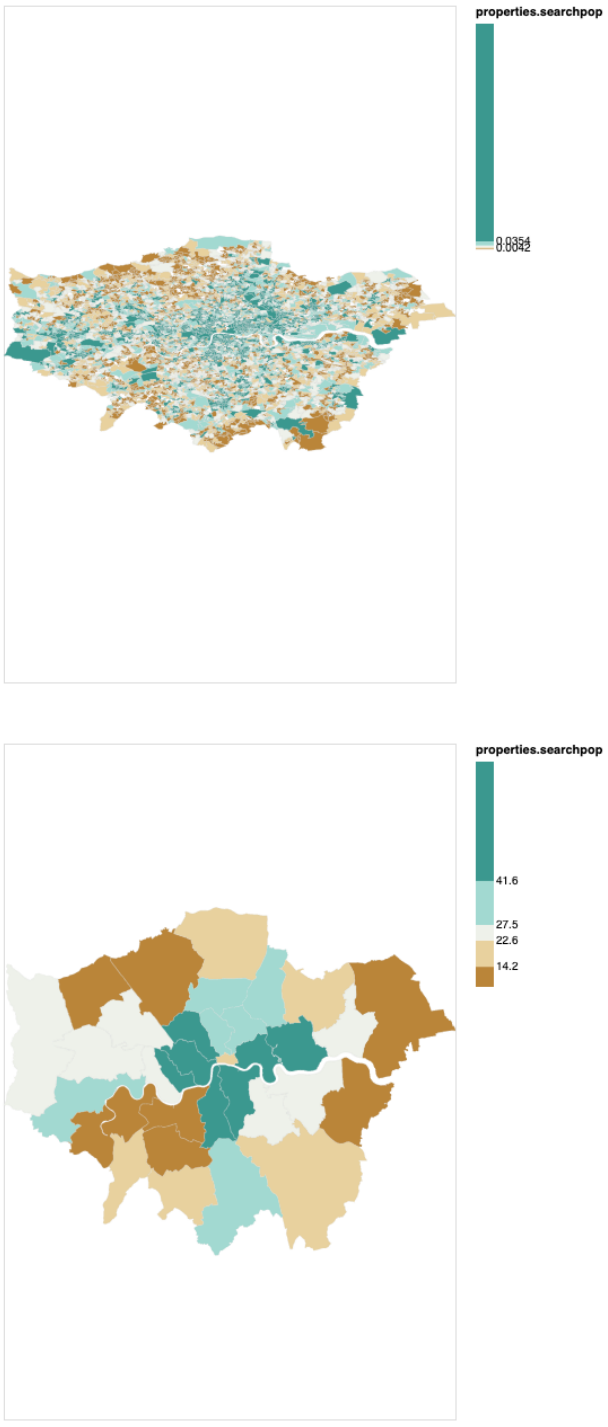

Figure 4(a) (top) and Figure 4(b) bottom

mixed non-white populations. Westminister was an outlier with two hex bins of 4223 and 2925 close together. For reference the 95th quantile is 166. This is probably less due to regular police concerns and more to do with national security matters as it home to most government buildings.

Spatial Analysis: Model

After the results of the exploratory analysis in the previous section we decided to explore the factors of the variability between the different areas, especially the difference in the number of search between the centre and outskirts. We decided to find a suitable census level for our analysis to focus our attention to demographic factors. Figure 4(a) illustrates the stop and search rate at the LSOA level normalised to the population. We use a diverging colour scheme with a quantile hue and the units show great variability in the amount of searches relative to the population. The mean is 0.03 and the median is 0.12. However the maximum is 1 for a LSOA in Westminster which is not possible. This could be an artefact from the ansonymyzation process but more probably this is because the day-time population of some units — especially non-residential ones– is much greater than the those living there. The distribution in Fig 4(a) generally follows that of Fig. 3, however, accounting for normalisation so we decide it is reasonable.

From the census demographic data we short-listed 11 variables which we assume to have explanatory power for the search difference between the units. Our focus here is on the demographic make up of the communities rather than crime statistics as we are analysing if there a specific aggregate characteristics which are more of a target of search and stops. Our investigation of scatter plots in Figure 5(a) revealed none of the variables had much of a relationship with our target variable however.

We tried to go up one level and use the MSOB. Although the scatter plots illustrated a slightly better relationship, they were still not statistically significant with our dependent variable. At the Borough level however, the relationships are clear as illustrates in Fig 5(b). We decided to use the Borough level as our unit of aggregation.

Our first model was ran on all 11 variables. Further investigation of the scatter plots and the correlation matrix in Fig. 6 revealed that several variables had high correlation with each other and some had low correlations with our dependent variable. We note that private transport to work is highly correlated with all variables and especially our dependent variable but we decide to exclude it as we expect that the percentage of people this describes are a minority in a metropolitan city like London. Moreover Black—the

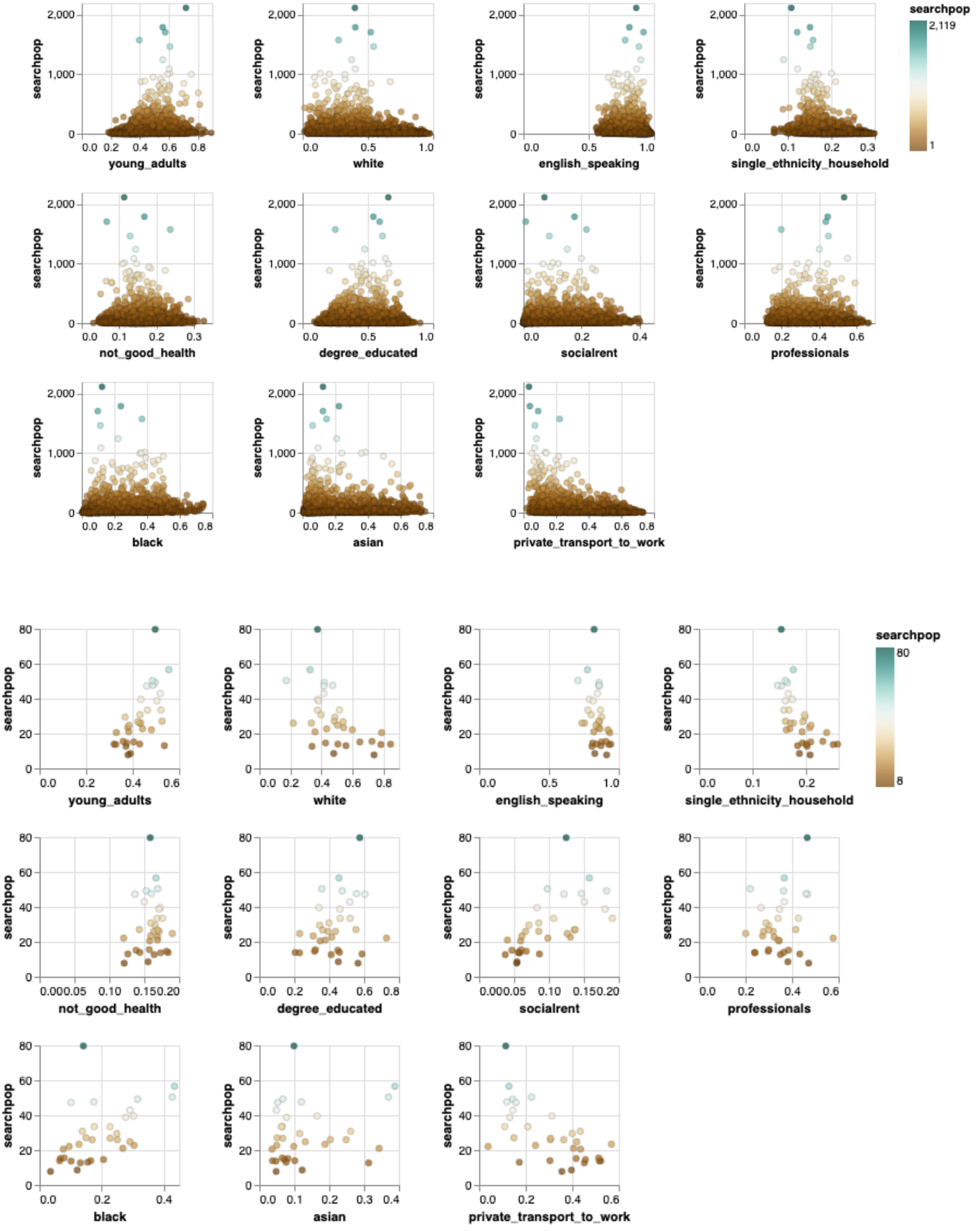

Figure 5(a) (Top) Scatter Plot LSOA and Figure 5(b). Scatter Plot Borough

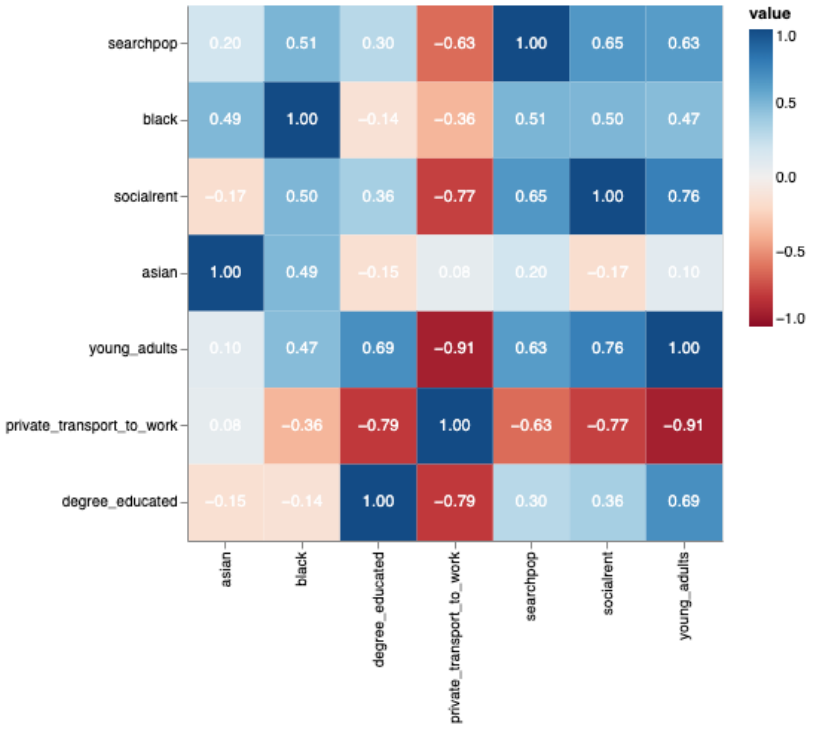

Figure 6 Correlation Plot

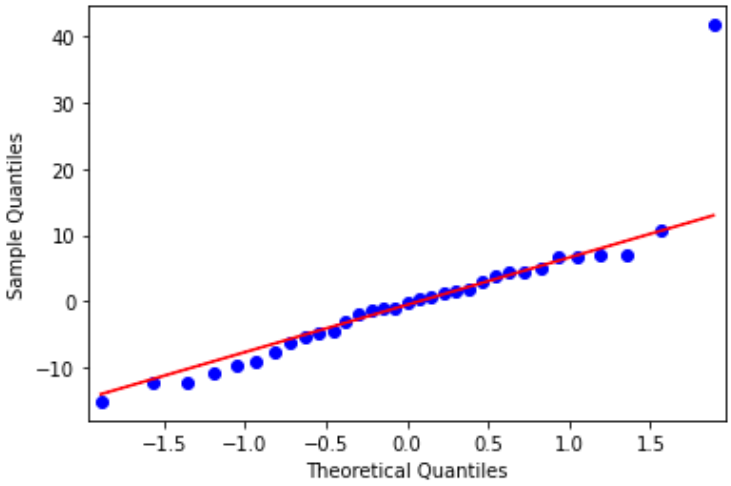

Figure 7 QQ plot of Borough level Regression

proportion of Black people in the borough— is correlated to the proportion of Asian people. As Multicollinearity makes coefficient of dependent variables unreliable in Linear Regression and we are aiming for interpretability, we decide to omit the Asian variable as we feel that is more appropriate to our research question and the general discussion around stop and search.

Further experimentation left us with 4 variables which were : the proportion of black people in the borough, proportion of people on social rent, the promotion of young adults in the borough, and the proportion of degree educated people in the Borough.

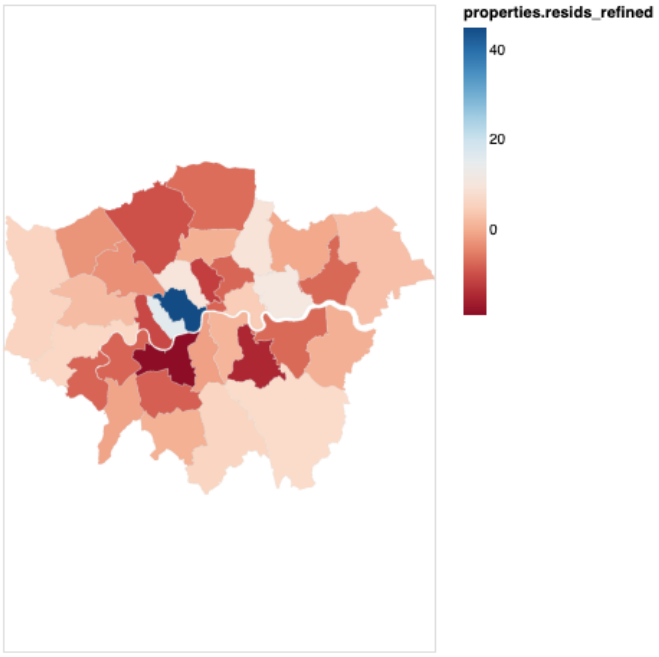

Figure 8 Residual Map.

The residual map of our final model is shown in Fig(). The R2 metric is 0.43 but as can be seen in the residual map, Figure 8, the model mostly underestimates most boroughs except for Westminster. However Westminster is an outlier as previously mentioned and the QQplot shows signs of normality.

4.3. Results

The answers to our research questions are:

-There are no temporal patterns as such to the stop and searches. The trend for total number of searches is pretty constant throughout the year and although there are fluctuations in individual categories of searches throughout the month, there is no apparent pattern. The correlation with crime rate is inconclusive. There is a correlation with reported crime but that is expected as more searches will lead to more crimes apprehended.

-The number of stop and searches vary through London with more searches in the central areas and fewer in the surroundings. More populated areas naturally have more searches and we can infer that police target populated areas more than isolated.

-We can infer a range of demographic factors which are correlated on the Borough level with more searches. However we were unable to provide a causal link. Some areas with some attributes on the aggregate level (such as more ethnic minority communities or more people on social rent) are targeted by the police for the searches.

-The data is inconclusive ablaut the local variations. We could not model the geographic weighted regression and the Borough level was too large to analyse any local variation.

5. Critical reflection

The result that the features were correlated on an aggregate level but the relationship broke down on a local level was clearly a case of Simpson's Paradox [5]. Clearly there was a background factor which we were missing which lead to such results. One problem is that the police just started releasing granular reports of stop and searches in 2018 and there is not much literature studying the causal effects of stop and searches in the UK. There is more research in the United States and there the link is more obvious, probably due to the fact communities live in a more segregated way.

Our main problem was one of aggregation. We failed to find a suitable metric to aggregate the geopoints. This could be a data integrity problem due to the nature of anonymised data. An alternative to GWR and Linear Regression could have been clustering the Lower level census units so find out the common attributes. However we were aiming for descriptive statistics and clustering is not as transparent as simple regression models and are quite prone to subjectivity.

Another alternative approach we could have taken was using the points themselves. Perhaps through some clustering algorithm the such as K-means, the data could have been aggregated so as to find suitable hotspots for our analysis. Or something along the lines of DBSCAN where the points and the noise are separated and we could have shortlisted clusters and focused our attention to those clusters discovered.

On the other hand we were looking for demographic reasons for stop and search and for that we naturally will have to use census or other aggregated metric to reference one dataset to another. The stop and search data does include some demographic data such as age and ethnicity which we could have focused on, so perhaps our analysis was not suitable to the data. Perhaps aggregating the points through some demographic data would have been more appropriate.

This is further aggregated by the fact that we were using census data from 2011 and we expect there has been demographic change in the intervening 10 years. The new census should be out soon in 2022 so we could make use of that. In addition the data on the breakup of daytime and nighttime population would have been quite helpful.

Table of word counts

Problem statement	195
State of the art	312
Properties of the data	373
Analysis: Approach	344
Analysis: Process	1106
Analysis: Results	181
Critical reflection	371

References

[1] Gelman, J. Fagan, and A. Kiss, 'An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias', *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 813–823, Sep. 2007, doi: 10.1198/016214506000001040.

[2] R. Shapiro and F. A. Pearman, 'Using the interaction geography slicer to visualize New York City Stop amp; Frisk', in *2017 IEEE VIS Arts Program (VISAP)*, Oct. 2017, pp. 1–8. doi: 10.1109/VISAP.2017.8282370.

[3]'Home | data.police.uk'. https://data.police.uk/ (accessed Jan. 09, 2022).

[4]'2011 Census data catalogue - Office for National Statistics'. https://www.ons.gov.uk/census/2011census/2011censusdata/2011censusdatacatalogue (accessed Jan. 09, 2022).

[5]A. Gelman, J. Fagan, and A. Kiss, 'An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias', *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 813–823, Sep. 2007, doi: 10.1198/016214506000001040.