Introduction

Method

This study uses data from LSOA atlas downloaded from London Datastore (2014). All the data was for the year of 2011 and was at LSOA (Lower Super Output Area) level. Before analysing the data, a data cleansing was performed, where all LSOAs with 0 value in median house price were removed, as there was no property in those LSOAs.

The median house price data (`MedianHP`) was from Land Registry. The variable `kMedianHP` was it divided by 1000, and was used in this study as an alternative for the purpose of concise. **Figure 1** shows its spatial distribution. In general, it reveals a double-hotspot pattern with exceptionally high median house prices in LSOAs in south Barnet, Camden, Westminster and Kensington and Chelsea. Median house prices are generally higher in the north London than in the south, and higher in the west than in the east.

The socio-economic predictors chosen for this study are listed in **table 1**. After checking their correlations with the dependent variable (**table 1**), `c_per_hhlds` and `Pct_CHDC` were excluded from the predictors. The multicollinearity of the remaining predictors was then checked using VIF. A VIF smaller than 5 indicates the predictor is safe from multicollinearity (citation on VIF). **Table 2** shows VIFs for all of the selected factors. All variables had VIFs smaller than 5, suggesting that there was no multicollinearity.

The socio-economic predictors used in this study include:

1. Median annual household income (`MedianIncome`). This factor represents the economic dimension in the process of gentrification. Figure 2.1 shows how it spreads across London. It has a very similar spatial distribution with the median house price data, where southwest London (*Richmond*, *Wandsworth* and *Merton*), central west London (*Westminster* and *Kensington and Chelsea*) and northwest London (*Camden* and *Barnet*).
2. Percentages of non-white population (`Pct_nonwhite`). It is calculated by dividing non-white population with total population for each LSOA and represents the ethnic dimension. Figure 2.2 shows its spatial distribution. It has four clear spatial clusters in west London (*Ealing* and *Hounslow*), northwest London (*Brent*), east London (*Redbridge* and *Newham*) and south London (*Croydon*) where non-white population percentages are the highest across London.
3. Public Transport Accessibility Level (PTAL) is a measure of public transport accessibility (Shah and Adhvaryu, 2016). The PTAL ranges from 0 (very poor) to 6b (very high).
   The average PTAL (`PTAL_average`) data was from TfL. It essentially reflects the density of public transport in the area. Figure 2.3 shows its distribution. It basically follows a radial pattern with higher PTAL average in inner London and lower in outer London, except some peripheral hotspots including *north Croydon*, *north Richmond* and some LSOAs in *Greenwich*, where the average PTAL is relatively high.
4. Percentage of highest level of qualification above level 4 (`Pct_qual_above_l4`). This is the percentage of people who hold a level 4 and above highest qualification. It is a proxy to determine the education backgrounds of residents in the area. Its spatial distribution pattern is demonstrated in figure 2.4, which shows higher overall education levels in inner and southwest London LSOAs while lower in other boroughs'.

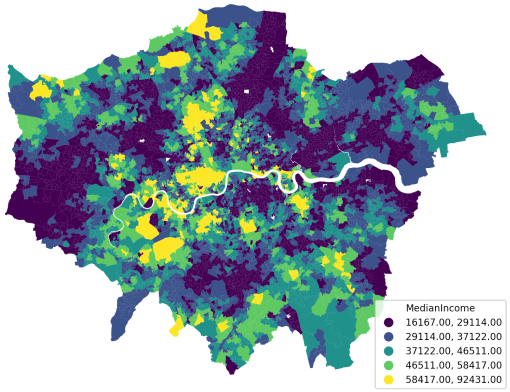

Figure 2.1

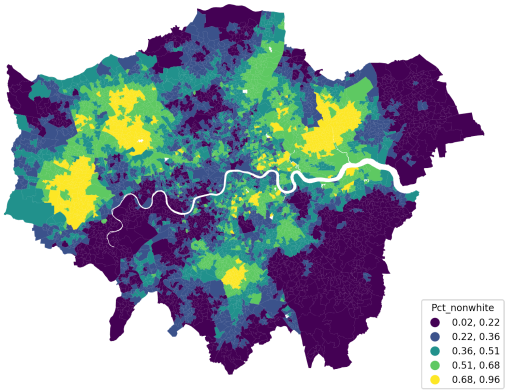

Figure 2.2

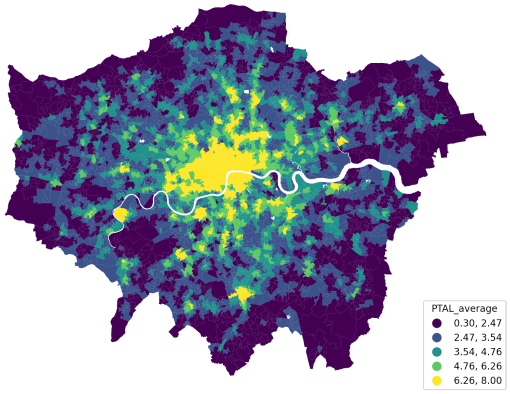

Figure 2.3

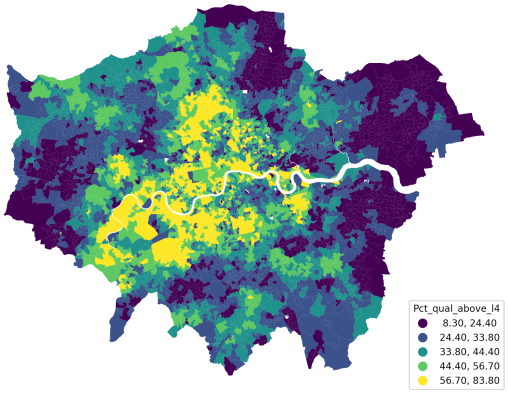

Figure 2.4

A simple Ordinary Least Square (OLS) regression model was first developed with the listed predictors as its independent variables and the LSOA median house price as its dependent variable. The model results are presented in **figure 3**. The global model has an adjusted R^2 value of 0.55 which means around 55% of the data can be explained by it.

However, a global model does not address spatial non-stationarity, which is the variation of the relation across space (citation on spatial non-stationarity). In order to capture the potential spatial variations, a two-step approach was carried out. The first step was checking spatial autocorrelation. This was achieved by testing the OLS residuals' global Moran's I, which is a (citation on moran's I). The global Moran's I of the OLS residuals was 0.37 (p value < 0.01) which indicated a spatial dependence. **Figure 4** shows the distribution of the residuals, from which a clear spatial pattern can be observed.

Since the existence of spatial autocorrelation, the global model was no longer effective in representing the relationship between the dependent variable and the predictors. As a result, the second step was establishing a local model, namely Geographically Weighted Regression (GWR) model. GWR is a local regression model that captures process's spatial heterogeneity (citation on GWR). GWR results contain estimated local parameters of each predictors for every regression point, which makes studying the relationships between different predictors and the dependent variable in different places much easier.

Building a GWR model requires determining a kernel which defines the 'local'. In this case, an adaptive gaussian kernel was selected. A gaussian kernel was used because the data is continuous across space, so it makes more sense to give closer data point higher weight.

Results

Conclusion