UCL DEPARTMENT OF GEOGRAPHY

YEAR 2016-17

EXAM CANDIDATE ID:	**QRSH1**
MODULE CODE:	**GEOGG125**
MODULE NAME:	**Principals of Spatial Analysis**
COURSE PAPER TITLE:	**Exploratory Spatial Data Analysis**
WORD COUNT:	**1995**

Are you registered as dyslexic with UCL Student Disability Services (SDS) and been given labels to 'flag' your written work

~~YES~~ / NO ***(please delete as applicable)***

Exploratory Spatial Data Analysis

Introduction

This report explores the spatial distribution of house prices across London using the Land Registry's Price Paid data (Land Registry, 2015). The analysis has been conducted at a ward scale and uses supporting demographic profile data as published by the London Data Store (London Datastore, 2015).

The aim of this report is to explore variance of house prices across London and investigate any secondary factors influencing their distribution through regression modelling. Analysis of the spatial distribution of house prices is approached using multiple correlation techniques. Kernel density estimation is used to analyse the density of house sale transactions and identify the regions with the highest turnover of house ownership. Spatial distribution can also be analysed by quantifying the similarities between adjacent wards via autocorellation. Spatial clustering will also be considered.

Demographic factors contributing to the mean house price will be investigated through both linear and geographically weighted regression modelling, with the aim of analysing the spatial influence of these contributing independent variables on the fit of the model. The indicators selected for analysis in conjunction with house price data are: median household income (2012/2013), crime rate per thousand population (2014/2015), average Public Transport Accessibility (PTA) score, and the percentage of the population from BAME backgrounds. These indicators were selected through prior correlation analysis between profile variables.

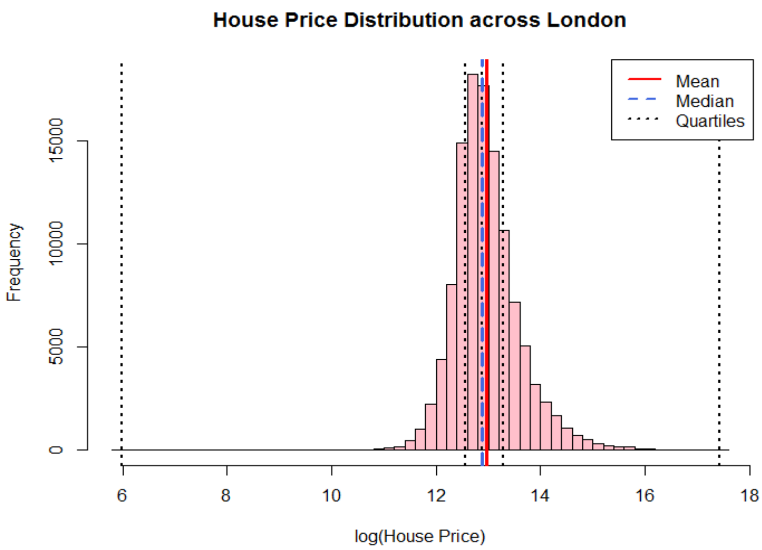
This analysis has been conducted using R.

Data Description

The house price data analysed in this report was obtained from the Land Registry (Land Registry, 2015) and comprises 115,379 records of house sales within the Greater London area during 2015. Each record is represented as point data and contains the sale price, dwelling type, address, postcode, and borough in which the building is located. The coordinates of the houses have been calculated from postcodes, and so represent the postcode location as opposed to the true street address. The statistical characteristics of this dataset are detailed and visualised in Figure 1.

The minimum and maximum recorded prices for houses sold during 2015 are £400 and £36,750,000 respectively. It is unlikely that the two houses recorded as being sold for £400 is correct as their estimated values listed by Zoopla are in the region of £400,000 (Zoopla, 2017) (Zoopla, 2017b). It is possible that these are errors in the dataset. Further investigation of other uncharacteristically low sale prices reveals a significant number of additional errors in the recording of house prices, however since this dataset is the officially published list of house sale prices for 2015, for the purposes of this analysis it is assumed that the figures are correct.

The distribution of house prices over London shows a log-normal distribution with a very small interquartile range of £300,000. The mean and median house prices are £552,600 and £395,000 respectively; the mean is greater than the median, therefore implying a positive skewness. This is confirmed through calculation of skewness, which indicates extreme positive skewness of 12.7. This is likely due to the small number of disproportionately expensive properties sold during 2015.



Statistic	Value
Minimum	400
1st Quartile	285000
Median	395000
Mean	552600
3rd Quartile	585000
Max	36750000
SD	712441
Skewness	12.7

Figure 1: (left) Histogram showing distribution of house price data (log) across London for 2015, and (right) Statistical characteristics for the prices of houses (£) sold during 2015. Note that a minimum house price of £400 is unlikely and is a potential error in the dataset.

The ward information used in this report was obtained from the London Datastore (London Datastore, 2015), and comprises 64 demographic indicators over 658 wards. Borough names will be used to aid in the spatial orientation of the description of the results throughout this report. Figure 2 shows a map of the London wards coloured by borough.

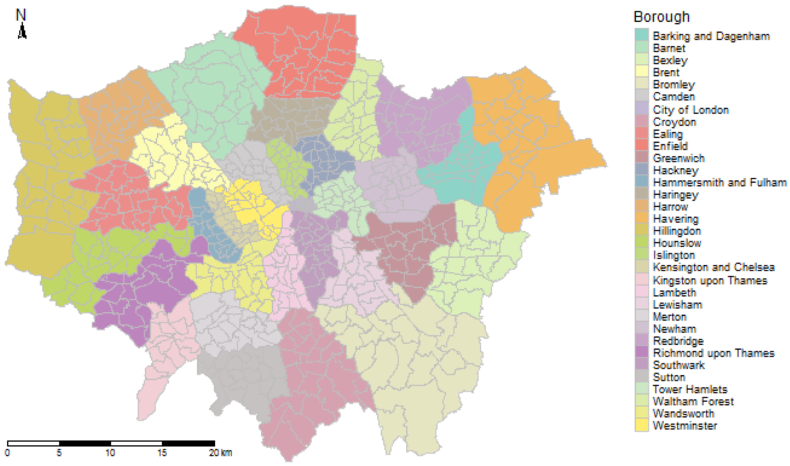


Figure 2: London wards coloured by borough.

Analysis

Initial spatial analysis of house price data on a ward and borough scale clearly shows clustering of high and low house prices across Greater London (Figure 3). The highest house prices are found in the north-western wards of central and inner London, stretching from the boroughs of Camden and Islington to the north, to Richmond and Kingston upon Thames to the south-west. Localised areas of high house prices are also found in select wards to the north-west of outer London within the boroughs of Harrow, Barnet, and Hillingdon. South eastern wards in Bromley also show relatively high mean house prices.

Areas of low house prices are found to the east of London across a large number of wards, and to the extreme west of Greater London across the boroughs of Hillingdon and Hounslow. Wards in north Croydon and Waltham Forest also have notably low mean house prices.

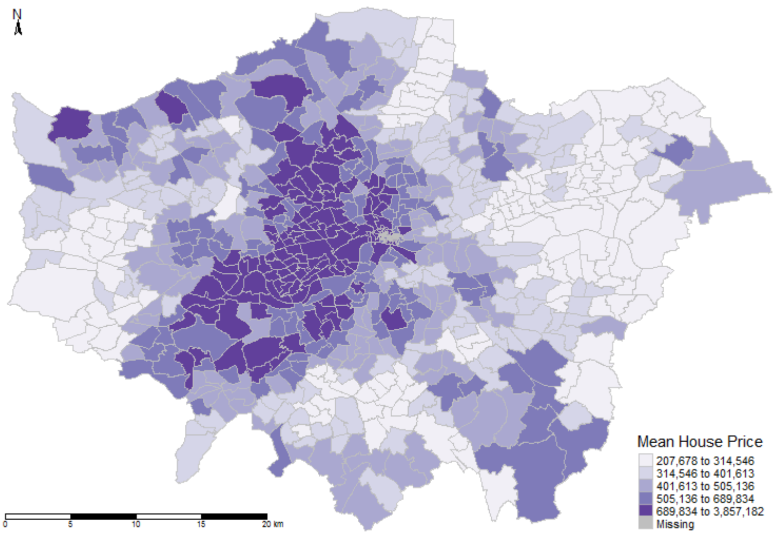


Figure 3: Spatial distribution of mean house price data across London wards for 2015. The highest house prices are found in the north-western wards of central and inner London, stretching from the boroughs of Camden and Islington to the north, to Richmond and Kingston upon Thames to the south-west

Density Estimation

Kernel density estimation was carried out on the house price point data to analyse the spatial distribution of house sales for 2015 (Figure 4). The highest density region (top 10%) of house sales occurred in west-central London, with the top 25% covering most of the remainder of Inner London, with the exception of Islington. The density of house sales correlates well with the PTA score, with the top 75% of transactions occurring in regions with a PTA score of 4 or greater. The density of house sales decreases, on average, with distance from central London. Zones where higher densities of house sales extend into the Outer London wards correlate visually with higher PTA scores, for example to the south in select wards in Sutton and Croydon, and to the east towards Barking and Dagenham, and Bexley.

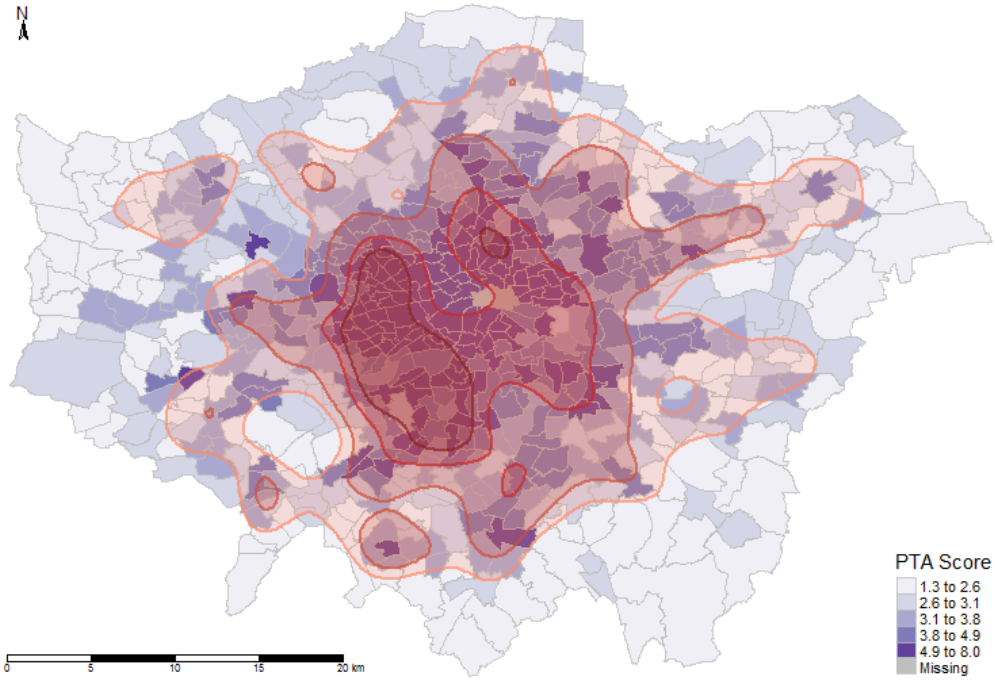


Figure 4: Kernel density estimation showing density of house sales in 2015. Contours are draw by density regions of 10%, 25%, 50%, and 75%. Wards have been coloured by PTA score, which correlates well with density of house sales. House sale hotspots appear to correlate with localised areas of higher PTA scores.

Spatial Autocorrelation

A Moran test was carried out to determine whether London house prices are spatially clustered. The Moran statistic for house price in London is 0.75, indicating that house prices are positively autocorrelated, and that the mean ward house prices are therefore related to that of their neighbours. The small P-value of < 2.2e-16 indicates that the Moran statistic is statistically significant, and that there is a negligible probability that the house prices in London are not autocorrelated (ESRI, 2016).

A local Moran test was carried out to explore the variation in house price autocorrelation; for London, the local Moran values range from -0.33 to 31.55 (Figure 5). The negative Moran statistics indicate minor negative spatial autocorrelation, whilst the positive values indicate wards where the mean house price is moderately to strongly related to those of their neighbours. Usually the Moran statistic should not exceed 1.0, however, as mentioned previously the house price data is strongly skewed. This may have affected the Moran statistic calculation and resulted in calculated values of >1.0 (ESRI, 2016).

The wards showing strong positive autocorrelation in house prices are those located in west-central London in the boroughs of Westminster, Kensington and Chelsea, and Hammersmith and Fulham, and to the east of London in Barking and Dagenham, Newnham, and parts of Havering. Moran statistics of zero indicate random spatial correlation.

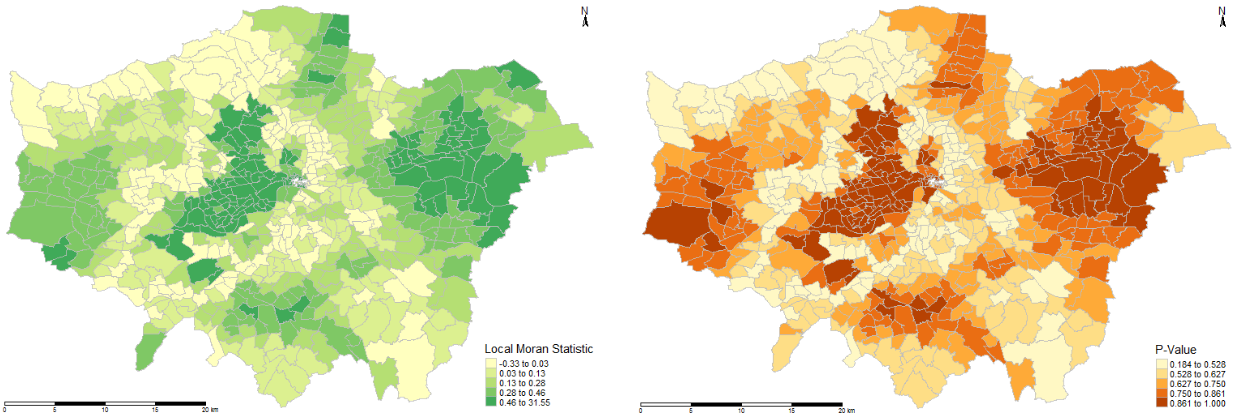


Figure 5: (left) Local Moran statistic (Ii), and (right) Local Moran P-value. Strongest clustering occurs in west-central London, and east London.

To investigate whether the observed spatial clustering is a result of high or low house prices, the Getis-Ord statistic (Gi*) was calculated. The Getis-Ord statistic defines neighbours based on proximity, and so a search radius of 5 km was used to ensure defined neighbours for each London ward. The resulting plot shows a strong positive clustering in the wards of Kensington and Chelsea, and Westminster (Figure 6), indicating that the house prices in these wards are both high, and spatially correlated to their adjacent wards (ESRI, 2016b). House prices across the remainder of London show weak negative correlation, indicating that both their mean house prices are low, and that their neighbours also show similarly low values. This could also be a result of the extreme high house price values found in the afore-mentioned areas, with every other value considered low in comparison.

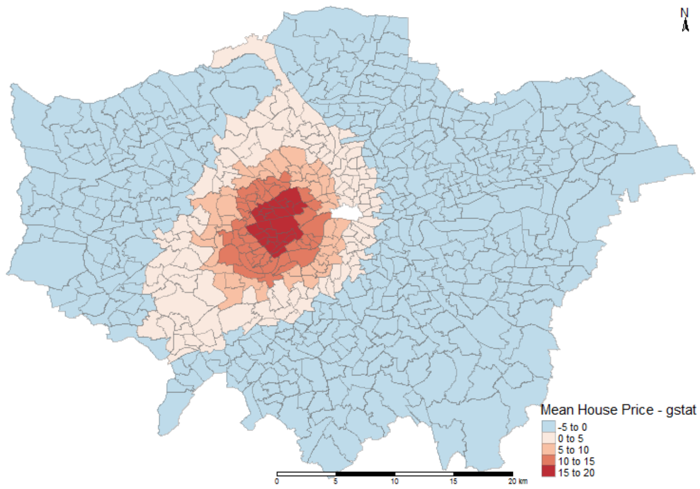


Figure 6: Getis-Ord statistic showing spatial clustering of autocorralated high and low house prices across London wards.

Regression Modelling

Having analysed the spatial characteristics of London house prices, the next stage is to explore what independent variables could contribute towards their values and distribution. Initially, a linear model was created to evaluate the global influence of mean household income, crime rate, PTA score, and the percentage of the population from BAME backgrounds. It was found that, combined, these four variables explained 72.5% of the variance in house prices. Each variable is statistically significant to the model to within 99.9%.

Geographically weighted regression (GWR) was performed to spatially analyse the fit of the model, and the explore the spatial variation in the influence of the predictor variables across the wards. GWR allows the coefficients to vary spatially in order to allow for the differing influence of each predictor on the dependent variable. Plotting the GWR R^2 values for each ward shows that model is a good estimator of house price over a significant proportion of London wards, with little-to-no clustering apparent (Figure 7). Wards showing low R^2 values appear to be localised and occur in small clusters, suggesting that house prices in these wards could be influenced by an additional, localised factor that has been unaccounted for in this model. Comparison of the predicted values of house price, as calculated using the model, against the real data shows significant similarities between the two.

The variation in the influence of the predictor variables on the model can be seen in Figure 8. Interestingly, PTA score appears to have least influence on the model in areas where the other three indicators show greatest influence, as indicated by the positive coefficients. The percentage of the ward population from BAME backgrounds appears to have the most significant influence on the model in areas when % BAME is very low. Crime rate also appears to have most significant influence in areas where either house price has not fully been explained by another variable, or where crime rate is low. Crime rate has little influence on house price in the "hot-spot" region.

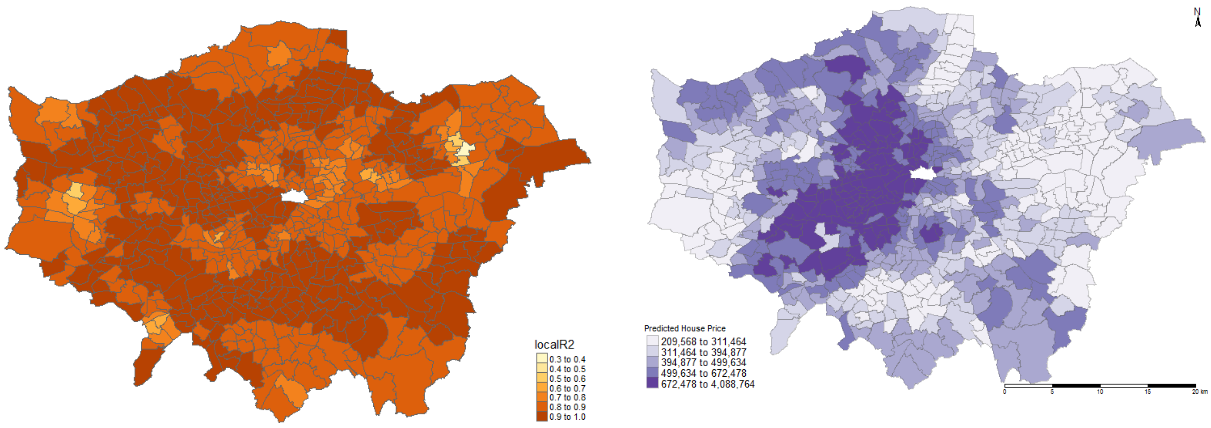


Figure 7: (left) R^2 value per ward as calculated using GWR, and (right) Predicted mean house price based on model.

Interpolation

Since the house price data for 2015 is not evenly distributed across London, interpolation can be used to create a continuous surface of the spatial variation in house price, using the known data to predict gaps. Figure 9 shows Inverse Distance Weighting (IDW) of the house price data along with the original data points used to interpolate the surface. Where there the house price density is greatest, the modelled surface can clearly be seen to match the general trend of the raw data points. However, where the density of known prices is low, the interpolated values appear to be greater than would be expected, given the closest known house prices. This is particularly evident to the north and east of London along the path of the Lee Valley and river Thames. There is no house price information in this area and the closest house price data indicates relatively low house prices in the region, yet IDW has predicted higher than expected house prices for these regions.

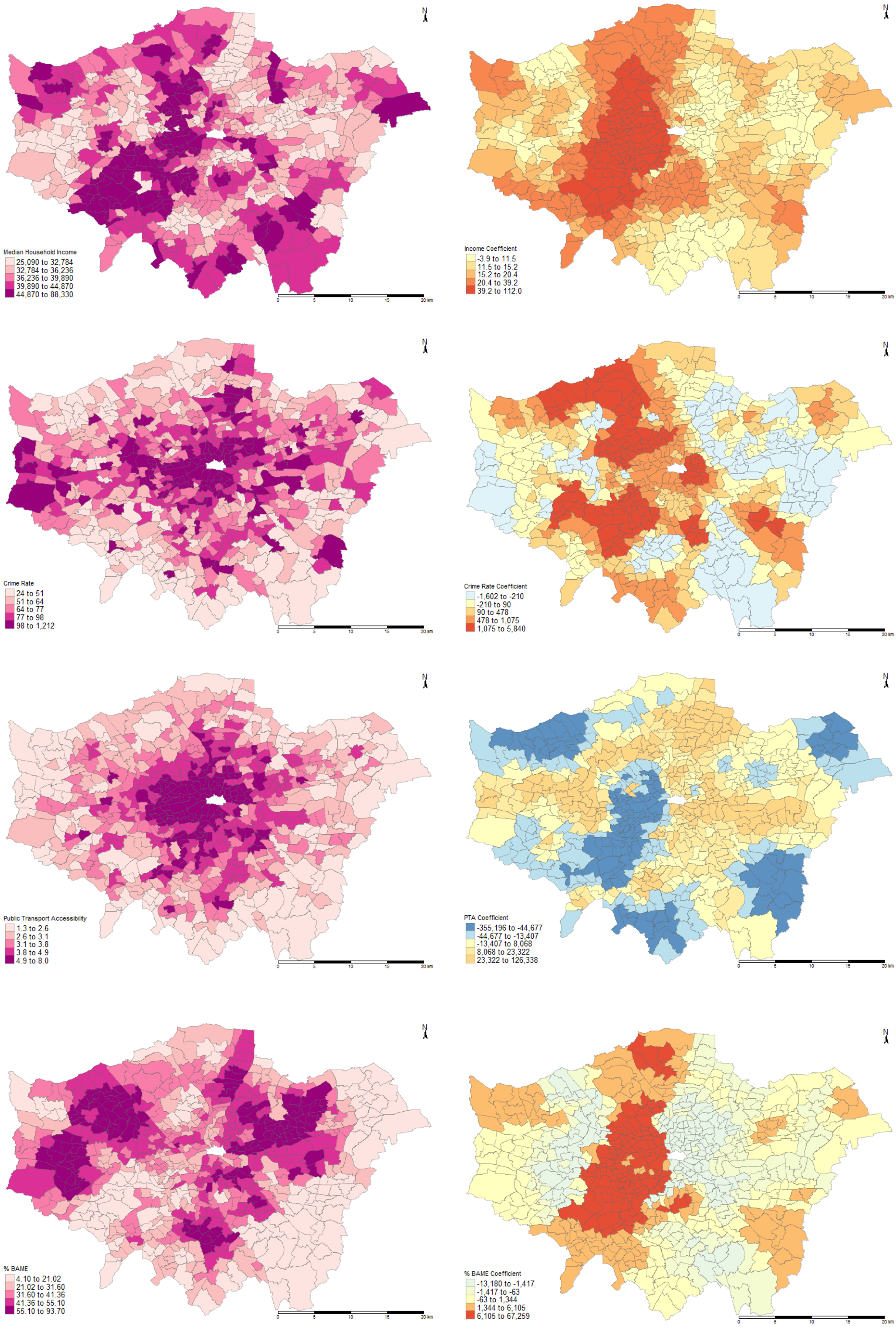


Figure 8: (left) Predictor variable spatial distribution, and (right) GWR coefficient. The highest spatial clustering of expensive housing is found in Kensington and Chelsea, despite the comparatively high crime rate found in this area. The prices in this area can be modelled solely based on median household income and % BAME.

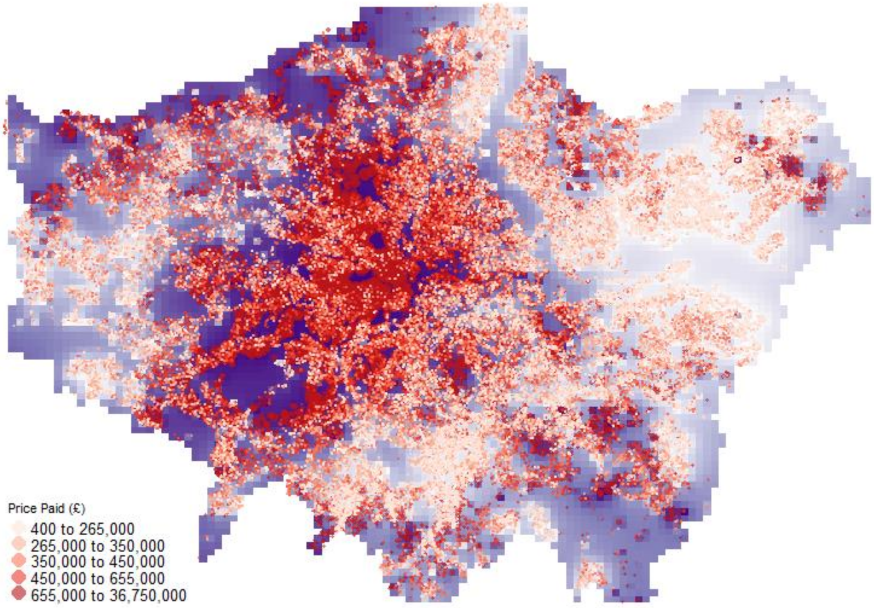


Figure 9: House price across London as estimated by IDW. Original house price data from which the surface was estimated is shown for comparison.

Discussion and Conclusion

Spatial analysis of London house price paid data for 2015 concludes that there are clear trends in the distribution of house prices. The highest priced houses are found in west-central London, and coincide with the areas with the largest PTA scores. The greatest influencing factors in the modelling of house prices in west-central London are income and %BAME, with the high crime rate in the area having little to no influence. It is possible the clustering of so many expensive houses occupied by such high-earning individuals has, it itself, created a crime hot-spot. This would explain the lack of influence that crime rate has on the model in this region.

The area with the highest density of house sales is also centred on this region, extending southwards. As house sale density also appears to correlated with PTA score, it can be concluded that house buyers prefer to purchase properties within easy access of public transport. Improvement of transport links across London could therefore lead to a rise in house prices across the affected areas. The perceived desirability of the region is also likely a contributing factor in the high-density of house sales in west-central London.

A limiting factors in this analysis is the uneven spread of house price data over London, and the lack of ward demographic values for the City of London. Inclusion of house price datasets for additional years could resolve this issue. Further investigation could include analysis of house sale density over time to identify any trends in the locations in which people buy houses over the years. This could be analysed in conjunction with change in demographics in order to better understand the factors influencing house sales in London over time, and therefore aid in the prediction of future house buying trends.

Bibliography

ESRI, 2016b. *How Hot Spot Analysis Works.* [Online]
Available at: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm
[Accessed 01 2017].

ESRI, 2016. *How Spatial Autocorellation Works.* [Online]
Available at: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm
[Accessed 01 2017].

Land Registry, 2015. *Land Registry Price Paid Data - Datasets.* [Online]
Available at: https://data.gov.uk/dataset/land-registry-monthly-price-paid-data
[Accessed 01 2017].

London Datastore, 2015. *London Datastore.* [Online]
Available at: https://data.london.gov.uk/dataset/london-borough-profiles
[Accessed 03 01 2017].

Zoopla, 2017b. *Property details for 161 Sumatra Road, London NW6 1PN.* [Online]
Available at: http://www.zoopla.co.uk/property/161-sumatra-road/london/nw6-1pn/17242755
[Accessed 01 2017].

Zoopla, 2017. *Property details for 14 Agamemnon Road, London NW6 1DY.* [Online]
Available at: http://www.zoopla.co.uk/property/14-agamemnon-road/london/nw6-1dy/17241276
[Accessed 01 2017].