Analysing Spatial Association of Geographical Data - Are there locally varying sociodemographic factors that drive the distribution of Anti-Social Behaviour rates in the London Borough of Croydon?

1. Introduction

Crime in London is an area of growing interest following continued policing and financial cuts in recent years (Sarjou, 2021). Studies have demonstrated that crime events tend to cluster in space but predicting specific behaviours is inherently difficult (Cahill & Mulligan, 2007). Nevertheless, understanding the local variation in sociodemographic factors that drive the distribution of crime events could allow more effective policy planning. This process is an important aspect of 'social crime prevention' (Silvestri, et al., 2009).

This study aims to contribute to the understanding of the relationship between the distribution of Anti-Social Behaviour (ASB) offences and locally varying sociodemographic factors in the London Borough of Croydon during 2019. Conclusions were made by interpreting the associations between ASB rates and various independent variables through non-spatial and spatial regression analyses.

2. Data and Methods

The methodology (Figure 1), utilised software GeoDa, RStudio, ArcMap and QGIS. Variables for analysis were selected based on a thorough literature review (Table 1). Throughout the study, the dependent variable was ASB offenses per 1000 inhabitants. The Index of Multiple Deprivation (IMD) score was the primary independent variable (see Appendix A), and additional variables were added during model re-specification.

An essential part of statistical studies prior to analysis is data exploration and log transformations were applied where appropriate (see Appendix B). Ordinary Least Squares (OLS) is a starting point for regression analysis (see Appendix C), and the presence/absence of spatial autocorrelation is displayed through the regression residuals, which were calculated using Moran's I (see Appendix D). The model was respecified (see Appendix E) and results interpreted to determine which model best represented the data.

Variable	Notes	Source
ASB rate	Anti-Social Behaviour offenses per 1000 inhabitants	Croydon Observatory
IMD score	Index of Multiple Deprivation Score	Croydon Observatory
Distance to nearest pub	Straight line distance to the nearest pub	Open Street Map
Dwelling type - flat/maisonette/apartment	Percentage of people living in a flat/maisonette/apartment	London Data Store
Lone parent households	Percentage of lone parent households	London Data Store
Social rented property	Percentage of people living in social rented property	London Data Store
Employment	Percentage of people in employment	London Data Store
No qualifications	Percentage of people with no qualifications	London Data Store
Median household income	Median annual household income	London Data Store
Median house price	Median house price	London Data Store

Table 1: Table of variables used in analysis including their data source; where blue is the dependent variable, green is the primary independent variable and orange are the additional independent variables.

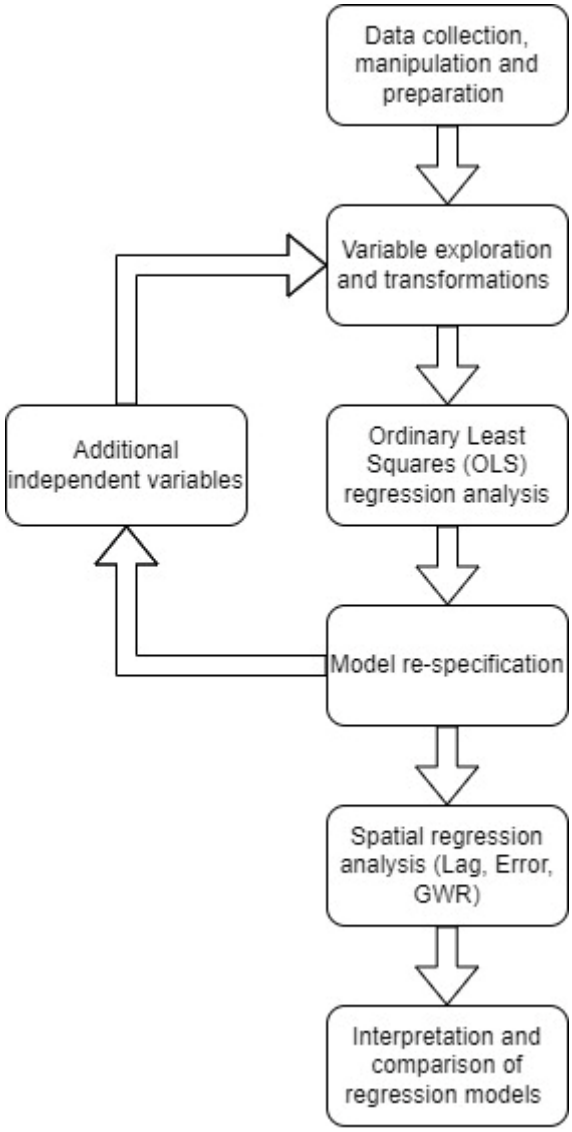

Figure 1: Visual methodology workflow created using draw.io.

3. Results and Discussion

i. Data Exploration

The London Borough of Croydon is made up of 219 LSOAs (Lower Super Output Areas) and a total of 8582 ASB offenses were recorded during 2019. Figure 2 displays higher levels of ASB rates in Croydon, Purley, Selhurst, South Norwood, Thornton Heath, and New Addington. The LSOA with the highest ASB rate (~127 offenses per 1,000 inhabitants) was Croydon 027B, specifically along central Croydon high street. The LSOA with the lowest rate (~1 offense per 1,000 inhabitants) was Croydon 042C, located in the south near Kenley Aerodrome.

High scores of IMD are visible in Croydon, particularly in the west and north of the town, Thornton Heath and New Addington. Conversely, areas of low IMD are found in the south and southwest of the borough. The LSOA with the highest IMD score (58.17) was Croydon 015D, located in northwest Croydon around Croydon University hospital whilst the lowest level of IMD (3.77) was Croydon 035A, located east of Sanderstead railway station, in the south-central area of the Borough.

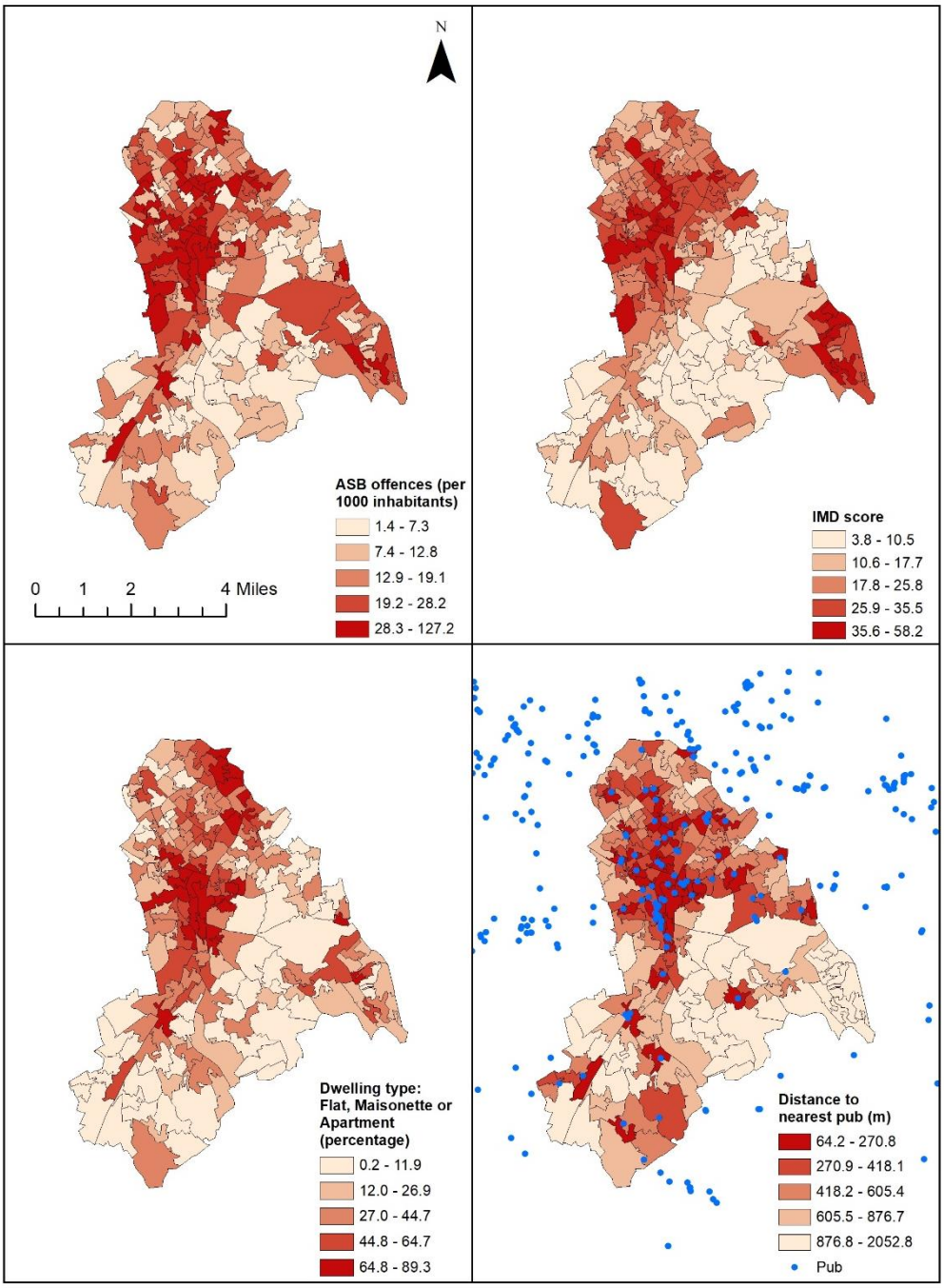

Figure 2: Exploratory variable maps created using ArcMap. ASB rate (top left), IMD score (top right), percentage of flats/maisonettes/apartments (bottom left), straight line distance (m) to the nearest pub (bottom right).

Figure 3 displays a positive correlation between the variables, with the ASB rate generally increasing with IMD score. Regression analysis can check the extent to which IMD can be used to predict ASB rates (Anselin, et al., 2004).

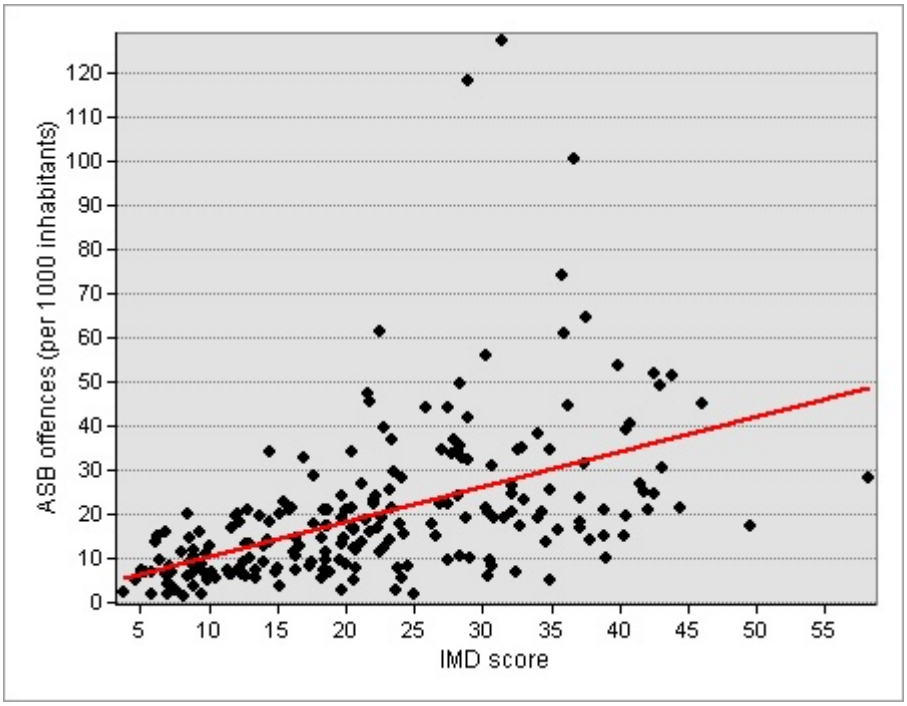

Figure 3: Scatter plot displaying IMD score and ASB offences created using ArcMap.

ii. Ordinary Least Squares (OLS)

OLS analysis produced a R-Squared value of 0.35 (Table 2), explaining 35% variance in the rate of ASB offenses in the study area. The coefficients for the intercept and slope display information on the least-squares estimates for the fitted line. The y-intercept (constant coefficient) of 0.75 and the slope (IMD coefficient) of 0.02, suggests a slight positive relationship between the variables. The statistically significant p-value indicates IMD will be reliable in predicting values of the dependent variable, using the following formula:

(Log ASB rate) = 0.75 + (0.02 * IMD score)

The multicollinearity condition number of 4.18 is well below the required threshold. Although a histogram of the residuals displays a close to normal distribution, a statistically significant Jarque-Bera value indicates a non-normal distribution of the error term. Jarque-Bera can also indicate missing explanatory variables from the model; unsurprising given the moderate R-Squared score. Notably, Jarqu-Bera results become less influential during interpretation as the sample size grows (Anselin, et al., 2004). As a result, the model is biased and likely missing explanatory variables. Conversely, the results from the Koenker test are not significant, suggesting residuals do not deviate from a normal distribution. Finally, the statistically insignificant Breusch-Pagan test, suggests heteroskedasticity is not present and the residuals are distributed with equal variance (Figure 4).

Mapping the residuals suggests a possible relationship with the LSOA regions (Figure 5). As a result, the estimators will be biased and therefore less significant than we might have originally thought. Mapping spatial patterns of the residuals can be misleading and need formal diagnostic (Anselin, et al., 2004). Running global and local spatial autocorrelation of the residuals will determine how the model can be improved.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ASB (Constant)	0.7506	0.0424	17.6891	0.0000
IMD	0.0189	0.0017	10.9014	0.0000

Model results	Value
R-Squared	0.3539
Adj. R-Squared	0.3509
Akaike Info Criterion	72.2521

Diagnostic tests	Value	Probability
Multicollinearity	4.1822	NA
Jarque-Bera	7.5293	0.0232
Breusch-Pagan	0.3191	0.5722
Koenker-Bassett	0.2529	0.6150

Table 2: OLS results using ASB rate as the dependent variable and IMD score as the independent variable. Calculated using GeoDa.

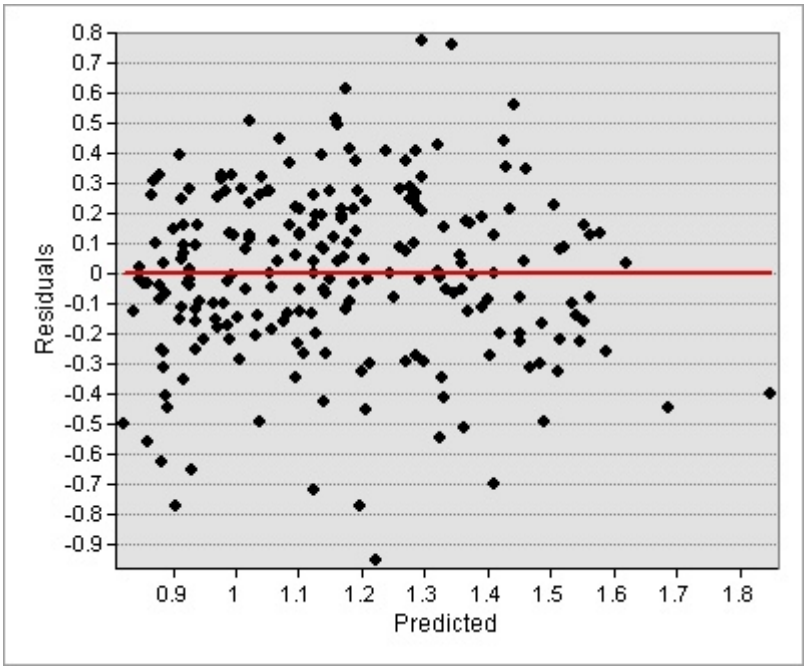

Figure 4: Scatter plot of OLS predicted and residuals values, providing a visual diagnostic of heteroskedasticity. Created using ArcMap.

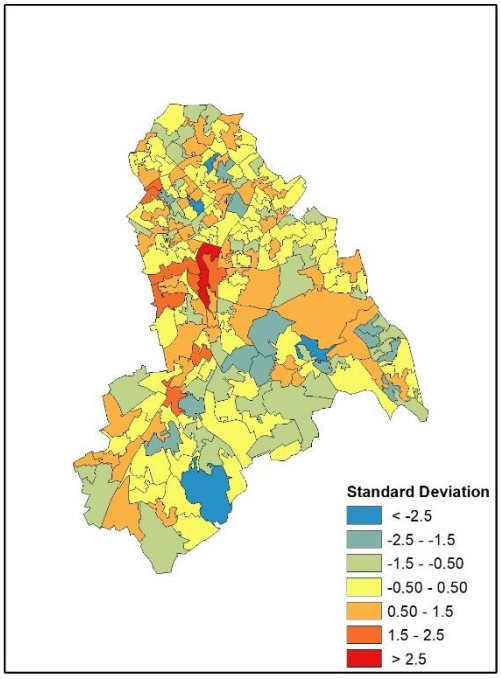

Figure 5: Map of OLS residuals (with one independent variable), created using ArcMap.

iii. Spatial Autocorrelation of Residuals

Although the p-value is significant, a Moran I's value of 0.069 (Table 3) does not suggest a strong spatial autocorrelation of the residuals. However, Local Indicators of Spatial Associations (LISA) displays clusters, where the residuals are predicting either too high, or too low, noticeably around Croydon town centre (Figure 6). Although a global test finds no significant deviation from randomness, a local test proves useful in uncovering isolated hotspots and coldspots (Rogerson, 2010). These results suggest regression residuals are influenced by geography and therefore must be considered to produce a reliable model.

Test	MI Score	Value	Probability
Moran's I	0.0689	1.8433	0.0453
Lagrange Multiplier (lag)	NA	6.7363	0.0095
Robust LM (lag)	NA	5.3500	0.0207
Lagrange Multiplier (error)	NA	2.6784	0.1017
Robust LM (error)	NA	1.2921	0.2557
Lagrange Multiplier (SARMA)	NA	8.0284	0.0181

Table 3: Diagnostic for spatial dependence based on OLS results using only IMD score as the independent variable.

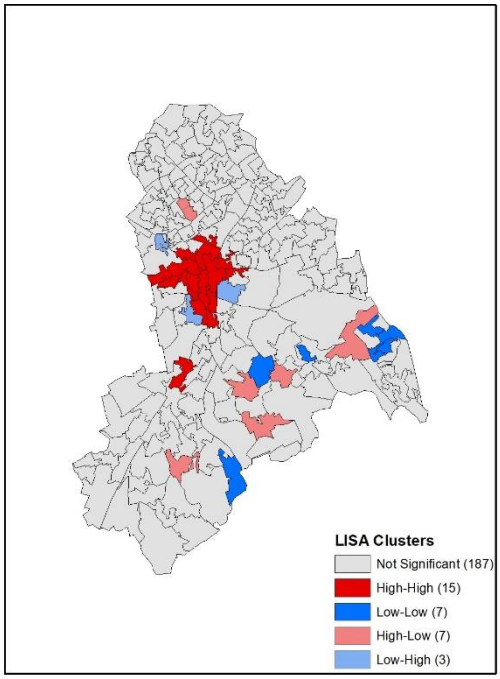

Figure 6: LISA cluster map of OLS residuals (with one independent variable) using Queens 1st order contiguity, where p < 0.05. Calculated using GeoDa and displayed in ArcMap.

iv. Re-specification of Regression Model - Additional Independent Variables

The use of census data can be conflicting. In the case of IMD, it draws on multiple sociodemographic factors and therefore the addition of the census variables does not necessarily provide a separate dimension of explanation regarding their relation (Anselin, 2017). The addition of eight new variables (Table 1) results in a multicollinarity score of 120.48, suggesting a high level of data redundancy. Following exploratory regression analysis, two additional variables were statistically significant, distance to the nearest pub and flat/maisonette/apartment dwelling type percentage. Correlation coefficient results between -0.23 and 0.57 imply the three independent variables are not highly correlated and are therefore suitable for multivariate regression analysis (Figure 7).

A re-specified regression model included the three statistically significant variables and produced an adjusted R-Squared score of 0.46 (Table 4), a 0.11 improvement from the previous model. AIC (Akaike Info Criterion) decreased from 72.3 to 45.6. Based on these values alone, this model is an improvement on the first iteration of OLS. The multicollinearity value of 27.7 is relatively high, although still below the value of 30 threshold suggested by Luc Anselin (2004). Both Breusch-Pagan and Koenker tests are not statistically significant. This suggests heteroscedasticity is not present and therefore the residuals are distributed with equal variance. A Moran's I value of -0.016 and a non-significant probability implies spatial autocorrelation in the residuals is not present, as we cannot reject the null hypothesis of spatial randomness. Furthermore, the LISA cluster map of the residuals sees a reduction in both the number of hotspots (15 to 8) and cold spots (3 to 2) (Figure 8). A statistically significant Jarque-Bera test suggests the residuals are not normally distributed, although a histogram reveals a very minor negative skew. ASB rate prediction using this model would apply the following formula:

(Log ASB rate) = 1.52 + (0.02 * IMD score) + (-0.31 * Log Distance to nearest pub) + (0.13 * Log Dwelling Type F/M/A)

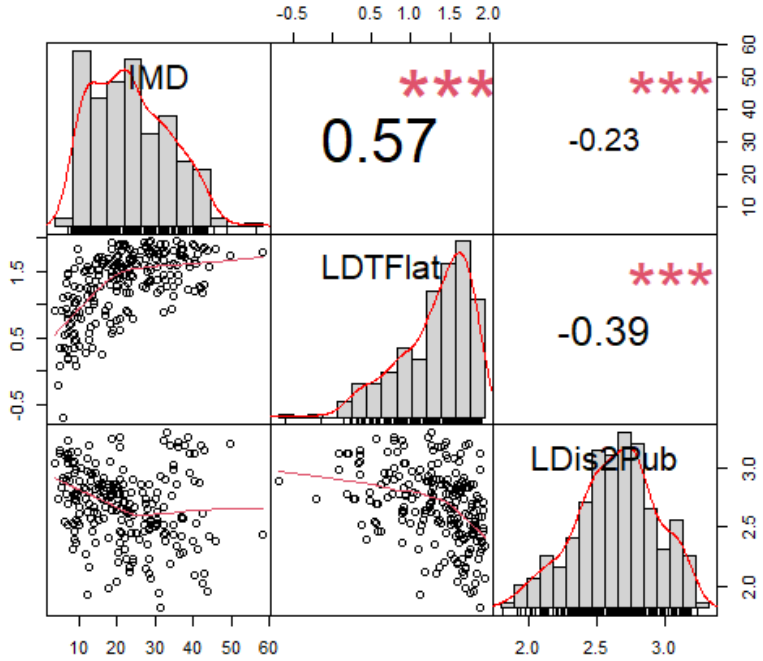

Figure 7: Scatter plot matrix displaying the linearity and distribution of the three statistically significant independent variables: IMD (Index of Multiple Deprivation), LDTFlat (Log Dwelling type - flat/maisonette/apartment), LDis2Pub (Log Distance to nearest pub). Created in RStudio using the Chart Correlation function.

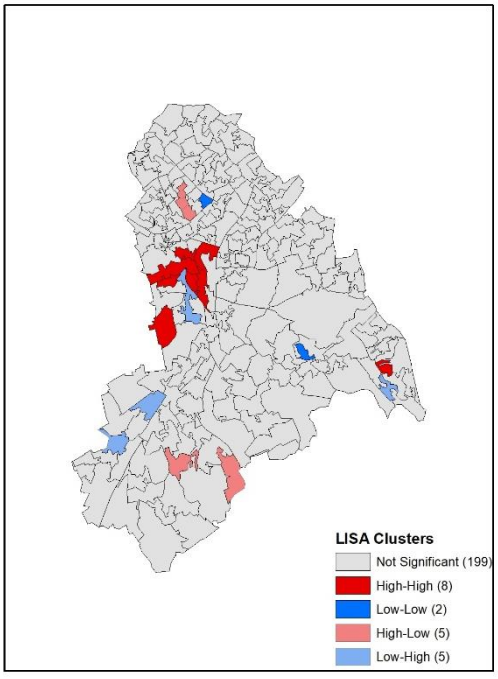

Figure 8: LISA cluster map of OLS residuals (with three independent variables) using Queens 1st order contiguity, where $p < 0.05$. Calculated using GeoDa and displayed in ArcMap.

Variable	Coefficient	Std. Error	t-Statistic	Probability
ASB (Constant)	1.5251	0.1885	8.0920	0.0000
IMD	0.0137	0.0019	7.1839	0.0000
Distance to nearest pub	-0.3083	0.0603	-5.1092	0.0000
Dwelling type - flat/maisonette/apartment	0.1259	0.0453	2.7762	0.0060

Model results	Value
R-Sqaured	0.4721
Adj. R-Squared	0.4647
Akaike Info Criterion	32.0000

Diagnostic tests	Value	Probability
Multicollinearity	27.6656	NA
Jarque-Bera	31.4853	0.0000
Breusch-Pagan	2.2515	0.5219
Koenker-Bassett	1.3585	0.7153

Test	MI Score	Value	Probability
Moran's I	-0.0156	-0.1235	0.9017
Lagrange Multiplier (lag)	NA	0.5902	0.4423
Robust LM (lag)	NA	3.1809	0.0745
Lagrange Multiplier (error)	NA	0.1375	0.7108
Robust LM (error)	NA	2.7281	0.0986
Lagrange Multiplier (SARMA)	NA	3.3183	0.1903

Table 4: OLS results using the three statistically significant independent variables: Index of Multiple Deprivation, Dwelling type - flat/maisonette/apartment and Distance to nearest pub. Calculated using Geoda.

v. Re-specification of Regression Model - Spatial Regression Analysis

a) Spatial Lag Model

Although the addition of independent variables increased the adjusted R-Squared value, this study firstly reverts to the original independent variable IMD, due to its need for spatial consideration. As per the methodology, LM Lag, which introduces a spatial lag of the dependent variable, was used due to the diagnostic results being statistically significant (Table 3); whereas LM Error is not.

When using all three independent variables, the diagnostic for spatial dependence did not suggest the use of a spatial model due to insignificant statistical results. As a means of analytical investigation both Lag and Error models were tested using these variables but the spatial autoregressive coefficients Rho and Lambda for each respective model were not statistically significant. Although producing higher R-Squared and lower AIC scores than the non-spatial OLS model, a Lag or Error spatial model would not be appropriate with this combination of variables. This is endorsed by Moran's I test results, which reveal spatial autocorrelation had almost no effect on the residuals (Table 4).

LM Lag model, using only IMD score as the independent variable produced an R-Squared value of 0.37 with the coefficients being statistically significant (Table 5). The spatial autoregressive coefficient is statistically significant, providing a clear indication that the results from the Lag model are preferred to the non-spatial model. This is further supported by a statistically significant Likelihood Ratio. A multicollinearity score of 4.10 suggests a low level of data redundancy. Additionally, the insignificant Breusch-Pagan test suggests heteroscedasticity is not present. The results imply that the spatial lag in the dependent variable improves the model, using the following formula:

(Log ASB rate) = 0.55 + (0.02 * IMD score) + 0.22

A reduced Moran's I value of -0.028 indicates the addition of a spatial component has reduced the global spatial autocorrelation of the residuals. Figure 9 displays a reduction in both hot spots and cold spots, indicating the model is less biased in comparison to the original model.

Variable	Coefficient	Std. Error	z-Value	Probability
W_LogASB	0.2211	0.0852	2.5942	0.0095
ASB (Constant)	0.5538	0.0872	6.3520	0.0000
IMD	0.0160	0.0020	8.1450	0.0000

Model results	Value
R-Squared	0.3782
Akaike Info Criterion	67.9245

Diagnostic tests	Value	Probability
Multicollinearity	4.1822	NA
Breusch-Pagan	0.0242	0.8764
Likelihood Ratio Test	6.3275	0.0119

Test	MI Score
Moran's I	-0.0280

Table 5: Spatial Lag model results using ASB rate as the dependent variable and IMD score as the independent variable. Calculated using GeoDa.

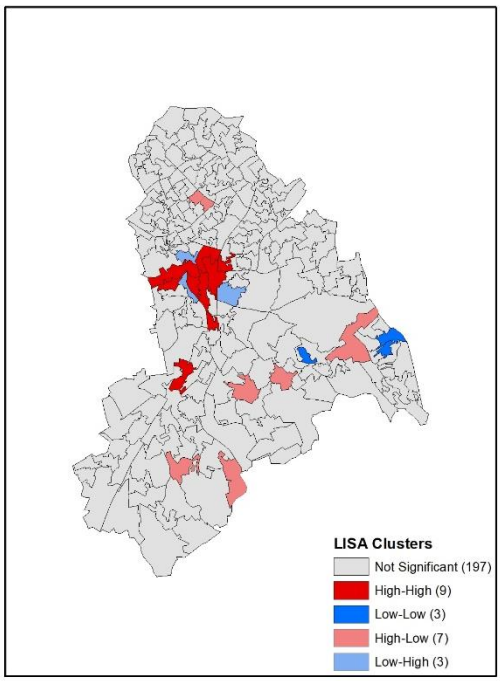

Figure 9: LISA cluster map of LM Lag residuals (with one independent variable) using Queens 1st order contiguity, where p < 0.05. Calculated using GeoDa and displayed in ArcMap.

b) Geographically Weighted Regression (GWR)

The R-Squared score of 0.41 from GWR using only IMD as the independent variable is higher than all previous models and the AIC score of 59.77 was also lower than the previous spatial model. A Moran's I value of -0.022 is also closer to 0 than the previous spatial model. This is consistent with previous literature suggesting GWR provides the best model when the relationship is spatially inconsistent (O'Sullivan & Unwin, 2010)

GWR was also run with the three independent variables used in the previous iterations of regression analysis. This produced an adjusted R-Squared value of 0.46 and an AIC value of 34.29 (Table 6), with a Moran's I value of -0.016 and LISA results almost identical to standard OLS with the same variables (Figure 10). Crucially, GWR generally strengthens the OLS findings as its formula is localised, depending on the area of interest. One of its main benefits is the ability to see variation in the model's performance, through local R-Squared across the study area. Additionally, it allows focus on areas where the model is not performing as expected (Cahill & Mulligan, 2007). GWR can be of particular importance to policy planners when attempting to account for local differences between communities not captured by standard measures and thus potential explanatory causes of crime. This may aid area-specific crime prevention interventions. Alternatively, GWR can be used to measure the success of an implemented intervention by determining areas where the mediation was more successful.

Local R-Squared maps (Figure 11) display goodness of fit across the study area, with higher value areas producing a more reliable model. Local R-Squared with only IMD as the independent variable shows a large variation in goodness of fit across the study area, with high values located in the central and west of the borough, providing useful information to policy planners. Conversely, although producing a higher R-Squared value, using three variables displays very little deviation in the value across the study area, enforcing previous findings that using the three statistically significant independent variables from multivariate regression removes the effects of spatial autocorrelation.

If the study were to increase in size, to cover multiple boroughs for instance, GWR would likely be the favourable model choice due to its ability to capture variation in relationships between variables of large datasets and the LSOA regions being relatively similar in size (Cahill & Mulligan, 2007).

Model results	**Value** (1 Var)	**Value** (3 Var)
R-Squared	0.4371	0.4721
Adj. R-Squared	0.4053	0.4647
Akaike Info Criterion	59.7699	34.2887

Diagnostic tests	**Value** (1 Var)	**Value** (3 Var)
Multicollinearity	4.1822	27.6656

Table 6: GWR results using one (left) and three (right) independent variables. Calculated using ArcMap.

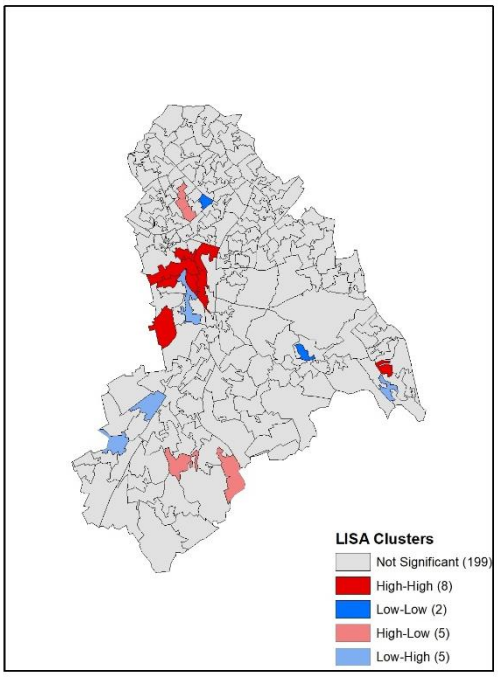

Figure 10: LISA cluster map of GWR residuals (with three independent variables) using Queens 1st order contiguity, where p < 0.05. Calculated using GeoDa and displayed in ArcMap.

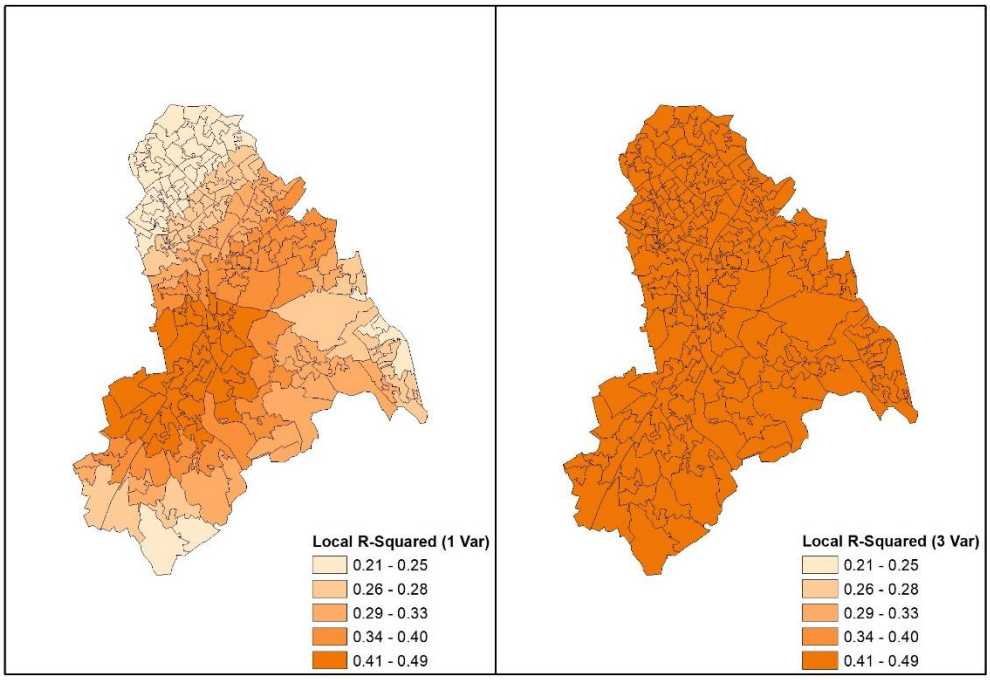

Figure 11: GWR local R-Squared maps using one (left) and three (right) variables with identical symbology classification. Created using ArcMap.

4. Conclusion

This study investigated whether there are locally varying sociodemographic factors that drive the distribution of ASB rates in the London Borough of Croydon at LSOA level. The analysis concludes there is a statistically significant relationship between IMD, distance to the nearest pub, flat/maisonette/apartment dwelling type percentage and ASB rates. Additionally, to improve understanding, results indicate that regression models should incorporate the effects of spatial autocorrelation, particularly at a local level.

The LM statistical test indicate that the Spatial Lag model was more appropriate than the Spatial Error model to consider the spatial effects of the relationship. GWR produced the better of the spatial models based on adjusted R-Squared and AIC scores as well as having a greater impact on cluster reduction of the residuals. Whilst retaining some local spatial autocorrelation, GWR is importantly tolerant of such autocorrelation and therefore provides more precise estimates. Although OLS and GWR produced similar statistical results, GWR generates the most favourable model, by reason of producing a truly local statistic.

These results can have various positive applications in aiding effective policy planning and mitigating future offenses. Moreover, it is important to continue to experiment with different variables to see how the relationship changes. Future studies may wish to analyse how these relationships change over time.

5. Figures

6. Tables

7. References

Anselin, L., 2005. *Exploring Spatial Data with GeoDa: a Workbook,* Urbana-Champaign, USA: Center for Spatially Integrated Social Science, Department of Geography, Universit of Illinois.

Anselin, L., 2017. *GeoDa Software: Review - Ordinary Least Squares and 2 Stage Least Squares.* [Online]
Available at: https://www.youtube.com/watch?v=FZGXswl63to
[Accessed 1 April 2022].

Anselin, L., Florax, R. & Rey, S. J., 2004. *Advances in Spatial Econometrics: Methodology, Tools and Applications.* 1st ed. Berlin: Springer Science & Business Media.

Bivand , R. S., Pebesma, E. & Gómez-Rubio, V., 2013. *Applied Spatial Data Analysis with R.* 2nd ed. New York: Springer Science+Business Media.

Cahill, M. & Mulligan, G., 2007. Using Geographically Weighted Regression to Explore Local Crime Patterns. *Social Science Computer Review,* II(23), pp. 174-193.

Croydon Observatory, 2019. *Index of Multiple Deprivation (IMD) Score - LSOA (2019).* [Online]
Available at: https://www.croydonobservatory.org/deprivation/map/
[Accessed 1 April 2022].

Data Police UK, 2019. *Data downloads.* [Online]
Available at: https://data.police.uk/data/
[Accessed 1 April 2022].

Fischer, M. M. & Getis, A., 2010. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications.* 1st ed. Berlin: Springer.

Fischer, M. M. & Wang, J., 2011. *Spatial Data Analysis: Models, Methods and Techniques.* 1st ed. s.l.:Springer Science & Business Media.

Fotheringham, A. S., Brunsdon, C. & Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* 1st ed. Chichester, England: Wiley.

London Datastore, 2020. *Statistical GIS Boundary Files for London.* [Online]
Available at: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london
[Accessed 1 April 2022].

O'Sullivan, D. & Unwin, D., 2010. *Geographic Information Analysis.* 2nd ed. New Jersey: Wiley.

Rogerson, P. A., 2010. *Statistical Methods for Geography.* 3rd ed. London: SAGE Publications, Limited.

Sarjou, A., 2021. Violent Crime in London: An Investigation using Geographically Weighted Regression. *arXiv.*

Silvestri, A., Oldfield, M., Squires, P. & Grimshaw, R., 2009. Young People, Knives and Guns. A Comprehensive Review, Analysis and Critique of Gun and Knife Crime Strategies. *Centre for Cime and Justice Studies.*

Tian, W., Jitian , S. & Zhanyong , L., 2014. Spatial Regression Analysis of Domestic Energy in Urban Areas. *Energy,* Issue 76, pp. 629-640.

Waldo, T., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography ,* 46(1), pp. 234-240.

8. Appendix

Appendix A - *Data Preparation and Manipulation*

Crime data, which is recorded monthly and maintained through a rigorous quality control process, was downloaded from the Data Police website (Data Police UK, 2019). The data was filtered by crime type, location and year to only include ASB offenses in the London Borough of Croydon during 2019, chosen due to it being the last full year not influenced by the Coronavirus pandemic. Offenses include personal, environmental and nuisance ASB with rates being maintained by the Croydon Observatory. Census data, which included the Lower Super Output Area (LSOA) spatial boundaries was downloaded from the London Data Store (London Datastore, 2020). Finally, Index of Multiple Deprivation (IMD) data was downloaded from the Croydon Observatory (Croydon Observatory, 2019). Data was manipulated and merged using Excel, R and GeoDa.

The variables used in this study were selected based on a thorough literature review. The primary independent variable used in the first iteration of OLS analysis is the IMD Score, where a higher score indicates a higher level of deprivation, the official measure of deprivation in England. It is comprised of seven distinct domains of deprivation - income, employment, health deprivation and disability, education and skills training, crime, barriers to housing and services, and living environment, these are combined to give an overall relative measure of deprivation. A variety of additional independent variables were included during the model re-specification (Table 1).

The straight line distance to the nearest pub was calculated using the 'Distance to Nearest Hub' tool in QGIS, using the centroid of the LSOA regions as the origin layer and the pub locations as the destination layer. Pub locations were obtained using the 'Quick OSM' (Open Street Map) plugin within QGIS.

Appendix B - *Data Transformations*

Prior to analysis, an essential part of statistical studies is data exploration. When carrying out hypothesis testing, generally the values of the variables are assumed to be normally distributed (O'Sullivan & Unwin, 2010). When variables are highly skewed a log transformation can move a distribution closer to normal. ASB rates had a highly positive skew, therefore a log transformation was applied using the GeoDa field calculator to transform its distribution closer to normal (Figure 12). Log transformations were also applied to other variables that displayed a highly skewed distribution. Additionally, a scatterplot matrix of all the independent variables identified their linearity. Correlation coefficients are useful in understanding how pairs of variables are related. In regression analysis, it is important that no 2 variables are highly correlated.

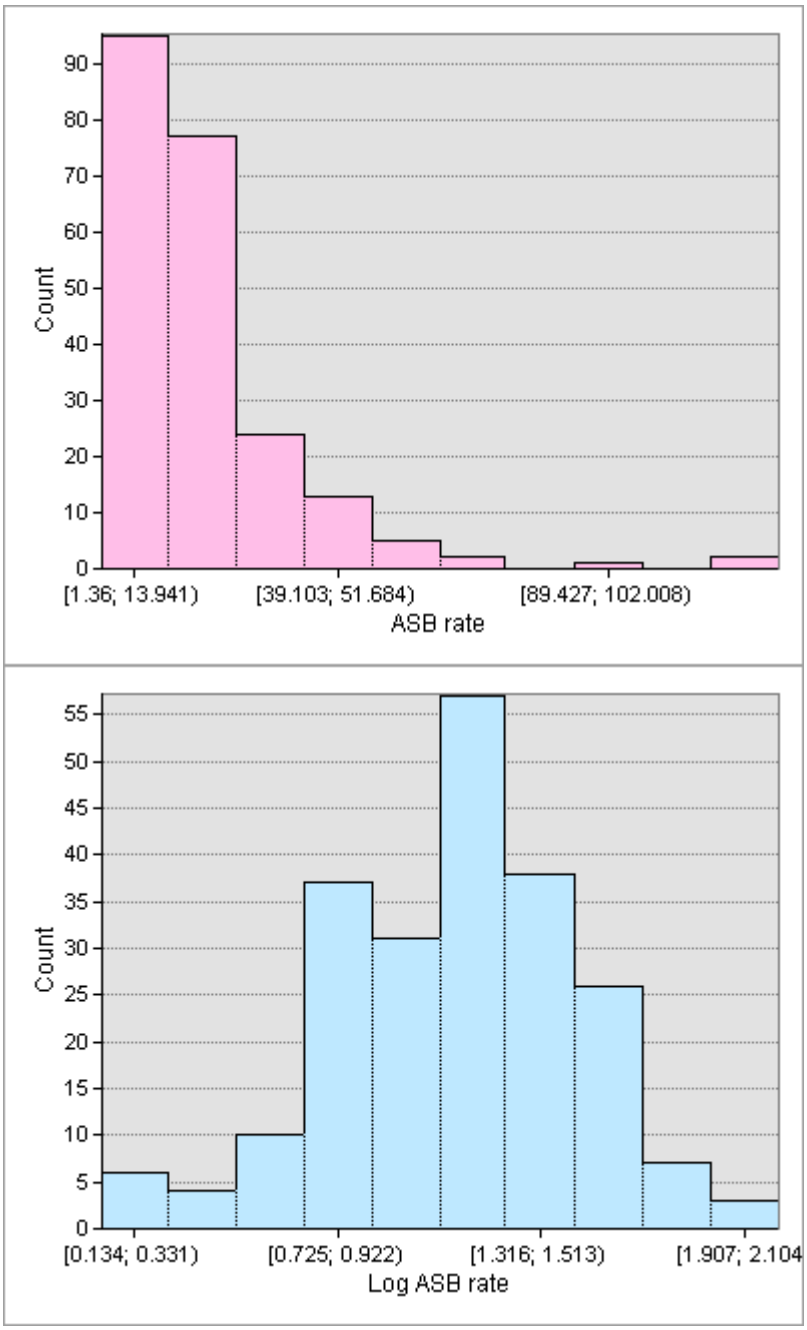

Figure 12: Distribution of Anti-Social Behaviour rate before (pink) and after (blue) applying a log transformation. Created using GeoDa and displayed in ArcMap.

Appendix C - *Ordinary Least Squares*

Ordinary Least Squares (OLS) is a starting point of regression analysis. Whereas correlation is used to measure the strength of the linear association between variables, regression analysis refers to the relationship between a dependent variable and independent variables (Rogerson, 2010). An important outcome of regression analysis is the production of one equation that allows the prediction of values of Y from values of X on a global level; whereby the relationship between the variables is assumed to apply with the same coefficient at all locations. The model is calculated by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line (O'Sullivan & Unwin, 2010). An important step in the evaluation of OLS models is to examine the residuals for evidence of trends related to the variables in the model. The residuals are the differences between the observed values of Y for each case minus the predicted or expected value of Y. When there is a discernible trend, the model is said to be misspecified (O'Sullivan & Unwin, 2010).

As per previous regression analyses, a typical significant level of 0.05 was used throughout this study (Rogerson, 2010). This confidence interval defines a range of values that we can be 95% certain contains the mean slope of the regression line. Significant p-values in the independent variables, suggest reliability at predicting values of the dependent variable.

To determine whether OLS produces a sound model, it must pass multiple statistical checks (Table 7) which are calculated and provided as part of the regression results. Identifying which model is most helpful involves comparing these checks between results.

Significance of the regression coefficients	
T-test and p-value	• The standard error and the t-value indicate how the p values were calculated. • OLS calculates a coefficient for each variable. If the p-value is less than the desirable threshold (0.05) the variable is statistically significant to the model at a 95% confidence level. • The p-value determines whether the estimates for the y intercept and the slope are equal to 0 or not (if they are equal to 0, they do not have much use in the model). • The sign of the coefficient determines whether the relationship is positive or negative. • A linear regression equation is created using these coefficients: dependent variable value = y intercept value + (slope * independent variable value)
Entire model	
R-Squared	• R-Squared value measures a model's strength. • Concerns variation in the dependent variables accounted for by the independent variables. Ranges from 0 (no relation) to 1 (independent variables explain all variation). • Adjusted R-Squared, which accounts for an increasing R-Squared value with the increasing number of independent variables in a model, is used to compare strength between different models.
AIC	• AIC (Akaike Information Criterion) is useful when comparing models. • Indicates how close the model is to reality. A lower value indicates a better fitting model.
Multi collinearity	
MCN (Multi Collinearity Number)	• Multi collinearity indicates the extent to which explanatory variables add an independent dimension to the analysis. •If MCN is greater than 30 it indicates significant multicollinearity.
VIF (Variance Inflation Factor)	• VIF is a measure of the amount of multicollinearity in a set of multiple regression variables. • A value larger than 7.5 means there could be redundancy among variables.
Normality of residuals	
Jarque-Bera	• Tests the normality of the residual distribution. • Test combines the effects of both skewness and Kurtosis. A low probability test score indicates non-normal distribution of the error term.

	• If statistically significant the model is biased and predictions cannot be fully trusted. • Can also indicate explanatory variables are missing from the model. • Particularly important with a small sample size.
Breusch-Pagan (Heteroskedasticity test)	• Used to determine whether heteroscedasticity is present in a regression model • Tests the squares of the explanatory variables to determine if there is nonconstant variance in the errors. • When statistically significant, heteroscedasticity is present (the residuals are not distributed with equal variance).
Koenker (Heteroskedasticity tests)	• Similar to the Breusch-Pagan test, except the residuals are studentized, meaning they are made robust to outliers/non-normality. • A statistically significant Koenker statistic indicates a nonstationary relationship in your model. • Can test the normality of residuals. If it is significant, the residuals deviate from a normal distribution.

Table 7: Table of statistical checks for properly specified models in regression analysis.

Appendix D - *Spatial Autocorrelation*

Spatial autocorrelation refers to data from locations near one another in space are more likely to be similar than data from locations remote from one another (O'Sullivan & Unwin, 2010). Moreover, as per Waldo Tober's (1970) first law of geography is that "Everything is related to everything else, but near things are more related than distant things"; therefore, we would expect most geographic phenomena to exert spatial autocorrelation. For example, in population data, this is often the case as persons of similar characteristics tend to reside in similar neighbourhoods.

Spatial autocorrelation in regression, calculated using GeoDa, looked at whether the error term is correlated with geography. If spatial autocorrelation is present, it breaks the assumption that the value of residuals is random. Therefore, one will likely need to re-specify the model and adjust for the geography in the relationships.

The most widely used method for checking the presence/absence of spatial autocorrelation is Moran's I, which describes the degree of spatial concentration or dispersion for the variables of interest (Tian, et al., 2014). Moran's I ranges from −1 (dispersed) to 1 (clustered). When the Moran's I for a specific variable is positive, the large values for the variable are surrounded by other large values. When the Moran's I for a variable is negative, the large values are surrounded by small values. Thus, a positive spatial autocorrelation implies a spatial clustering for a variable, whereas a negative spatial autocorrelation suggests a spatial dispersion (Fischer & Wang, 2011).

If the errors are correlated with geography, the estimators of the analysis will be biased. If spatial autocorrelation is ignored, then all inference (including R-Squared) based on the classical regression model will be unsound. To check spatial autocorrelation, Moran's I and spatial dependence of the error term were calculated to a significance level of 5% ($p < 0.05$).

Spatial weights matrices, representing the geometric relationship between regions, are required to calculate the Moran's I for a given variable. There are numerous different ways to define the spatial weights matrix and the choice is an important analysis step (Bivand , et al., 2013). Queen's case considers neighbours based on adjacency and includes both areas that share a boundary edge or corner vertex as a neighbour. Alternatively, contiguity between polygons may be ignored and a measure of distance used instead. It's advisable to work with simple adjacency-based approaches in the exploratory phase of analysis, particularly when the underlying processes are not well understood, hence 1st order Queens contiguity was used in this analysis (O'Sullivan & Unwin, 2010). Additionally, even though the LSOA regions aren't entirely uniform, Queen's contiguity also proves useful in urban areas where spatial regions are a similar size and have multiple neighbours (Anselin, et al., 2004). Using Queen's contiguity, 79.4% of the observations have between 4-7 neighbours. Other spatial weights such as Rooks and distance matrices were checked through a distance correlogram and considered in the explanatory phase but did not produce much variation in the results.

Mis-specified models can be improved by 1) removing and adding variables. In a geographic setting, when we observe spatial structure in a model's residuals, 2) spatial dependence of the variables should be included, or the model should vary spatially. This study uses both adjustments to improve the model. Notably, mis-specification doesn't necessarily mean the model is of no use (O'Sullivan & Unwin, 2010).

Appendix E - *Model Re-specification*

Additional Variables

In an initial attempt to develop the model, additional variables were added and OLS rerun. Adding additional variables to explain more of the variation in the dependent variable aims to eliminate any residual spatial autocorrelation. Furthermore, adding additional variables into a model helps to control spurious correlations as correlation does not equal causation (Anselin, 2017).

Spatial Regression Analysis

When residuals are influenced by the geography of a study area, a spatial regression model will need to be implemented. This study considers three spatial regression analyses: LM (Lagrange Multiplier) Lag, LM Error and Geographically Weighted Regression (GWR). Determining which of these models is most suitable is a key process in spatial analysis (Fischer & Getis, 2010).

The LM test, which requires the results from OLS and a spatial weights matrix, produces four test statistics in the analysis diagnostic output: LM Lag, LM Error, Robust LM Error, and Robust LM Lag. The LM Error and robust LM error statistics are used to test whether the spatial error models are best suited. Conversely, the LM lag and robust LM Lag tests consider whether the spatial lag models are appropriate. When both the LM Error and Lag statistics display no spatial dependence (not significant), only the OLS results should be used. When both LM Error and Lag statistics are insignificant, the robust LM test can be implemented. If both robust LM Error and Lag test statistics are highly significant, the model with the largest statistic value should be used (Tian, et al., 2014).

Although LM tests explicitly include spatial dependency in the model, they generally consider spatial dependence itself to be uniform across the whole study area producing a 'semi-local' rather than a truly local statistic. In contrast, GWR is a local form of weighted linear regression allowing coefficients in the model to vary from place to place (O'Sullivan & Unwin, 2010). One can estimate the location regression for a particular location, where weights are attached to observation surrounding a location. Larger weights are assigned to point near the location and smaller weights are assigned to observations far from the location. Simply, GWR involves partitioning the data into several regions and estimating a local regression model for each region individually (Fotheringham, et al., 2002).