Crime and census data relationship exploration in London

Nalmpantis Alexandros-Dimitrios, City University of London

Abstract— This study aims to explore the undelying characteristics both that explaine crime in London. The data utilised was mainly demographic data from Census 2011 and crime statistics. The data relations were examined both statistically and through geographically spatial correlations.

Index Terms— Crime, Census, Geographically Weighted Statistics

1 Motivation, data, research questions

1.1 Introduction

Crime in London can be dependent on multiple factors both residential and non. A metropolis such as London which shows continuous development and growth would naturally appear to have increasing crime rates. Variables such population density and various income related factors could be directly affecting those trends. The assumption though that crime could be related to residential variables geographically located in a hotspot area crime can appear as not true. Thus this study will examine the potential factors affecting the relationships of crime trends to specific variables that might be affecting those.

On Figure 1 the London population vs the crime rates is depicted. It is clear that the number of total offences is not inverse to the increasing population in London. Since 2001 London population is rising steadily while total crime rates are decreasing. This could potentially be an effect of non-residential variables according to [1]. For example factors such as average income of a ward in London could not be necessarily related to decreased crimes.

1.2 Data

In order to test if there is any relationship of crime to residential factors, statistical data from Census 2011 were drawn. The Census data is an ongoing statistical demographic study which has been drawn from early 1800 to this day. [2] The census covers various themes from demographic such as total population average age of population, to housing and employment characteristics per population. In addition to the census demographics the crime demographics will be utilised broken down by year since 2001 to 2012 at various crime categories. The data are broken down on ward level and were provided in comma separated format provided by City University within the Visual Analytics course for the purpose of this study. [1], [2]

Furthermore, the study aims to cove also the crime rate census geographically. Thus the GIS shapefile of London wards were utilised. In order to get the latest ward level and boundaries the data provided were related to the latest boundaries published on Ward level. [3]

1.3 Research Questions

On this study the relationship examination of crime offences to the various residential demographics of ward areas in London, was conducted. The correlation of data was analysed both on a plot and aggregated level and also to spatial level through the utilisation of geographically weighted model.

Research questions:

- Which demographic variables are strongly correlated to the total offences crime rates in London?
- How demographic subcategories such as age distribution affect crime rates in London?
- How the correlation is geographically distributed within London wards?

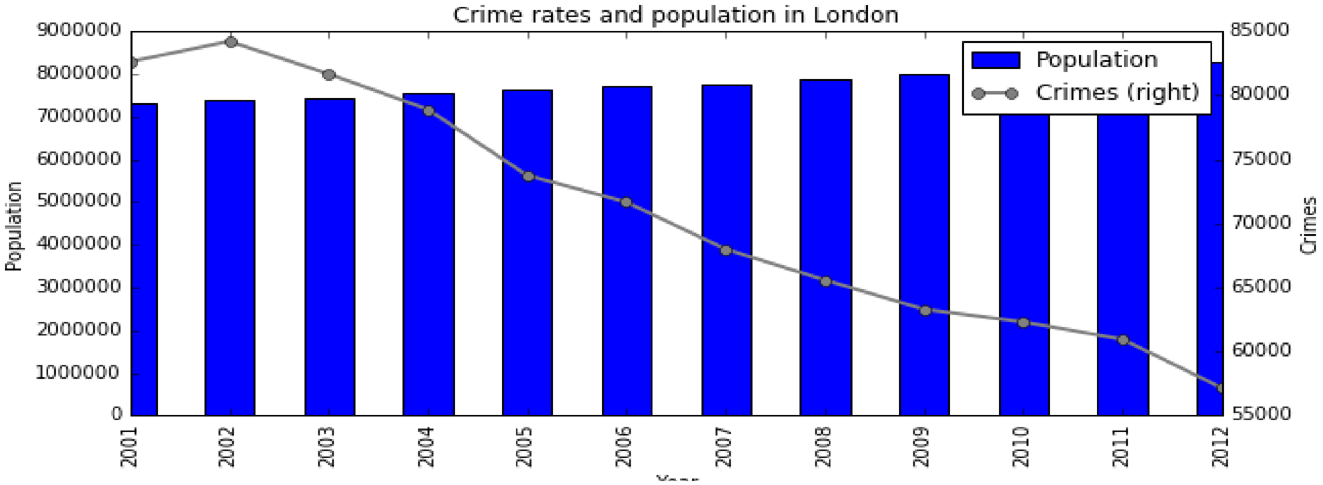

Figure 1: Crime Rates in London vs Population from 2001-2012. The population of London is represented to the y axis while crime numbers are on the secondary Y axis. The data range from 2001 to 2012.

2 TASKS AND APPROACH

2.1 Data Wrangling

The initial step of the study was to assess the suitability of the available data. The dataset was tested for missing values. The level of completeness of the dataset was investigated in order to establish the level the analysis could be affected. Furthermore, several unwanted columns were cleaned with data not required during the investigation and analysis. For example chronological data were excluded from the main database and included in separate file to access the variability of the data on yearly level.

In order to assess the spatial relations of the data, GIS shapefiles were utilised. The data were divided to London Wards and thus the spatial dataset for London wards in 2011 was utilised. For the assessment of the spatial relations both Tableau and R was utilised. The data were transformed accordingly to be readable by both of these tools utilising QGIS to transform to the required format and projection systems.

To assess the data further more databases were connected at the ward level. The average house price for each London wards was inserted and the London Output Area Classification (LOAC) was utilised. The LOAC utilises census datasets to cluster and reveal characteristics of London population classified in generalised categories. The categories reveal the similarities of population which in turn reveal the geospatial characteristics of London. Figure 3[2]

2.2 Data exploration

For the initial investigation steps of the dataset Python and Pandas tools were utilized. The dataset consists of several columns/variables which appear to have been structured in a wide format. Some initial investigation through utilising correlation matrix revealed that accessing the dataset in this form could be challenging and incomprehensible. Thus several generalised columns were selected and two correlation matrixes constructed. One correlation matrix examined the relation of various rates (% of population of a variable) to the total offences and subcategories of crimes. The matrixes then were reduced to a level of comparison of crimes to the selected variables to reveal potentially strong relations. Two correlations tests were used. Both a Pearson and a Spearman correlation test were utilised. Figure 2, Figure 4

Figure 2: Correlation Matrix of initial dataset. The pallet is diverging. Positive correlation is represented by red color while negative by blue color.

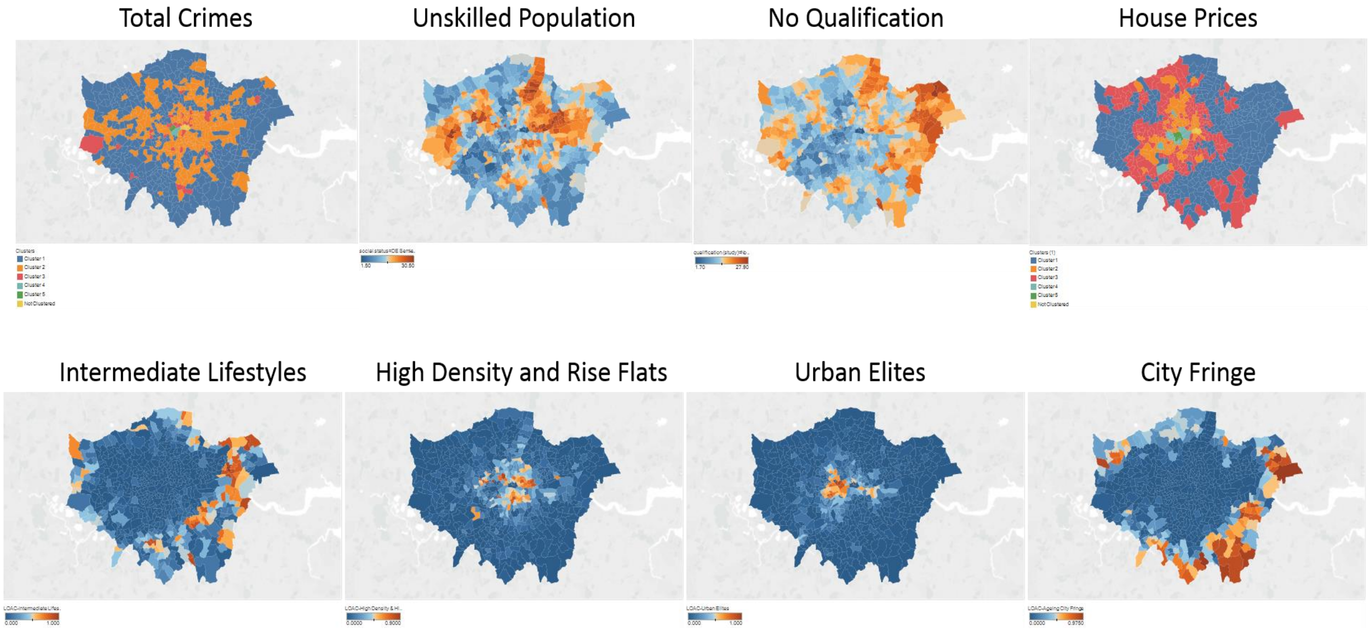

Figure 3: On the first row of maps the selected variables are drawn as choropleth maps with crime rates being the first one. Patterns that are not visible directly due to high variance of data by geography. For those clustering was utilized to reveal the potential patterns. On the second row of maps the London output are classification was used as a % of ward on the choropleth maps. For these maps the 2014 ward boundaries were utilized. QGIS software was used to convert the shapefiles to point outlines that could be read by Tableau. [4] [3] [7]

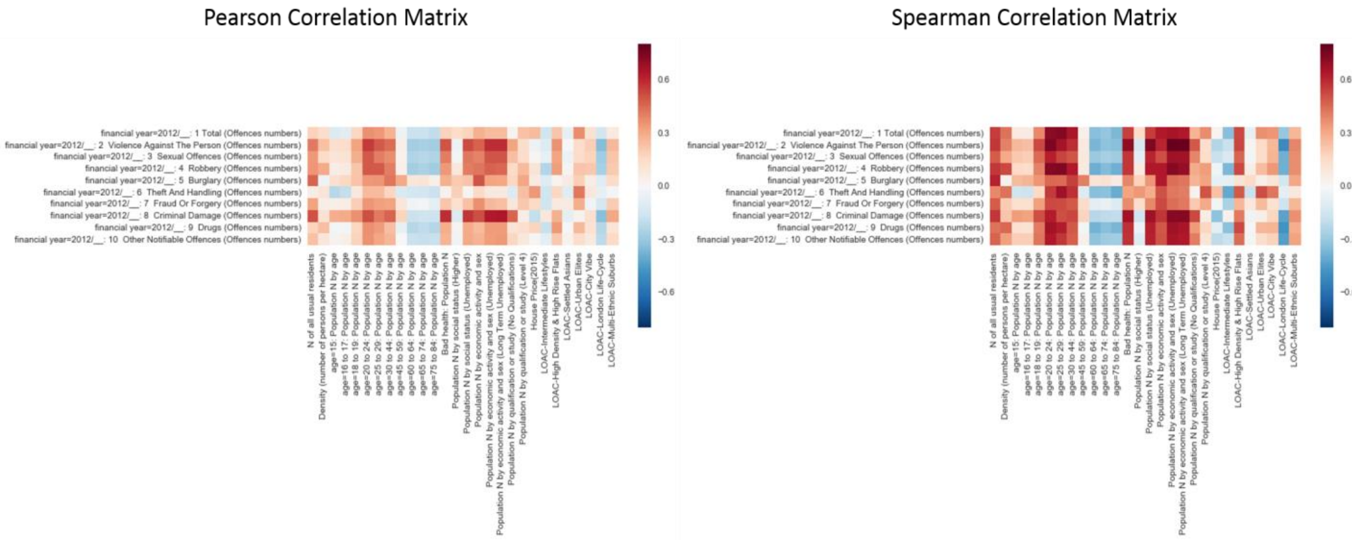

Figure 4: Reduced database correlation matrices representing the various crimes on the Y axis vs the variables on the X axis. A diverging pallet was used with red representing positive correlation coefficient and blue negative. The matrix on the right is a spearman and on the left it is a Pearson test.

2.3 Spatial Data Exploration

During this step the selected variables from the previous step were plotted on the London ward map utilizing Tableau. Those maps were used to assess the variations of those variables in London Wards. The geographical variations of the variables could be visually inspected if there are any clear correlations with the total offences geographically plotted.

Due to the level of granularity of some variables the trends are not clearly visible. Thus clustering of the areas/variables were necessary. Tableau was utilised to calculate the clusters of the geographically distributed variables. By applying k mean clustering to offences for example on the total offences data reveal the geographical patterns and variations by each ward. Figure 3

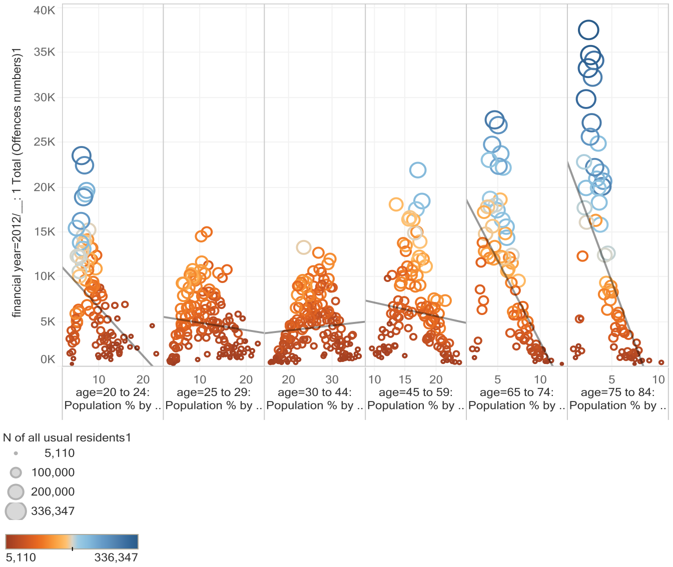

Figure 5: Crimes vs age group scatterplot. On the Y axis the crime number was aggregated for each age group while on the x axis the age groups were plotted. The bubble and the color represents the total number of residents in each ward.

2.4 Subcategory analysis

Following the correlation of the variable categories and reduction of the dimensions, the selected variables subcategories were plotted on a scatterplot along with the trend line to reveal linear relationships within the category. In addition R was calculated. For example on Figure 5 the relationship of various age category groups were plotted against total crimes to determine the relationship of age and crime.

2.5 Assess the geographical weighted statistics variations.

Finally utilizing the GWModel (Reference) package in R the geographically weighted statics were calculated for the selected variables on each Wards. Then these were plotted on a choropleth map to determine potential London area trends on the spearman correlation coefficients between total offences and selected variables. [5] [6]

3 ANALYTICAL STEPS

3.1 Data Wrangling of census and crime database

During this stage as described above the data were explored to reveal potential gaps and the state of values that could affect the analysis. The initial database contained over 600 columns of data. After initial investigation utilizing Python-Pandas columns containing multiple year of the same variable and variables that were irrelevant to the study were excluded. The clean database was reduced to a total of 370 column variables. Furthermore, several datasets were created with cleaned data from missing values. The missing values on the crime data asses at only 3% of the total dataset thus it wouldn't affect the results significantly.

The spatial data were converted utilising QGIS and were depicted on Tableau and R. [5], [6] After the shapefiles were prepared the LOAC were imported. The classification of London is clear according to the LOAC when depicted on a map. Figure 3 By examining the groups look distributed mostly between inner and outer London. Groups like city fringe are

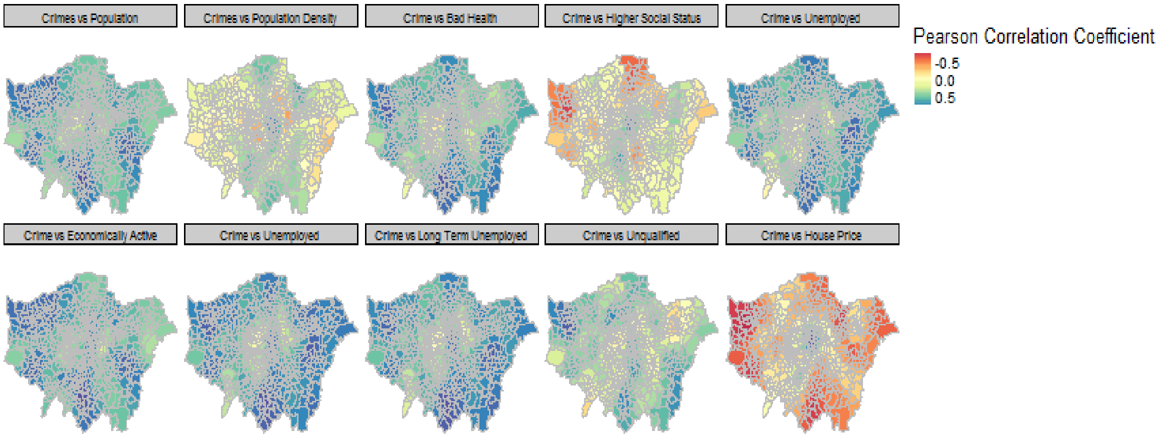
Crimes vs Population
Crimes vs Population Density
Crime vs Bad Health
Crime vs Higher Social Status
Crime vs Unemployed
Crime vs Economically Active
Crime vs Unemployed
Crime vs Long Term Unemployed
Crime vs Unqualified
Crime vs House Price
Pearson Correlation Coefficient
-0.5
0.0
0.5

Figure 6: Geographically weighted statistics by London Wards. Each facet is a choropleth map of the spearman correlation coefficient calculated vs selected variables from the initial reduction steps. The color scale is a diverging pallet with blue as positive correlation and red as negative.

concentrated to outer skirts of London while more urban groups are gravitated towards inner London. The groups express various demographic variables but in this case the economic characteristic is depicted within these groups.

3.2 Data exploration and correlation of crime statistics to census data

Initial correlation tests were conducted to determine correlations visually on the clean database. Figure 2 Due to the large size of the database though it was impossible to determine the correlations visually and thus more dimension reduction took place in order to reach to a narrower selection of variables. So the main categories of the data were selected for the initial correlation investigation instead of the bin breakdown of each category.

Then the correlation matrix was plotted with the selected variables and assumptions. The correlation matrixes were conducted for both Pearson and Spearman tests and for both % rates of each variable and Population (N) of each variable. Figure 4

The examination of the correlation matrices was sequential and refined/confined to the selected variables. A diverging colour scale was selected to make the results more visually possible to determine the positives negative correlation coefficient.

3.3 Spatial investigation of aggregated data and clustering of results to reveal patterns

The categories with the most correlated data determined on the previous step were depicted on Tableau on a choropleth map. The London ward map was formatted as a csv file so that tableau can read the outlines of each ward.

Each ward then was coloured according to the variables selected during the previous steps. Due to the high number of wards (>600) on the choropleth map the dispersion of the results for some variables made it difficult to reveal patterns. Thus the clustering of the data were required. Data were clustered using the clustering (K-Mean) feature of Tableau. On Figure 3 example the total number of offences don't appear as to have any pattern. This is because central areas in London have very high crime rates and skew the data. Thus by clustering the total number of offence, patterns start to reveal. Figure 3

3.4 In depth analysis of the selected subcategories from previous steps

On this step the column data that appear to have some correlation to the crime statics will be analyzed further. In order to determine the relationship the data were plotted on Tableau as scatterplots along with trend line and colored by the population number. For example on Figure 5 the age distribution appears to have linear relationship with the crime statics in certain cases.

This step gave a more in depth understanding of the independent variables and how the variation appears between certain bining/ categories of those data. R squared was also calculated along with summary statistics to determine the linearity of the relationship.

3.5 Geographically weighted statics analysis of crime data to selected variables

On the final step of the analysis the geographically weighted statistics were calculated. This will show the geographical distribution of the spearman correlation coefficient for the selected variables from previous steps.

For this step the shapefiles of London Wards along with the extract of selected variables from the previous steps were connected. The geographically weighted statistical summary was then calculated by utilizing the GWModel in R. Then the variables of crime vs variable Spearman correlation coefficient is plotted on a choropleth map. Figure 6

4 FINDINGS

The steps above has given an overview of the most important factors explaining the total crime numbers in London. The factors were selected in terms of Spearman coefficient. The initial matrix plots revealed strong positive correlation of variables such as age and economic activities. From Figure 4 unemployment and education as well appear to positively correlate to crime offences while certain older age groups appear to be negatively correlated.

Apart from the census data data such as house prices and London Output Area Classification % by ward was tested. It appeared that mean house prices are not directly correlated to crimes other than positive correlation to theft. Same results appear to LOAC, where more affluent and central London classifications appear to be positively affected by crime. On the other hand classifications that are located mostly outside central London appear to less affect by the number of crimes.

Following the initial investigation of the variables the spatial relations were examined. The variables were mapped utilising Tableau on Figure 3The maps observed on the provided figure represent some variables with strong correlation to the total number of offences. Clustering was used in the case of total offences and median house prices maps due to the dispersity of the data. There are spatial relations between crime and for example median house prices and skill level of population. Furthermore, the bottom rack of maps represent the classification of London where the spatial relations can be observed. The spatial correlation appear to be confirming the previous correlation matrix results.

Moreover, the the spatial relations and correlation matrices categories such as age were examined in depth to reveal the relationships to crime. For this purpose the Figure 5 was used to assess on a scatterplot each age group to identify how crime rates trend. From the figure it is apparent how the selected variables are distributed geographically. For example while in most London the house prices are negatively correlated in central London it is positively correlated. Similar to this example but inverse is the population density of wards in London.

5 REFLECTION

This study was targeting on explaining the variations that affect crime rates in London. The crime rates were tested against multiple criteria which started from a larger census database with added data from various databases. Following some initial exploration the database were required to be reduced in order to utilise visual analysis.

The large number of variables to be tested made the visual representation of the data cluttered. Early on the analysis it was apparent that the dataset required to be reduced into general categories to have a better/clearer view of the visualisation analysis. For example the initial correlation matrix appeared to be so cluttered that any kind of analysis would be impossible due to the fact that the variables were 600 columns wide.

Following the initial representation of data, several categories of data cleaned. Multiple columns with multiple years of the same variable in wide format were excluded. After the reduction the same correlation matrices were recalculated and were filtered to the variables to be examined. The correlation matrices were tested with two correlations tests and Spearman test appeared to provide stronger correlations between the results. This could explain a not necessarily linear relationship in which spearman is more suitable to determine the correlation.

The variables chosen above were than tested on a geospatial level. Initial the variables were displayed on a ward level by utilising Tableau. But due to the high variance between ward values in London for some variables clustering was required. For example by clustering total crimes in London into 5 clusters it became clear that crime levels in central London are disproportional to the crime levels outside central London. This is a concern and potential limitation of this study since it could reveal that crime is not necessarily related to residential characteristics of one ward. The increased crime levels in central London could be attributed to the high number of visitors in those areas and not to the residents. Thus testing the correlation of crime rates to demographic data in those areas could result into false assumptions. [4], [7]

Thus in order to assess the dataset as geographically weighted, the GW-Model on R was utilised. The summary statistics of the selected variables were tested by alternating various bandwidths to reveal the potential patterns.

The analysis revealed that there are factors that explain the crime rates within London. A deeper dive analysis coupled by grouping for example on Output Area Classification could reveal further insights. The disproportionate crime rates also reveal that additional significant factors affect the patterns which cannot be examined by demographic variables only.

6 BIBLIOGRAPHY

[1] GLA_Intelligence, "Census Information Scheme," Greater London Authority, 2011. [Online]. Available: https://data.london.gov.uk/census/. [Accessed November 2016].

[2] GLA_Intelligence, "London Output Area Classification," Greater London Authority, [Online]. Available: https://data.london.gov.uk/dataset/london-area-classification. [Accessed November 2016].

[3] O. o. N. Statistics, "Open Geography Portal," [Online]. Available: http://geoportal.statistics.gov.uk/. [Accessed November 2016].

[4] N. Malleson and M. Andresenb, "Exploring the impact of ambient populationmeasures on London crime hotspots," *Journal of Criminal Justice,* vol. 46, pp. 52-63, 2016.

[5] CRAN-R, "Geographically Weighted Models," [Online]. Available: https://cran.r-project.org/web/packages/GWmodel/GWmodel.pdf. [Accessed December 2016].

[6] Tableau, "Create Tableau Maps from Shapefiles," [Online]. Available: https://onlinehelp.tableau.com/current/pro/desktop/en-us/maps_shapefiles.html. [Accessed November 2016].

[7] V. Spicer, J. Song, P. Brantingham, A. Park and M. Andresen, "Street profile analysis: A new method for mapping crime on major roadways," *Applied Geography,* vol. 69, pp. 65-75, 2016.

[8] L. Registry, "Average House Prices, Ward, LSOA, MSOA," Land Registry, [Online]. Available: https://data.london.gov.uk/dataset/average-house-prices-ward-lsoa-msoa. [Accessed November 2016].

[9] M. Inc, "A comparison of the Pearson and Spearman correlation methods," Minitab, [Online]. Available: http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/. [Accessed November 2016].

[10] I. Gollini, B. Lu, M. Charlton, C. Brunsdon and P. Harris, "GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models," *Journal of Statistical Software,* vol. 63, no. 17, 2015.