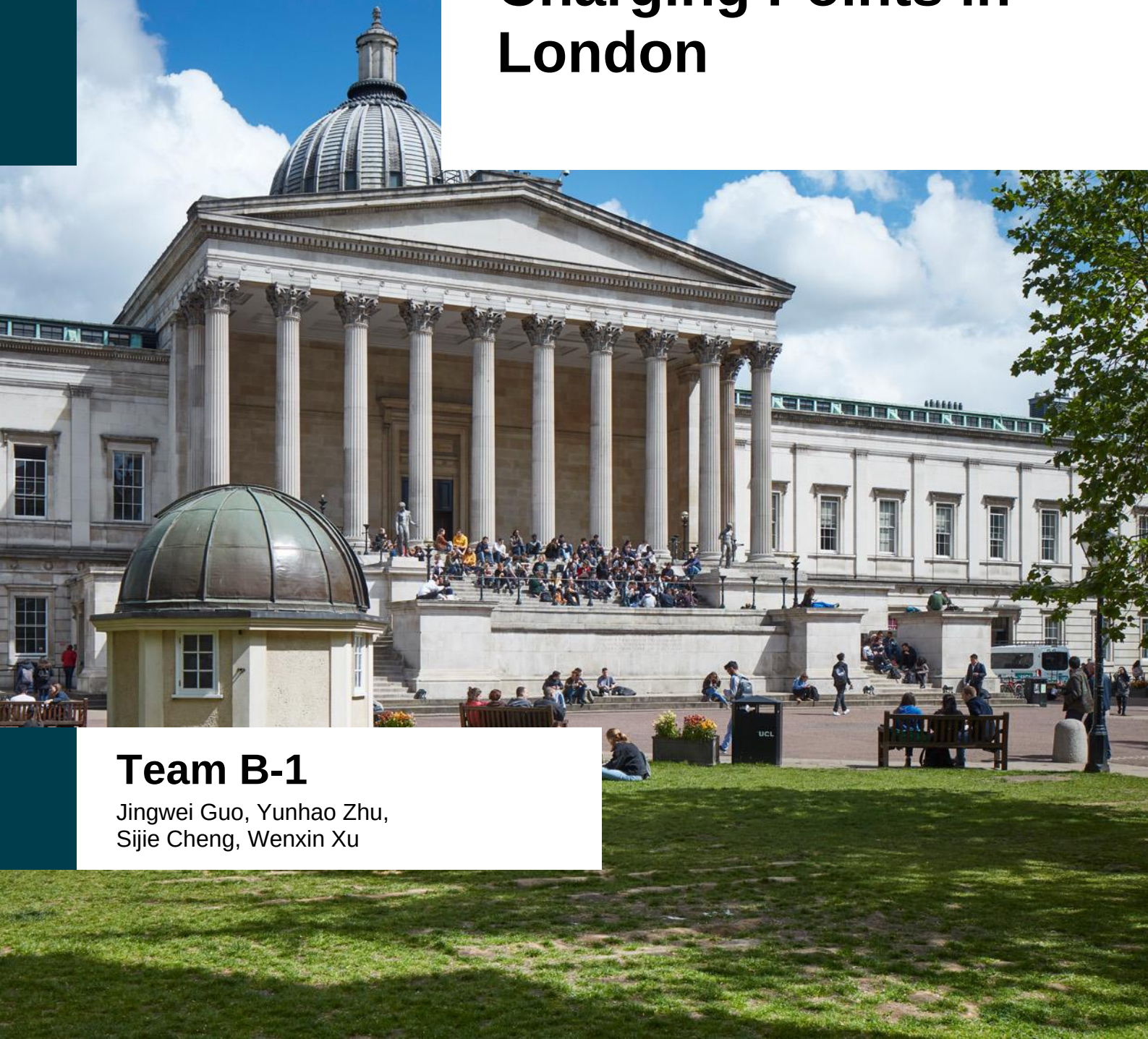
Civil, Environment and Geomatic Engineering
CEGE0097: Spatial Analysis and Geocomputation

Word Count: 5794

Site Selection for Electric Vehicle Charging Points in London



Team B-1

Jingwei Guo, Yunhao Zhu,
Sijie Cheng, Wenxin Xu

Contents

1. Introduction

In the current era, most nations around the globe have come to an agreement that the reduction of greenhouse gas emissions is the major key to the global climate change [1]. Since about 25% of the world's total greenhouse gases emissions come from traditional transportation [2], the International Energy Agency (IEA) proposed a global plan to achieve zero carbon emissions (net zero) by 2050, with no internal combustion engine cars expected to be sold by 2035 and all cars to be electrified by 2050 [3].

The demand for charging piles as an infrastructure for electric vehicles is also expected to grow rapidly in the future. According to the summary book *London's 2030 Electric Vehicle Infrastructure Strategy* [4], it is predicted that there should be 40,000 to 60,000 charging points in London in 2030.

As a result, researchers have proposed different models for siting public charging piles from various aspects. Yang et al [5] derived a method for siting charging stations with zonal charging demand; Zafar et al [6] sited charging stations based on Maximum Coverage Location Problem and QGIS software; Kaya et al [6] used Analytic Hierarchy Process (AHP), Preference Ranking Organization Method and multi-criteria decision making for the siting analysis of charging stations.

The aim of this project is to perform a multi-criteria site selection for public charging points in London, whose results may provide suitable areas for the installation of charging stations in the future. Based on 11 sets of parameters, using Lower Layer Super Output Area (LSOA) as the scale, the article first pre-processes the data with Kernel Density Estimation (KDE), Euclidean distance and Zonal Statistics. Then, spatial regression models are constructed to identify the factors influencing the current location of public charging points in London. Finally, referring to the analysis results, Analytic Hierarchical Analysis (AHP) is used to select suitable charging post locations.

The project is done by group B-1 of CEGE0097. Part 1, 2, 3 and 8 are written by all the 4 members, while the others are the individual parts. Part 4, 5, 6 and 7 are done by Sijie Cheng , Jingwei Guo, Wenxin Xu and Yunhao Zhu respectively.

2. Data Description

The choices of data for this project are fully considered and the sources are consistently chosen from authoritative organizations' websites. All the data is extracted from governmental sources: UK's Department of Transport, Greater London Authority (GLA), National Census,

and OpenStreetMap.

Taking the four aspects of transportation, population, economy and zoning into account, the project chooses 11 indicators for each. Their sources are shown in Table 2.1.

Table 2.1 Data definitions and sources

Theme	Variable Name	Description	Source
Dependent variable	Public charging points	Location of public charging points	UK government Department for Transport (2020)
Transportation	Traffic flow on major roads	Join the major roads with the number of motor vehicles	UK government Department for Transport (2020)
	Cars per Household	The number of cars owned by each family	Greater London Authority (GLA) (2010)
	Public transportation accessibility	Average Score from 0 to 6	GLA (2014)
Population	Population density	Population living per sq. km	GLA (2013)
Economy	House price	Median Price (£)	GLA (2014)
	Household income	Median Annual Household Income estimate (£)	GLA (2012)
Regional	Hospital	Shape and location of hospital	OpenStreetMap (2021)
	School	Shape and location of school	
	Green Space	Shape and location of park	
	Residential	Shape and location of residential	
	Substation	Shape and location of substation	National Grid Group (2021)

3. Exploratory Spatial Data Analysis

The data sources and their classifications are shown in Table 2.1. The downloaded data can be grouped into two main categories: vector data contained in shapefiles and numeric values.

The dependent variable, public charge point data, is presented as vector points, which is clipped to the London range for further analysis.

The average traffic flow data for London's main roads is a combination of two datasets, a vector layer of the UK's major traffic roads and a CSV table of motor vehicle flows on each

road. Project links these two datasets and clips them into the extent of London for further analysis.

The data for LSOA is from GLA. The LSOA boundary vector data (2011) is contained in a shapefile, and each borough has a unique LSOA code. Statistics within the LSOA are stored in CSV tables, and according to references and CSV table article selects population density, public transport accessibility, car availability and household income as explanatory variables for the analysis. By linking the data values to the underlying LSOA layer, these non-spatial data can be used in further spatial analysis procedures.

OpenStreetMap provides spatial data of land use classification for the whole London area. Except the roads, each piece of land is digitized and categorized into different land use classifications. By selecting and extracting attributes, the article creates vector layers containing polygons that represent hospitals, schools, green spaces, residential areas and substations.

4. Data Pre-processing (Sijie Cheng's Personal Part)

As discussed in the data description part, the data consists of two types: vector data containing points, lines and polygons, and numerical data in the scale of LSOA grids. In this part, our article applies data pre-processing procedures to unify their spatial extent and the presenting formats. The summary of the data and their pre-processing methods are listed in Table 4.1.

Table 4.1 Data types and pre-processing methods

Data Type	Data Name	Pre-processing Methods
Point	Public Charging Points	Kernel Density Estimation (KDE) + Zonal Statistics
Line	Traffic Flow on Major Roads	
Polygon	Hospital School Green space Residential Area Substation	Euclidean Distance + Zonal Statistics
Numerical Values of each LSOA	Population Density Cars per Household Public Transportation Accessibility House Price Household Income	Join

4.1. Point and Line Data Pre-processing

For the Point and Line data, the article uses KDE to convert them into raster layers. Then, Zonal Statistic is used to assign a value to each LSOA grid by calculating the mean value of the raster data in that area.

4.1.1. Methodology

Kernel Density Estimation (KDE) is a non-parametric method for estimating the density of a random variable. It obtains the shape of the density from the data rather than the parameters. The KDE method [7] assumes that density is available at any location in the study area where the point data is located, rather than having values only at locations where events occur [8]. Due to its applicability for interpolation and spatial visualization within continuous surfaces, KDE is widely used in the geospatial domain [9]. KDE formulation is shown as Equation 4.1.

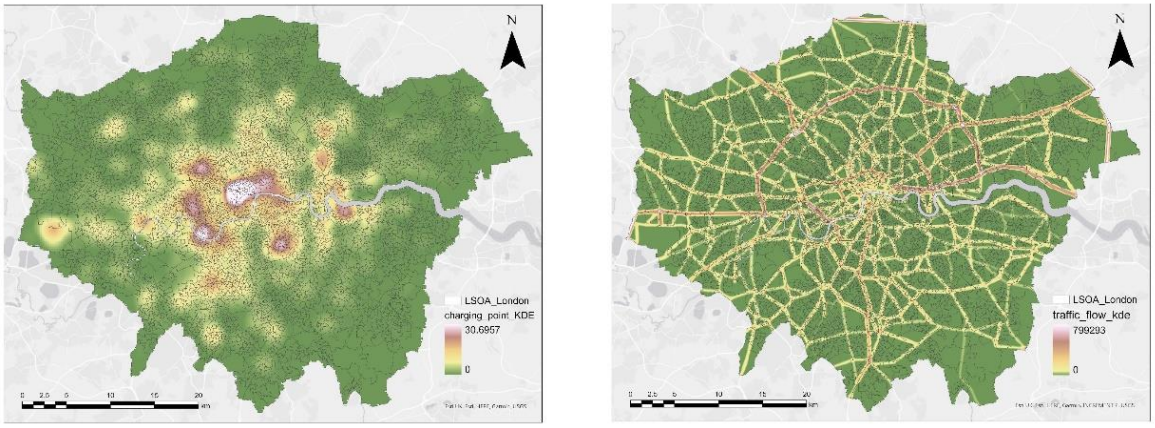
$$\text{density} = \frac{1}{(\text{radius})^2} \sum_{i=1}^{n} \frac{3}{\pi} \cdot \text{pop}_i \left(1 - \left(\frac{\text{dist}_i}{\text{radius}}\right)^2\right)^2 \tag{4.1}$$

Where i (= 1, 2, …, n) means the input points. If they lie within a radius distance of the (x, y) position, only the points in the sum are included. The pop_i is the population field value of point i, which is an optional parameter. The dist_i is the distance between point i and the (x, y) position.

When KDE is used on lines data, the effect of line segments on density is equal to the effect of the value of the kernel surface at the center of the raster image element on density.

The Zonal Statistic is a tool to compile statistics (like average, sum, min or max) in the defined areas. It only calculates one value for raster output at a time. The value then becomes the image value of the raster output for the image element corresponding to the region [10].

4.1.2. Results



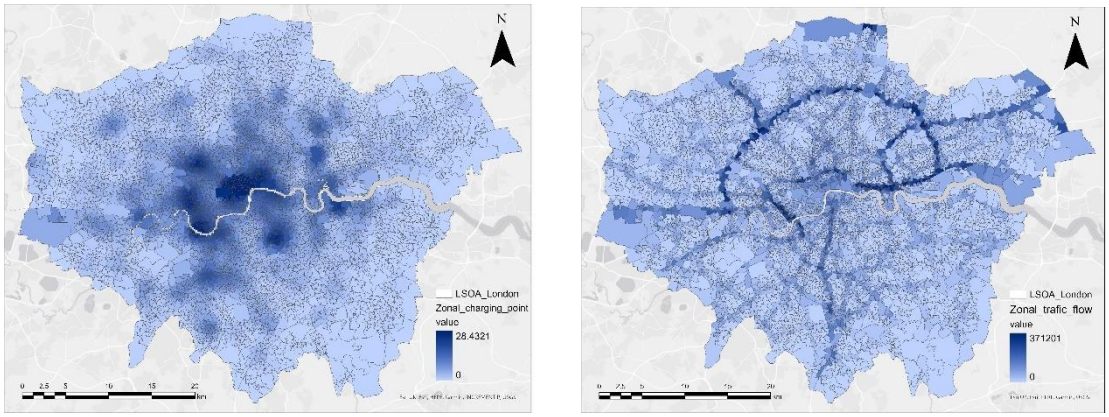
a) charging point

b) main road traffic flow

Figure 4.1 KDE output of charging points and traffic flow data

Kernel Density Analysis tool in ArcGIS pro is performed on public charging points and traffic flow on major roads data, whose results are shown as Figure 4.1.

Then, the Zonal Statistic Tool is used to calculate a mean value of the output raster layer for each LOSA zone, as shown in Figure 4.2.



a) charging point

b) main road traffic flow

Figure 4.2 Zonal Statistic output of charging points and traffic flow data

4.2. Polygon Data Pre-processing

To study the impact of the distance to some buildings on the number of charging points, the article calculates the Euclidean distance from the polygon data (hospital, school, green space, residential area and substation). Then, Zonal Statistic is used again to assign a distance value for each LSOA area.

4.2.1. Methodology

The Euclidean distance is the shortest distance between two points or vectors, it is the "ordinary" straight-line distance between two points in everyday life, as shown in Equation 4.2.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{4.2}$$

Where x_1, x_2 are calculated distances from the relative horizontal positions of two points while y_1, y_2 are those from the relative vertical positions of two points.

4.2.2. Results

This article uses the Euclidean distance tool of ArcGIS Pro to realize the calculation. It obtains the straight-line distance between a pixel to every source and choose the shortest one as the output.

Since the distribution of the polygon data is too dense so that there is no significant difference in the output value in different LSOA grids, this article screens out the polygons whose area is larger than the average for calculation. The output Euclidean distance raster layers are shown as Figure 4.3.

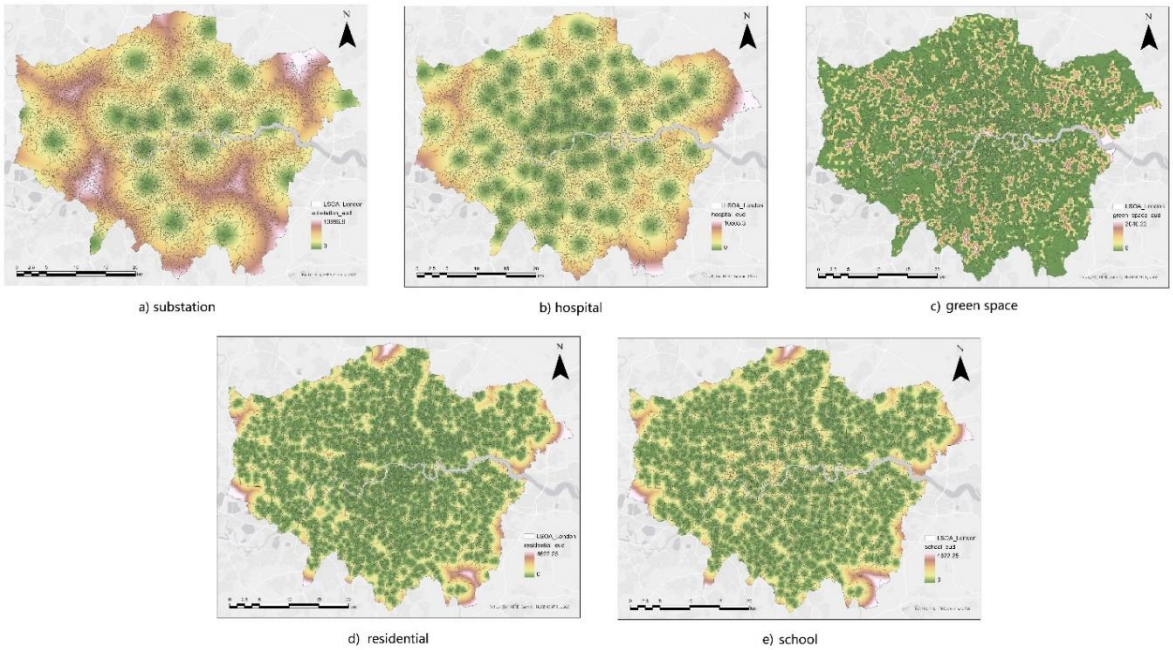


Figure 4.3 Euclidean Distance output of polygon data

Then, the Zonal Statistic Tool is used again to calculate a value for each LOSA grid, as shown below (Figure 4.4).

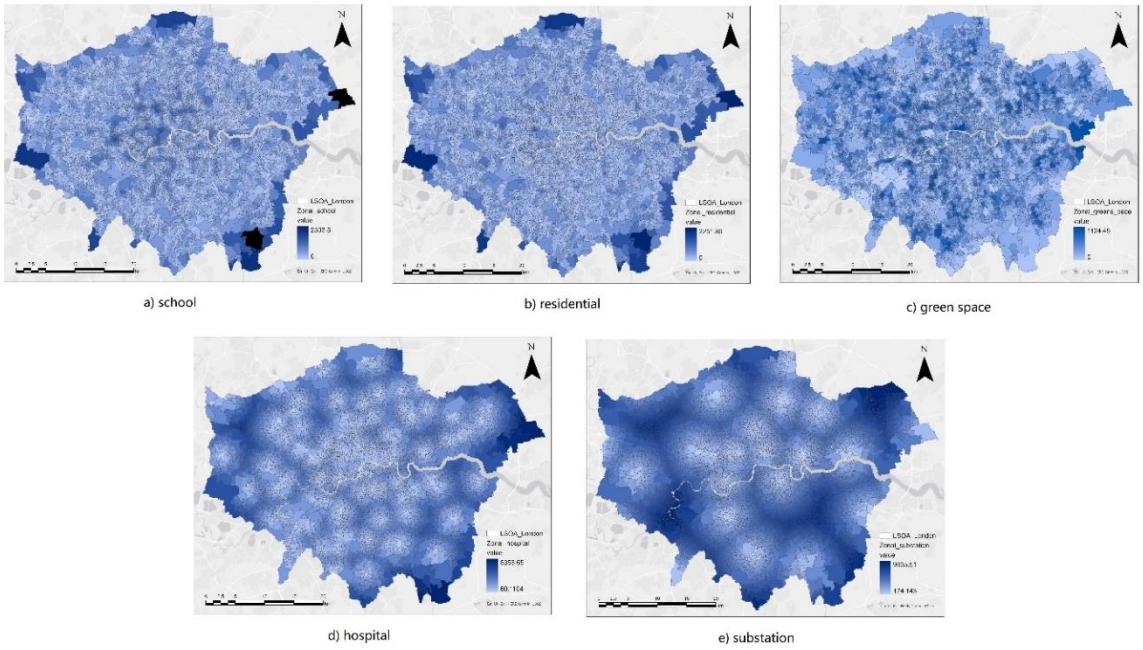


Figure 4.4 Zonal Statistic output of polygon data

4.3. Numerical Data (of each LSOA) Pre-processing

The numerical data is originally of the LSOA scale, thus is only needed to be joined to the shapefile of the LSOA zones.

4.3.1. Methodology

The join functionality in ArcGIS Pro is used to relate the base layer of LSOA 2011 grid zones and the Census Data by setting the same LSOA field code as the common attribute.

4.3.2. Results

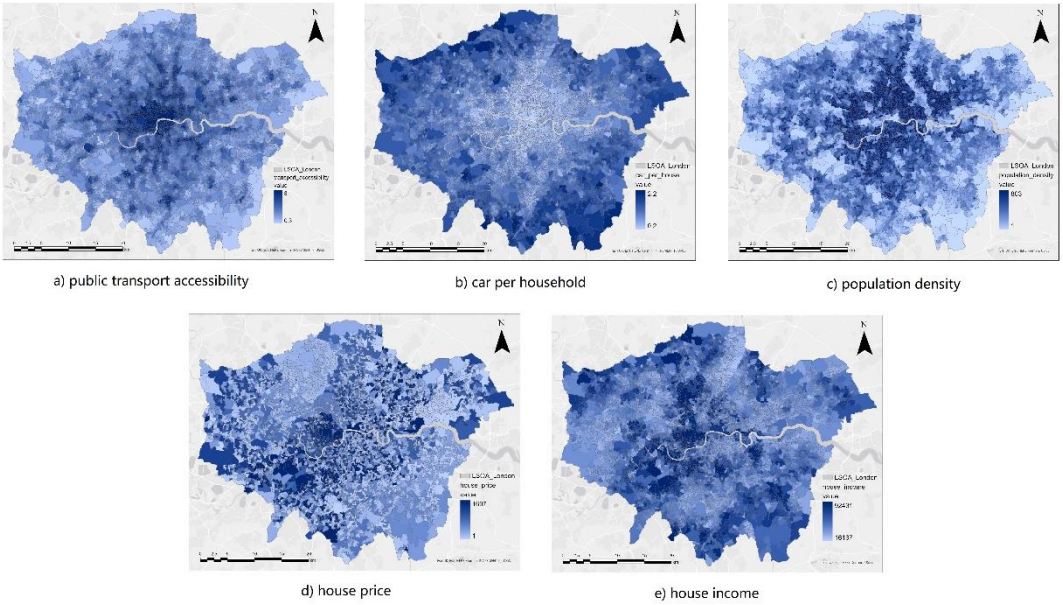


Figure 4.5 Visualized results of numerical data

The following maps in Figure 4.5 show the visualized results of population, household income, mean house prices, cars per household, and transportation accessibility.

5. Global Spatial Regression (Jingwei Guo's Personal Part)

In order to discover what factors are influencing the existing location of charging points in London, and thus provide reference for the weight evaluation of subsequent site selection, this article conduct global spatial regression analysis with the data which we have processed in Part 4. These data are shown in Table 5.1.

Table 5.1 Name and meaning of the pre-processed data

Data Type	Name	Meaning
Dependent Variable	Charging	Density of charging stations

Independent Variables	TrafficFlo	Impact of traffic flow on the major road
	Population	Population density
	PublicTran	Public transport accessibility
	CarsperHou	Number of cars per household
	HousePrice	Mean house price
	HouseholdI	Household income
	Dis2Substa	Distance to substations
	Dis2Green	Distance to green space
	Dis2Hospit	Distance to hospitals
	Dis2Reside	Distance to residential area
	Dis2School	Distance to schools

In this part, all the variables are normalized first. Then, some independent variables are filtered out with multicollinearity measurement to enhance the reliability of the regression. After that, this paper conducts regression analysis with OLS, SLM and SEM, and compares their results.

5.1. Data Normalization and Cleansing

5.1.1. Methodology

Before conducting regression analysis, we first cleanse it by deleting the rows with null values.

What's more, since the size and unit of the independent variables are quite different, if we conduct regression on these data directly, the output coefficients may also vary greatly. For the purpose of the spatial regression here is to discover the impact of each factor on the location of charging points by comparing their coefficients, the article normalizes the data to the range of 0 to 1 with the equation below, where x means a column of data, x_i means each data in the column and v_i is the normalized data.

$$v_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{5.1}$$

5.1.2. Results

After the data cleansing, 127 rows of data in our dataset have been deleted. Then, the remaining data are normalized to 0 to 1 by implementing Equation 5.1 with Excel formulas.

5.2. Variable Filtering by Multicollinearity Detection

5.2.1. Methodology

In a linear regression model, multicollinearity stands for the phenomenon in which one explanatory variable can be explained by more than one explanatory variables with a linear equation, as shown below in Equation 5.2.

$$X_{1i} = k_1 X_{2i} + k_2 X_{3i} + \cdots + k_n X_{(n+1)i} + b \tag{5.2}$$

If multicollinearity exists in the explanatory variables, the coefficients of the regression can vary greatly from each other, which may make the precision meaningless [11]. To improve the reliability of the regression results, it is necessary to filter out some variables with high multicollinearity before the regression.

This article chooses variance inflation factor (VIF) for multicollinearity detection. Independent variable X_i's VIF can be calculated with Equation 5.3 [12], in which $R^2{}_i$ is the coefficient of determination of the linear regression equation that can predict X_i with all the other Xs.

$$VIF_i = \frac{1}{1 - R^2{}_i} \tag{5.3}$$

If VIF_i is larger than 5, then X_i is considered to be high in multicollinearity, thus needs to be deleted from the variables. Once a variable is excluded, VIF should be calculated again until all the VIFs are less than 5.

5.2.2. Results

In this article, vif() method in the car package of R was utilized to calculate the VIFs of all the explanatory variables, whose results are shown in Table 5.2.

Table 5.2 Results of multicollinearity detection

Independent variables	Variance Inflation Factor (VIF)
TrafficFlo	1.123512
Population	1.943430
PublicTran	2.463148
CarsperHou	3.606109
HousePrice	2.160577
HouseholdI	2.333472
Dis2Substa	1.390437
Dis2Green	1.079728
Dis2Hospit	1.294374
Dis2Reside	2.215040
Dis2School	2.148970

According to the results, the VIFs of all the variables are less than 5, which means there is no significant multicollinearity among them. So, this article uses all these 11 variables as the explanatory variables in the regressions.

5.3. Linear Regression with Ordinary Least Squares

Since linear regression's results can indicate whether it is necessary to use other kind of models, this article start with it by using ordinary least squares (OLS).

5.3.1. Methodology

A linear regression model [13] can be expressed in the following form (Equation 5.4), where Y is an n*1 column vector which contains all the observations of the dependent variable, X is an n*m matrix of explanatory variables, β is an m*1 vector with all the coefficients included, and ε contains n random errors.

$$Y = X\beta + \varepsilon \tag{5.4}$$

To find the best β for this equation, OLS is always used whose purpose is to find values of β which minimize the sum of squared errors of Y. OLS's calculation method is shown as below.

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{5.5}$$

5.3.2. Results

By constructing the linear regression model in R, this article gets the diagnostics indicators of it, as shown in Table 5.3.

Table 5.3 Results of linear regression

Diagnostics indicators		Value
Residuals	Min	-0.53928
	1Q	-0.03861
	Median	-0.00763
	3Q	0.01979
	Max	0.81205
Multiple R^2		0.4412
Adjusted R^2		0.4399
p-value		0

According to Table 5.3, the adjusted R^2 is 0.44, which indicates that the model can only explain about 44% of the distribution of the charging stations in London.

Although the residuals seem to be normally distributed according to the indicators, the Quantile-Quantile plot (Figure 5.1) shows that there is deviation from the normal distribution at the tails. What's more, the distribution map of the residuals (Figure 5.2) implies that there may be obvious spatial autocorrelation.

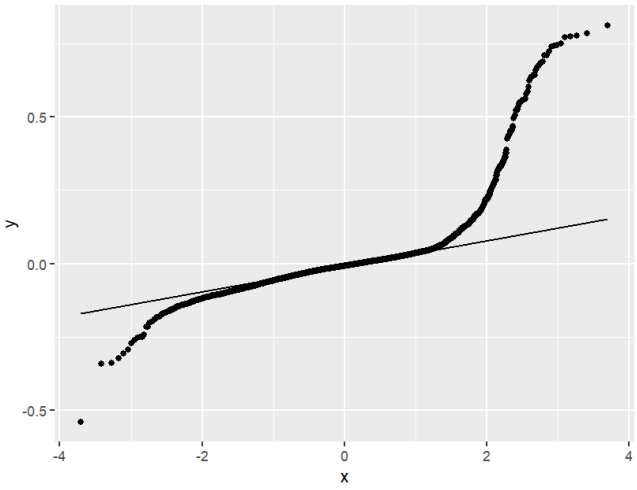


Figure 5.1 Quantile-Quantile plot of the residuals of linear regression

Thus, the article conducted diagnostic tests with Moran's I and Lagrange Multiplier tests for spatial lag and error dependence, to discover whether there is spatial autocorrelation in the regression residuals, whose results are shown in Table 5.4.

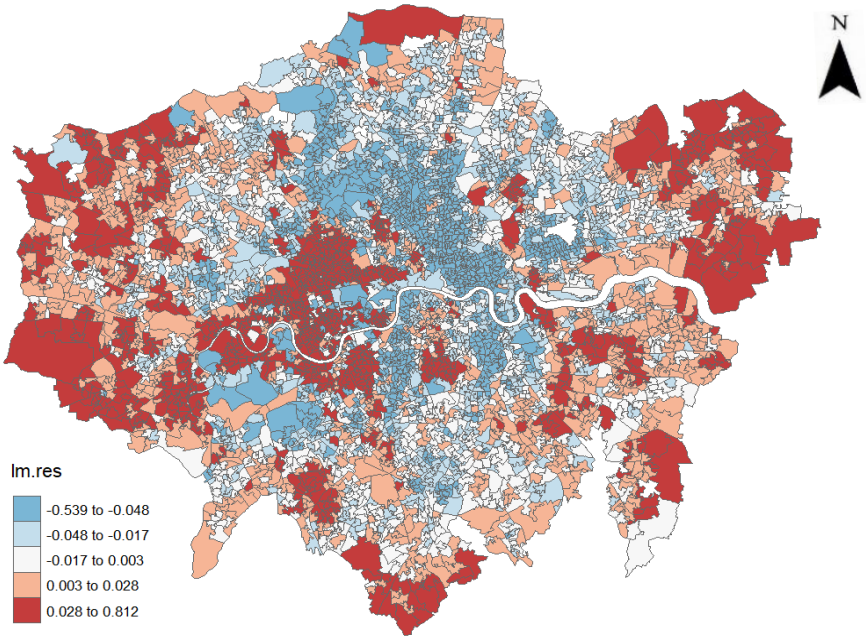


Figure 5.2 Distribution map of the residuals of linear regression

Table 5.4 Diagnostic tests on linear regression

Method	Output Type	Value	P-Value
Moran's I	Observed Moran I	0.84	0.00007
LM test for spatial lag dependence	RLMlag	568.03	0
LM test for spatial error dependence	RLMerr	308.42	0

According to Table 5.4, the output Moran's I is 0.84 with a very small P-value. This indicates the significant autocorrelation in the residuals, which is against the basic assumption of linear regression. In this case, the model is not reliable for its estimation may be biased or wrong. Since the chi-squared values of RLMlag and RLMerr [14] are both large and significant, the article tries to use both spatial lag model and spatial error model to explain the existence of spatial autocorrelation in the residuals.

5.4. Global Spatial Regression with Spatial Lag Model and Spatial Error Model

In 5.3, this article has proved the limitation of the linear regression for there is significant autocorrelation in its residuals. Hence, spatial regression is conducted on the same dataset with spatial lag model and spatial error model to explain this autocorrelation.

5.4.1. Methodology

The spatial lag model (SLM) takes the autocorrelation of the dependent variable into account. On the basis of linear regression, it adds a spatial lag term into the equation [15]. The model can be expressed in the following form, where Y, X and β are the same as those in Equation 5.4, ρ stands for the autocorrelation parameter, W is a matrix of weighted average of the nearby values, and ε is a spatially uncorrelated error vector.

$$Y = \rho WY + X\beta + \varepsilon \tag{5.6}$$

The spatial error model (SEM) assumes the spatial autocorrelation is due to some unobserved explanatory variables. It adds a spatial error term into the equation of linear regression which is shown as follows [15]. In this equation, u is an n*1 vector of residuals, λ stands for the autocorrelation parameter, W is the matrix of spatial weight or neighborhood connectivity, and ε is a normally distributed error vector.

$$Y = X\beta + \varepsilon + \lambda Wu \tag{5.7}$$

For both of these two models, maximum likelihood estimate (MLE) is used to estimate the best parameters [16].

5.4.2. Results

By constructing both of the two models with lagsarlm() and errorsarlm() function in R, this article gets the results of them, which is shown in Table 5.5.

Table 5.5 Results of spatial lag model and spatial error model

Diagnostics indicators		Spatial Lag Model	Spatial Error Model
Residuals	Min	-0.14972498	-0.15232087
	1Q	-0.00535061	-0.00536550

	Median	-0.00059314	-0.00097324
	3Q	0.00409598	0.00386934
	Max	0.20924347	0.21372215
Log likelihood		11890.89	11891.22
AIC		-23754	-23754
Spatial Autocorrelation Parameter (Rho for SLM and Lambda for SEM)	value	0.9942	0.99746
	LR test value	14529	14529
	p-value of LR test	0	0
	Asymptotic standard error	0.0011128	0.00074069
	z-value of asymptotic t-test	893.4	1346.7
	p-value of asymptotic t-test	0	0

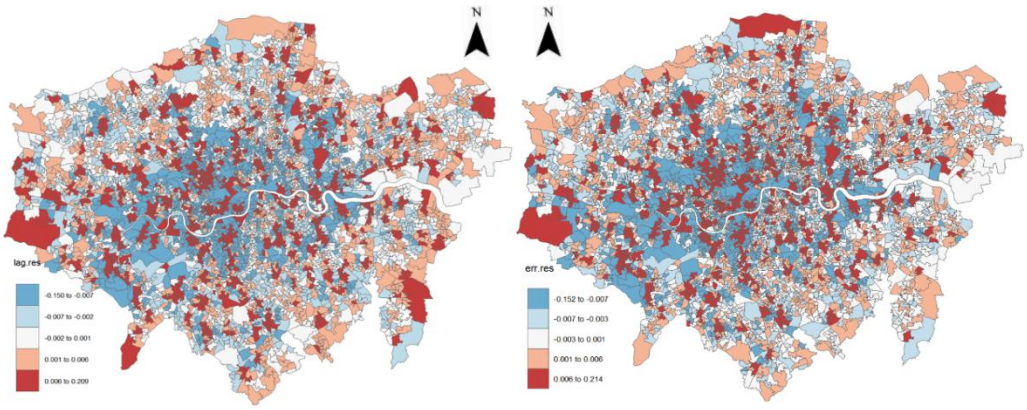
In both of the two models, the spatial autocorrelation parameters (Rho & Lambda) are statistically significant as expected, since the p-values of t-test and Likelihood Ratio test on the parameters are very close to zero.

For the definition of R^2 in these two models is different from that in the linear regression, to compare the fitting level among the three models, Akaike Information Criterion (AIC) [17] is used. According to Table 5.5, the AICs of the two models are both -23754, which are much smaller than that of the linear regression (-9227.1). This indicates that the two models fit the dataset better.

This article also checks the autocorrelation of the residuals with Moran's I. As shown in Table 5.6, the incorporation of the spatial components effectively mitigates the spatial autocorrelation in the residuals since the Moran's I drops from 0.84 to about 0.05. This obvious reduction in autocorrelation can also be seen visually in the distribution maps of residuals, if comparing Figure 5.2 with Figure 5.3.

Table 5.6 Moran's I of the three regression models

Model	Moran's I	P-Value
Linear Regression	0.84	0.00007
Spatial Lag Model	0.050389	0.0008
Spatial Error Model	0.046238	0.0009



(a) Spatial Lag Model (b) Spatial Error Model

Figure 5.3 Distribution map of residuals of SLM and SEM

Although the AIC and Log likelihood of the two models are similar, this article chooses SEM as our final global regression model, since interpreting the impact of the variables in SLM is much more complicated than in SEM due to the presence of the spatial multiplier [18]. This means the estimated coefficients of the SEM is more suitable for providing reference for the site selection.

Table 5.7 shows the estimated coefficients of SEM. According to this table, among all the variables, population density, public transport accessibility, distance to green space, residential area and schools' P-values are lower than our significance level of 0.01, which means these variables' impacts are significant. If we sort these significant variables by the absolute values of their coefficients, they can be listed as: Population > Dis2Reside > PublicTran > Dis2School > Dis2Green.

Table 5.7 Estimated coefficients of SEM

Independent variables	Coefficients	StdError	t-Statistic	P-value
(Intercept)	0.0724307	0.0940626	0.7700	0.4412838
TrafficFlo	0.0033997	0.0028983	1.1730	0.2408016
Population	0.0301636	0.0047202	6.3904	1.655e-10*
PublicTran	0.0195000	0.0029310	6.6529	2.874e-11*
CarsperHou	-0.0078163	0.0040309	-1.9391	0.0524877
HousePrice	0.0051651	0.0067307	0.7674	0.4428456
HouseholdI	0.0086851	0.0036795	2.3604	0.0182562
Dis2Substa	-0.0143049	0.0178610	-0.8009	0.4231877
Dis2Green	0.0083374	0.0019499	4.2757	1.905e-05*
Dis2Hospit	0.0052264	0.0078524	0.6656	0.5056801

Dis2Reside	-0.0251038	0.0054469	-4.6088	4.050e-06*
Dis2School	0.0185948	0.0050180	3.7056	0.0002109*

N.B. * next to a number indicates a statistically significant p-value at the level of 0.01.

Although the improved fitting performance of SEM has been proved, the model still has some limitations. According to the Quantile-Quantile plot of the residuals, there is still some deviation from the normal distribution line at both ends, which inspires us to try local spatial regression models that takes spatial heterogeneity into account.

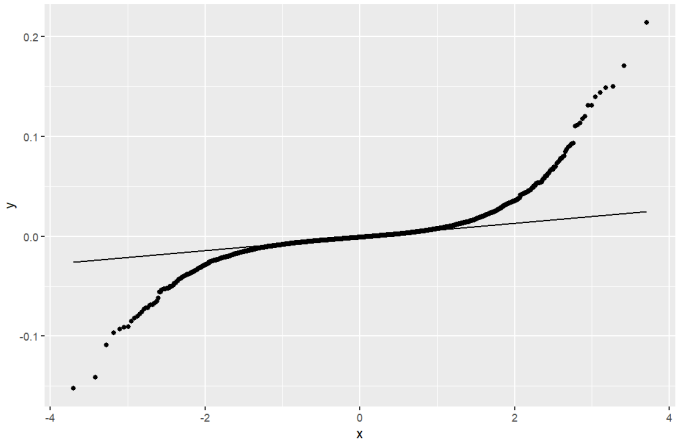


Figure 5.4 Quantile-Quantile plot of the residuals of SEM

6. Local Spatial Regression **(Wenxin Xu's Personal Part 21157882)**

Since the regression results of the global models still have some limitations, the article decides to try local models which take spatial heterogeneity into account.

In this part, the project first uses global Moran's I to test the spatial autocorrelation on each variable. Then, Geographically Weighted Regression (GWR) model is constructed to explain how the parameters vary spatially.

6.1. Spatial Autocorrelation Test

6.1.1. Methodology

Moran's I is a spatial autocorrelation method proposed by Patrick Alfred Pierce for exploring the spatial distribution of the study objects [19]. Among all the outputs of it, Moran's I index shows the degree of the spatial autocorrelation with a value range of [-1, 1]. The degree of correlation is judged by its absolute value and the positive and negative correlations are determined by the sign of it. P-value shows the significance, while Z-score is the standard deviation. The formula for Moran's I is shown as Equation 6.1.

$$\text{Moran's I} = \frac{n\sum_{i=1}^{n}\sum_{j\neq i}^{n}\omega_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}\sum_{j\neq i}^{n}\omega_{ij}(x_i-\bar{x})(x_i-\bar{x})} \quad (6.1)$$

Where n is the number of objects; ω_{ij} is the weight of the spatial elements i and j; $(x-\bar{x})$ are the deviation of the observed value from the mean value of the spatial units i and j.

6.1.2. Results

By using Moran's I tool in ArcGIS pro, the article checks whether the variables are spatially autocorrelated. The results are shown in Table 6.1.

Table 6.1 Moran's I of the variables

Variable name	Moran I	z-score	p-value
charging point	0.84	238.89	0
car per house	0.7	197.85	0
traffic flow	0.26	73.15	0
population dencity	0.48	137.31	0
house income	0.43	121.17	0
house price	0.5	142.69	0
public transport accessibility	0.59	167.59	0
substation	0.92	260.27	0
hospital	0.75	212.22	0
green space	0.19	55.07	0
residential	0.16	45.80	0
school	0.21	59.83	0

According to the table, all the variables show a high positive spatial correlation since their Moran's I are all greater than 0 with a significant P-value. Such results prove the necessity and validity of the GWR model which takes the spatial situation into account.

6.2. Local Spatial Regression with GWR

6.2.1. Methodology

GWR is a local spatial regression model. By implementing regression for each zone in an area, GWR explains the different effects of each variable on different positions [20]. Its formula can be expressed as Equation 6.2.

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{m}\beta_j(u_i, v_i)X_{ij} + \varepsilon_i \quad , i = 1, 2, \dots, n \quad (6.2)$$

Where y_i is the dependent variable; n is the total number of study areas; j is the number of

independent variables; (u_i, v_i) is the coordinates of the i^{th} study area's spatial location; $\beta_0(u_i, v_i)$ is the intercept of the i^{th} area; $\beta_j(u_i, v_i)$ is the k^{th} regression parameter estimate for the i^{th} area; X_{ij} is the value of the j^{th} explanatory variable and ε_i is the random error.

The estimation of β is calculated as Equation 6.3.

$$\hat{\beta}(i) = (X'W(i)X)^{-1}X'W(i)y \tag{6.3}$$

Where X is a matrix of explanatory variables, y is a vector of observations, and W(i) is a diagonal matrix of spatial weights which is constructed with the distances from each observation to position i [21].

6.2.2. Results

- **Coefficients Analysis**

The spatial distribution of the estimated coefficients of the explanatory variables is shown in Figure 6.1. According to it, the estimated coefficients differ significantly from region to region, which means that there is significant spatial heterogeneity in the study area.

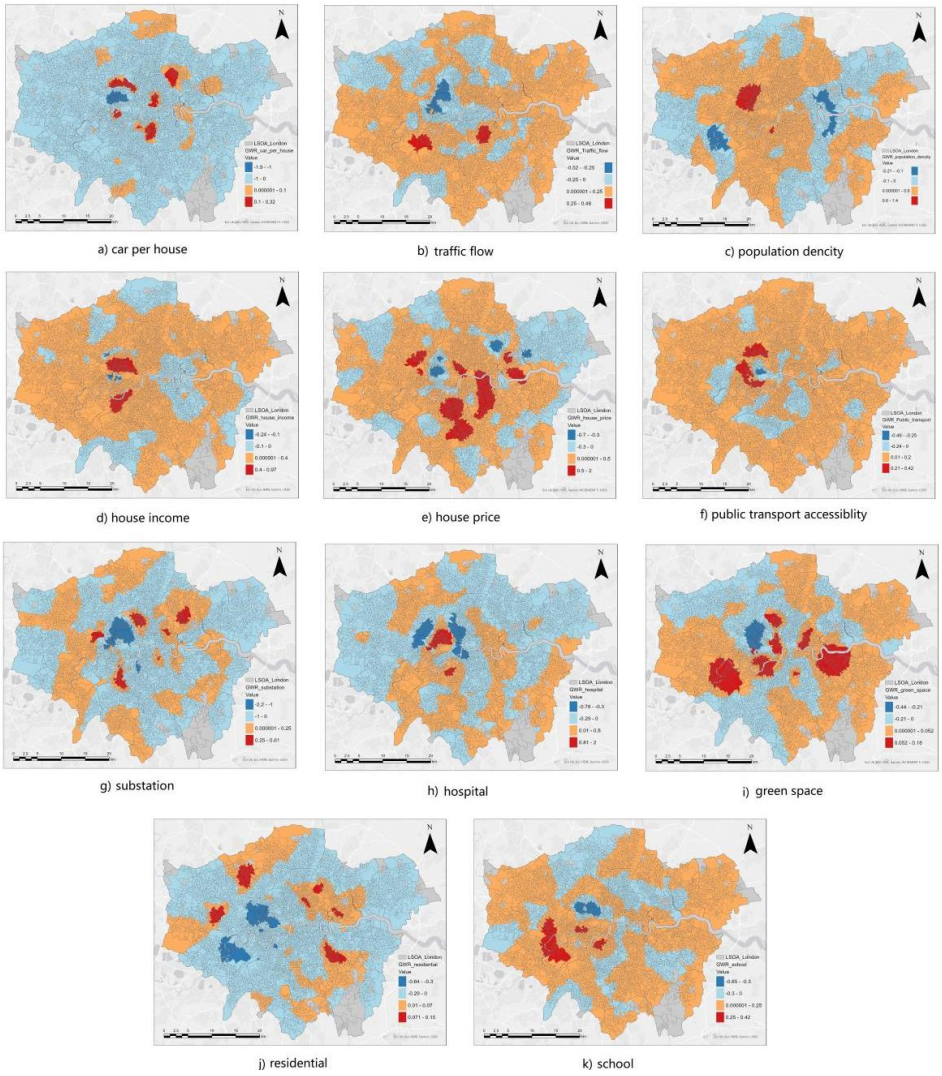


Figure 6.1 Estimated coefficients of the explanatory variables with GWR

- **Reliability Analysis**

The performance parameters of OLS, SEM, SLM and GWR are shown in Table 6.2.

Table 6.2 Parameters showing the fitting level of OLS, SEM, SLM and GWR

Criterion	OLS	SEM	SLM	GWR
R^2	0.4399	-	-	0.8787
AIC	-9227	-23754	-23754	-16125

According to the table, compared to the global regression models, although the overall R^2 of GWR is larger than that of OLS, its AIC is much bigger than that of SEM and SLM, which indicates GWR's poor fit to our dataset.

Figure 6.2 shows the distribution of R^2 in the GWR model, which reveals the spatial heterogeneity of it. In the vast majority of London, the value is larger than 0.3. What's more, LSOA grids with value greater than 0.6 covers more than half of the area. However, there are still regions with R^2 less than 0.3 (rendered in white and light pink), which may be responsible for GWR's worse fitting performance.

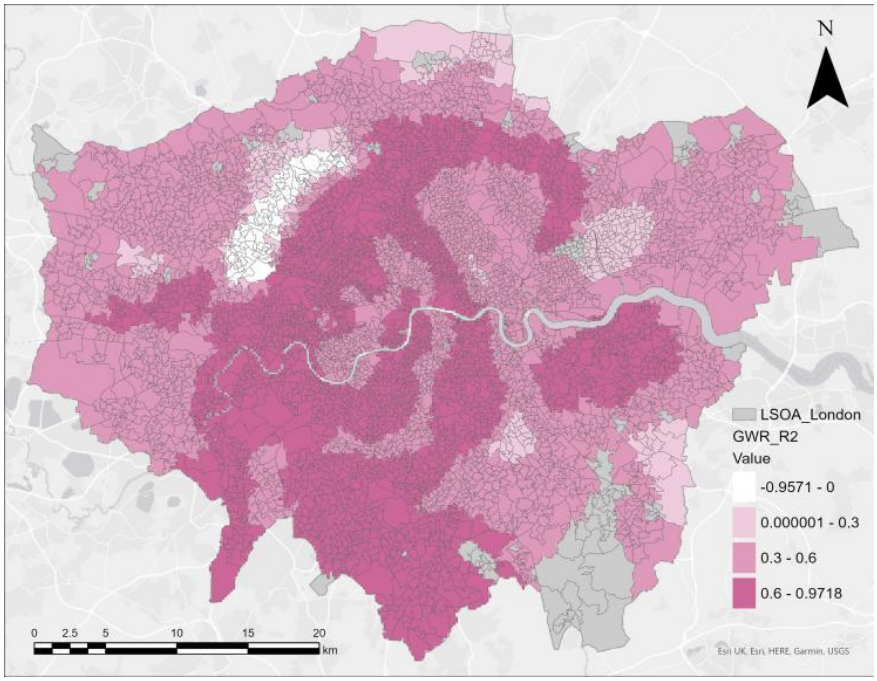


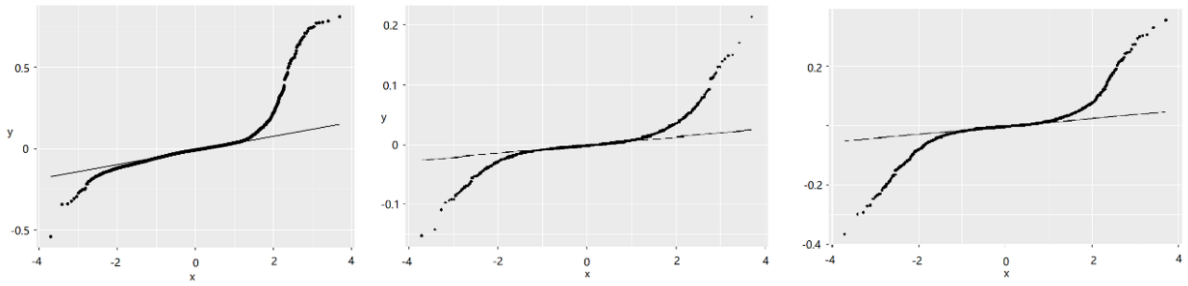
Figure 6.2 Spatial distribution of R^2 in the GWR model

The project also analyzes the autocorrelation among the residuals of GWR. Table 6.3 lists the Moran's I of GWR, OLS, SLM and SEM. According to the table, although the Moran's I of GWR (0.135) is much smaller than that of OLS, it's still bigger and more significant than that of SLM and GEM, which indicates the autocorrelation of the residuals can't be explained by data's spatial heterogeneity.

Table 6.3 Moran's I of OLS, SEM, SLM and GWR

Model	Moran's I	p-value
OLS	0.840	0.00007
SLM	0.050	0.0008
GEM	0.046	0.0009
GWR	0.135	0

In addition, by comparing the Quantile-Quantile plot (Figure 6.3) of the residuals of the regression models, it is obvious that the GWR's curve offsets more at the tails than that of spatial error model.



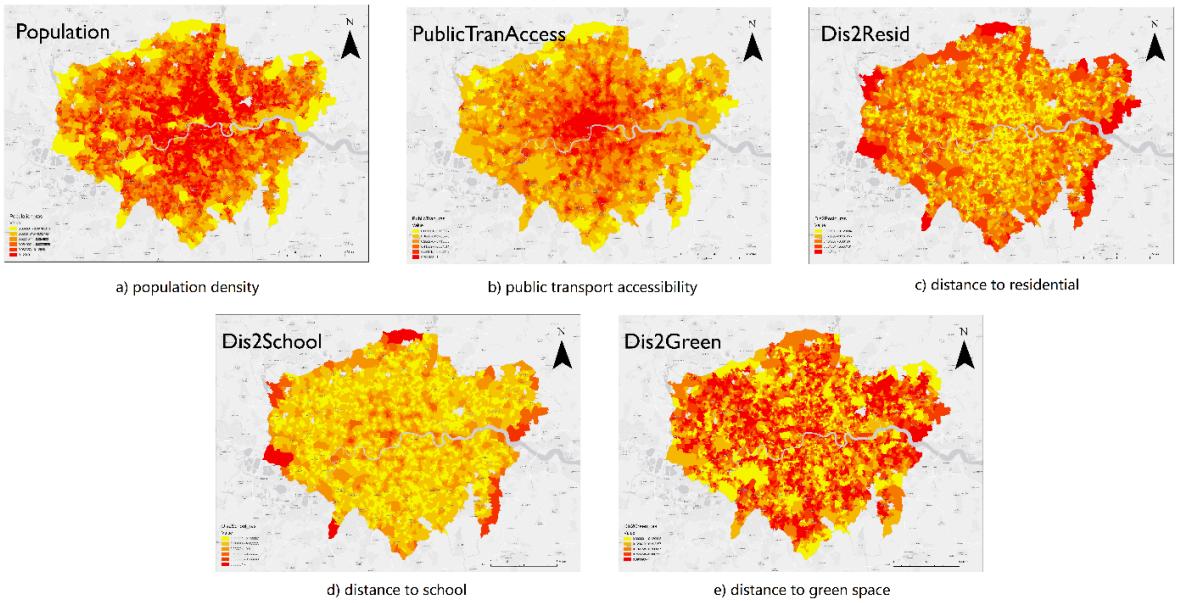
(a) OLS (b) SEM (c) GWR

Figure 6.3 Quantile-Quantile plot of the residuals of OLS, SEM and GWR

In conclusion, local spatial regression model does not fit our dataset as well as the global spatial regression models.

7. Weights Determination for Site Selection (Yunhao Zhu's Personal Part)

The final step for this project is to determine the weights for site selection. Considering the aim of our project is to select the best location for charging points with multiple affecting criteria, AHP's approximate method is the optimal method for the analysis. In general, processes include polygon to raster conversion, weight calculation, consistency test, and weighted overlay. The transformed raster layers are shown as Figure 7.1.



a) population density

b) public transport accessibility

c) distance to residential

d) distance to school

e) distance to green space

Figure 7.1 Transformed raster layers

7.1. Methodology

Analytic Hierarchy Process (AHP) is the most widely used method for multicriteria problem solving and analysis. In detail, it splits up the complicated decision-making problem and compares their affecting criteria pairwise. Individual weights are classified into different levels of intensity, and associated to the criteria that define the overall goal. For example, if we have two criteria called C_j and C_k, by creating matrix that calculates corresponding eigenvector and weighed summation, the overall weights for the alternatives are obtained. In this case, the levels of intensity are classified and converted into numeric values a_{jk} in the matrix. The overall weights are then computed using the formula below [22].

$$(Overall\ weight\ of\ alternative\ i) = \sum_j \begin{pmatrix} Weight\ of\ alternative\ i\ with\ respect\ to\ C_j \\ *\ weight\ of\ C_j\ with\ repect\ to\ the\ goal \end{pmatrix} \tag{7.1}$$

To perform AHP in ArcGIS Pro, several preparations need to be made. First, all the data should be converted to raster layers. Then, AHP calculation is performed to get the weight based on the significant results from the regressions model, which are population, public transportation accessibility, distance to residential area, schools and green space. Finally, raster calculator is used to stack up all the layers with their weights multiplied. The final output is a raster layer containing the scores that indicates priority of installing additional charging points for the London area. When converting into raster layers, a cell size setting of 90 is used, which is the half of the default size. This allows the cell size to be much smaller than a LSOA zone, so that projection of value will not be affected. During the process of AHP, data is directly used in

raster calculators instead of reclassifying for the data normalized into range of 0 to 1 in the previous steps.

7.2. Results

7.2.1. Weight Calculation Results

First, level of intensity has to be assigned to each criterion with reference from Saaty's pairwise comparison scale: population density, Euclidean distance to residential area, public transportation accessibility, distance to schools and green spaces. The scale is listed in Table 7.1. In this case, the importance ranking for the five criteria are: Population > Distance to Residential Area > Public Transportation Accessibility > Distance to Schools > Distance to Green Spaces.

Table 7.1 Saaty's pairwise comparison scale of AHP

Verbal judgment	Numeric Value
Extremely important	9
	8
Very Strongly more important	7
	6
Strongly more important	5
	4
Moderately more important	3
	2
Equally important	1

Thus, an AHP table is created with relative importance filled, in which each cell represents the relative importance between the two criteria on the row heading and column heading. From the rankings determined previously, the table is filled as shown below in Table 7.2. For example, the Population(row) - Dis2Resid(column) cell can be explained as "Population is moderately more important than Distance to Residential", and thus the Dis2Resid – Population cell should be filled with the reciprocal of the previous value.

Table 7.2 AHP table of the criteria

AHP	Population	Dis2Resid	PublicTran	Dis2School	Dis2Green
Population	1.0000	3.0000	5.0000	7.0000	9.0000
Dis2Resid	0.3333	1.0000	3.0000	5.0000	7.0000
PublicTran	0.2000	0.3333	1.0000	2.0000	5.0000
Dis2School	0.1429	0.2000	0.5000	1.0000	3.0000
Dis2Green	0.1111	0.1429	0.2000	0.3333	1.0000

After normalizing the previous table and calculate the average for each row, the priority value

(weight) for each criterion is determined, as shown in Table 7.3 below. We can verify the calculation by adding up all the weights and get a result of 1.

Table 7.3 Priority value (weight) for each criterion

Normalize	Population	Dis2Resid	PublicTran	Dis2School	Dis2Green	priority/ weights
Population	0.5595	0.6415	0.5155	0.4565	0.3600	**0.5066**
Dis2Resid	0.1865	0.2138	0.3093	0.3261	0.2800	**0.2631**
PublicTran	0.1119	0.0713	0.1031	0.1304	0.2000	**0.1233**
Dis2School	0.0799	0.0428	0.0515	0.0652	0.1200	**0.0719**
Dis2Green	0.0622	0.0305	0.0206	0.0217	0.0400	**0.0350**

7.2.2. Consistency Testing Results

The next step is to test the consistency. The approximate method in AHP requires the matrix to have a very low inconsistency in data. Consistency Ratio (CR) is calculated by Consistency Index (CI) / Random Index (RI). In this case, RI for 5 criteria is 1.12, according to Saaty (1980). From the equation for calculating CI:

$$C.I = \frac{\lambda_{max} - n}{n - 1} \tag{7.2}$$

Where λ_{max} is the average of the weighted sum calculated from the normalized fields for each criterion. From calculation, the result of CR is 0.0438383, which does not exceed the cap of 0.10. Therefore, this AHP model is usable and reliable. Hence, the AHP processes can proceed to the final step of raster calculation.

7.2.3. Raster Calculator Results

In the raster calculator, all the required layers are simply multiplied by their assigned weights and added to get the result. The only notable point is that for these two criteria: distance to school and distance to green spaces, they have a negative relationship with charging points' location from the previous regression results, while the other criteria "distance to residences" have positive relations. Considering this, the score of value for Dis2Resid should be inverted since charging points are more likely to be installed at low value areas (shorter distance from residential area), and the calculation for the other 2 distancing criteria remains the same (charging points are preferred to be installed at longer distance areas).

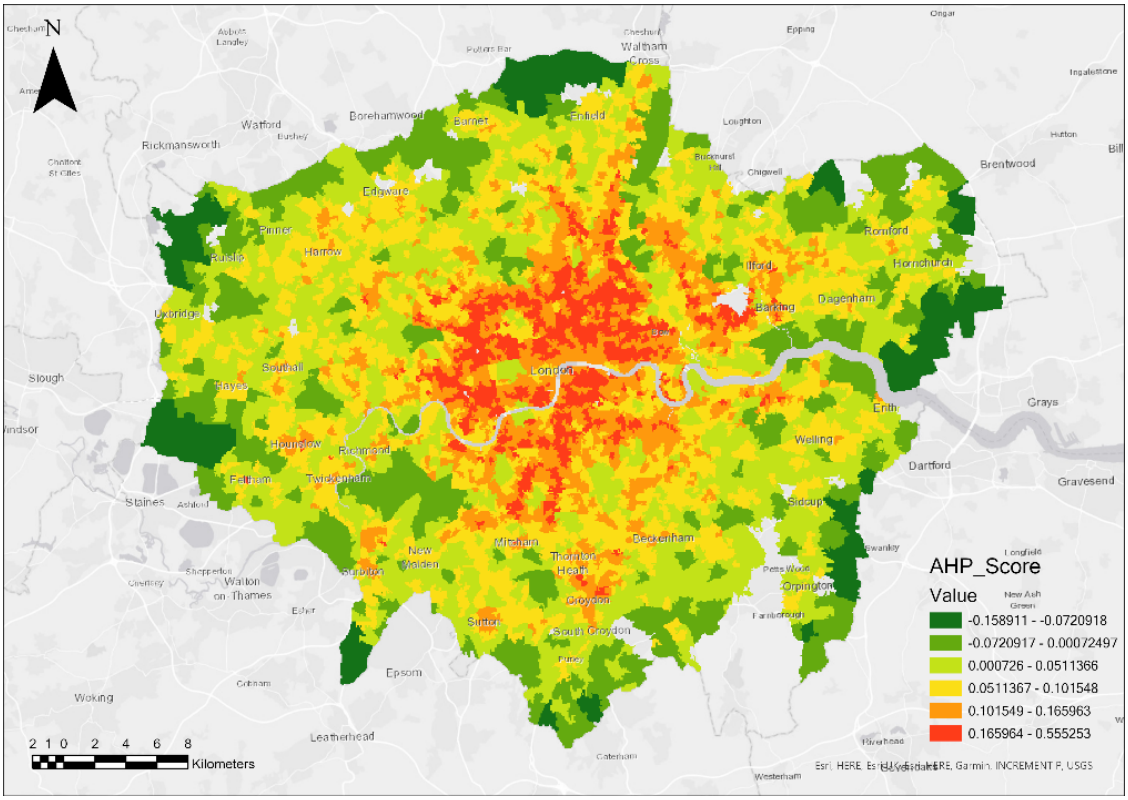


Figure 7.2 Final output raster containing the AHP scores

The final output raster containing scores is shown in Figure 7.2. The zones rendered in red and orange are the locations with highest priority in installing new public charging points based on those five criteria.

8. Discussion and Conclusion

With the results of AHP, the final task of this multi-criteria location selection is to find the most suitable places for installing public E-vehicle charging facilities. Thus, areas with highest scores from the AHP result must be determined and extracted. In this project, the threshold is set to 10%, which means the top 10% among all the areas are selected with the quantile reclassification method in ArcGIS Pro. The selected areas are then converted into polygon as shown below in Figure 8.1.

According to *London's electric vehicle charge point installation guidance* [23], it is a good choice to install public charging points on the street, and carriageway is an ideal place for these on-street charging facilities. Thus, to get a more specific result, the project first filters out the carriageways in London by excluding all the pedestrian paths, tunnels, bridges, motorways and the roads with a speed limit of 0 mile per hour. Then, the remaining carriageways are clipped with the most suitable areas that are shown in Figure 8.1. The result

is presented in Figure 8.2. To fulfill the required amount of charging posts by 2030 (which is 40,000 to 60,000), the London Government can give priority to installing new public charging points on the orange roads in Figure 8.2.

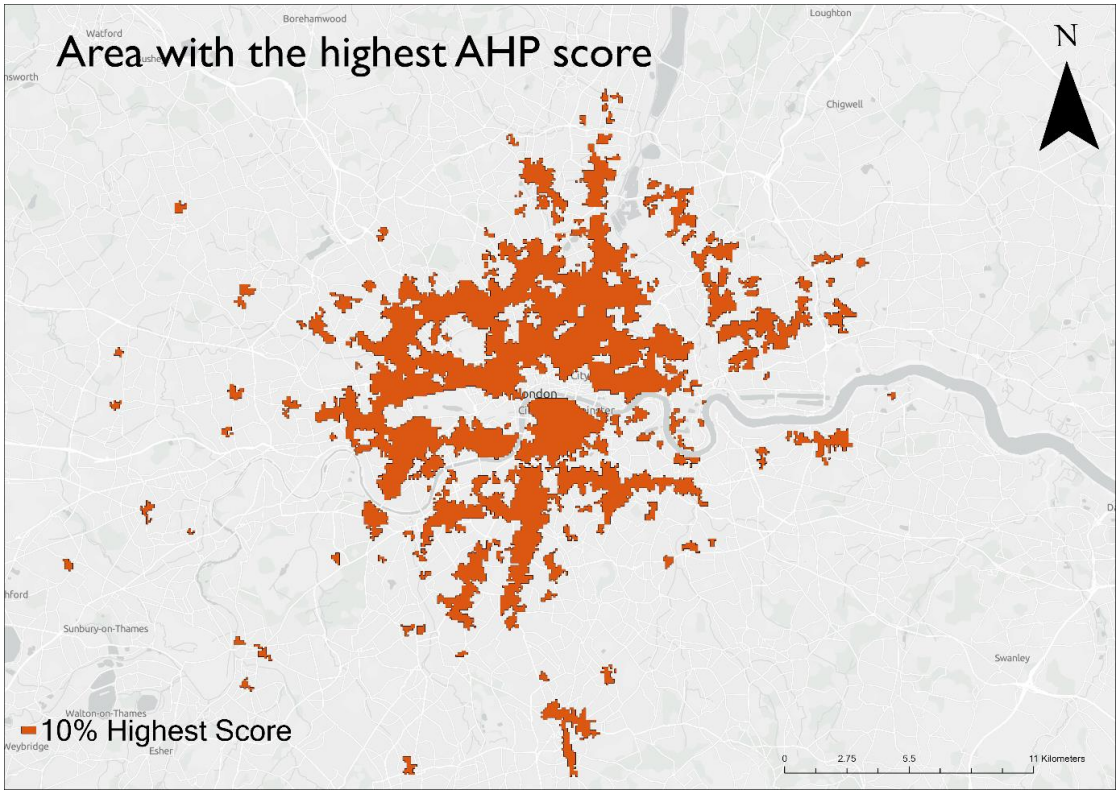


Figure 8.1 Areas with the highest (top 10%) AHP score

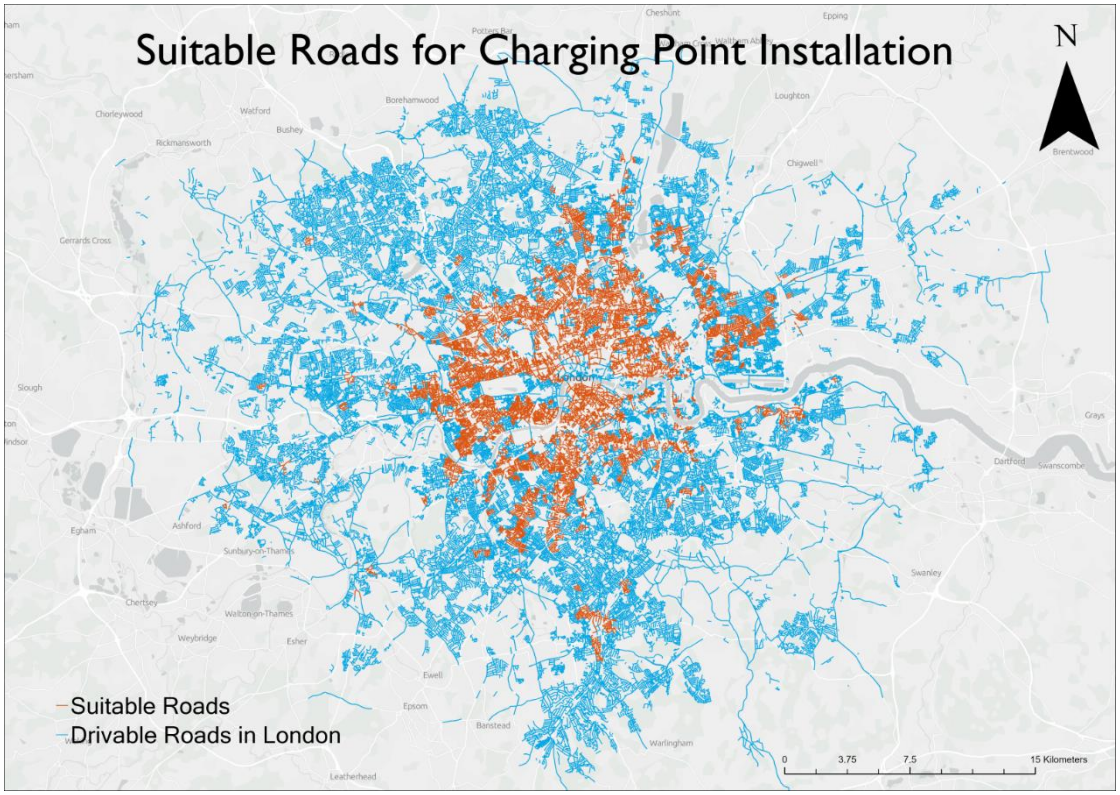


Figure 8.2 Roads suitable for charging points installation

In conclusion, by implementing data pre-processing and spatial regression with R and ArcGIS Pro, the project first analyses the factors affecting the existing position of public charging stations in London. Then, on the basis of the regression's results, the article uses AHP to select the most suitable areas for future installations. Although every step is performed deliberately with considerations in criteria selection, the scale of data, and choice of methodologies, limitations still exist in this project.

First, although OA's size and spatial density is better for more accurate location selection, considering its high demands on computer hardware for processing, the article chooses LSOA as the scale of data, which limits the accuracy of the results.

In addition, public charging points' location in real life are not only limited to roadsides but can also be in off-street spaces such as parking lots. Due to the limitation of data and detailed regulation on installing charging facilities for E-vehicles, more specific estimations on suitable sites can hardly be done. In future works, the project can possibly gain more information about the most common charging facility installation location by performing in-person field reviews in London, as well as record the most popular charging place among E-vehicle users, which may enhance the reliability of the results.

References

[1] Quinteros-Condoretty, A.R., Albareda, L., Barbiellini, B. and Soyer, A., 2020. A Socio-technical Transition of Sustainable Lithium Industry in Latin America. Procedia Manufacturing, 51, pp.1737-1747.

[2] Quinteros-Condoretty, A.R., Golroudbary, S.R., Albareda, L., Barbiellini, B. and Soyer, A., 2021. Impact of circular design of lithium-ion batteries on supply of lithium for electric cars towards a sustainable mobility and energy transition. *Procedia CIRP*, *100*, pp.73-78.

[3] IEA. 2021. Net Zero by 2050 – Analysis - IEA. [online] Available at: <https://www.iea.org/reports/net-zero-by-2050> [Accessed 16 January 2022].

[4] Transport for London. 2022. Electric vehicles and charge points. [online] Available at: <https://tfl.gov.uk/modes/driving/electric-vehicles-and-rapid-charging#on-this-page-3> [Accessed 16 January 2022].

[5] Yang, J., Liao, B.J., Wang, X.L., Wen, F.S., Zhang, X.Z. and Wang, L., 2015. Planning of charging facilities of electric vehicles based on geographical zonal charging demand coefficients. Electric Power Construction, 36(7), pp.52-60.

[6] Kaya, Ö., Tortum, A., Alemdar, K.D. and Çodur, M.Y., 2020. Site selection for EVCS in Istanbul by GIS and multi-criteria decision-making. Transportation Research Part D: Transport and Environment, 80, p.102271.

[7] Xu, Z., Xu, C., Hu, J. and Meng, Z., 2021. Robust resistance to noise and outliers: Screened Poisson Surface Reconstruction using adaptive kernel density estimation. Computers & Graphics, 97, pp.19-27.

[8] Shafabakhsh, G.A., Famili, A. and Bahadori, M.S., 2017. GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. Journal of traffic and transportation engineering (English edition), 4(3), pp.290-299.

[9] Kafi, K.M., Barau, A.S. and Aliyu, A., 2021. The effects of windstorm in African medium-sized cities: An analysis of the degree of damage using KDE hotspots and EF-scale matrix. International Journal of Disaster Risk Reduction, 55, p.102070.

[10] Desktop.arcgis.com. n.d. Zonal Statistics—Help | ArcGIS for Desktop. [online] Available at: <https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/zonal-statistics.htm> [Accessed 16 January 2022].

[11] Allen, M.P., 1997. The problem of multicollinearity. Understanding regression analysis, pp.176-180.

[12] O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & quantity, 41(5), pp.673-690.

[13] Seber, G.A. and Lee, A.J., 2012. Linear regression analysis (Vol. 329). John Wiley & Sons.

[14] Kelejian, H.H. and Prucha, I.R., 2010. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of econometrics, 157(1), pp.53-67.

[15] Fang, C., Liu, H., Li, G., Sun, D. and Miao, Z., 2015. Estimating the impact of urbanization on air quality in China using spatial regression models. Sustainability, 7(11), pp.15570-15592.

[16] Anselin, L. and Bera, A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. Statistics textbooks and monographs, 155, pp.237-290.

[17] Sakamoto, Y., Ishiguro, M. and Kitagawa, G., 1986. Akaike information criterion statistics. Dordrecht, The Netherlands: D. Reidel, 81(10.5555), p.26853.

[18] Crabtree, C., Darmofal, D. and Kern, H.L., 2015. A spatial analysis of the impact of West German television on protest mobilization during the East German revolution. Journal of Peace Research, 52(3), pp.269-284.

[19] Zhang, Q., Shen, J. and Sun, F., 2021. Spatiotemporal differentiation of coupling coordination degree between economic development and water environment and its influencing factors using GWR in China's province. Ecological Modelling, 462, p.109794.

[20] Kashki, A., Karami, M., Zandi, R. and Roki, Z., 2021. Evaluation of the effect of geographical parameters on the formation of the land surface temperature by applying OLS and GWR, A case study Shiraz City, Iran. Urban Climate, 37, p.100832.

[21] Mollalo, A., Vahedi, B. and Rivera, K.M., 2020. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the total environment*, *728*, p.138884.

[22] Kostagiolas, P., 2012. Managing Intellectual Capital in Libraries: beyond the balance sheet. Elsevier.

[23] Lruc.content.tfl.gov.uk. 2019. London's electric vehicle charge point installation guidance. [online] Available at: <https://lruc.content.tfl.gov.uk/london-electric-vehicle-charge-point-installation-guidance-december-2019.pdf> [Accessed 16 January 2022].