Spatial Econometrics - Project

GIS boundary analysis of Greater London

Author: Michele Guderzo

Course Coordinator: Dr Hab. Prof. Ucz. Katarzyna Magdalena Kopczewska

Warsaw, Academic Year 2021/22

Contents

1 Abstract

This project analyses and highlights the relationship that exist between the boroughs and the city centre of Greater London. After plotting and guessing a possible dependency between the area of certain boroughs and their distance to a certain coordinate (more generally, their position in space), two models were calculated which, although not completely correct, were quite satisfactory, especially in certain aspects.

2 Introduction

2.1 Background

The city of London, as we know it today, was significantly different in past centuries. Over the years, London has undergone a number of changes that altered its appearance not only in terms of government, but especially geographically.

Until 1889 the only part of London that had an administrative existence apart from the historic counties was the historic City of London, which was confined to the area of the medieval city. During the period 1889–1965, the County of London, carved from parts of the historic counties of Middlesex, Surrey, and Kent, administered an area that comprised present-day Inner London plus the outer boroughs of Newham and Haringey. The 1889 boundaries had been adopted in response to the rapid development of suburban areas in the 19th century. By the mid-20th century, however, the suburban population of London had spread far beyond the boundaries of the County of London. In an attempt to address that shift, the present boroughs were established in 1965 by amalgamating several existing boroughs and districts, at the expense of the surrounding counties, to form the new metropolitan county of Greater London.

The present-day City of London covers an area of 2.9 square km at the heart of Greater London and is a centre of world finance. Greater London forms the core of a larger metropolitan area (with a proportionately larger population) that extends as far as 70 km from the centre.

2.2 Aim

The aim of this project is to analyse the geographical situation of modern London, paying particular attention to the shape and conformation of boroughs. It is of interest to study whether and how the structure of today's boroughs can be influenced by their geographical location.

3 Data

The data was retrieved from the official London Datastore website (`https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london`), looking specifically for shape data. The data consists of a range of key GIS boundary files for ESRI and Map Info covering Greater London, including:

- *Output Area* (OA) of 2011;
- *Lower Super Output Area* (LSOA) of 2004 and 2011;
- *Middle Super Output Area* (MSOA) of 2004 and 2011;
- *London Wards* (two files: City of London merged into single area and split into seperate wards);
- *London Boroughs*.

However, for the purpose of my analysis, I only considered the variables *London Wards* (the one with City of London merged into single area) and *London Boroughs*, leaving out the others.
There were no changes to the original dataset before importing the data.
The data analysis was conducted with RStudio software.

4 Analysis

4.1 Exploratory analysis

After installing (if necessary) and loading the required libraries, I proceeded to import the data, making the appropriate transformations in order to read them correctly.
First of all, in order to have an overview, I made an initial plot of the graphs of the London wards and boroughs and compared them.

Figure 1 shows how, in the left-hand graph, the wards tend to thicken in the city centre, which is not evident in the right-hand graph (boroughs). On the other hand, the wards plot is confusing: it has too many subdivisions that are not immediately comprehensible, hence I decided to focus on the boroughs plot. This visualisation, however, is primitive and does not give us much insight into the data. Thus, I proceeded to use the *choropleth* function to provide clarity and more information, by choosing *hectares* as the study variable.

As can be seen from Figure 2, now we can understand more from the data. In addition, it appears that the boroughs with a larger area are those that are at a greater distance from the city centre. I then proceed to test this hypothesis by performing a spatial analysis.

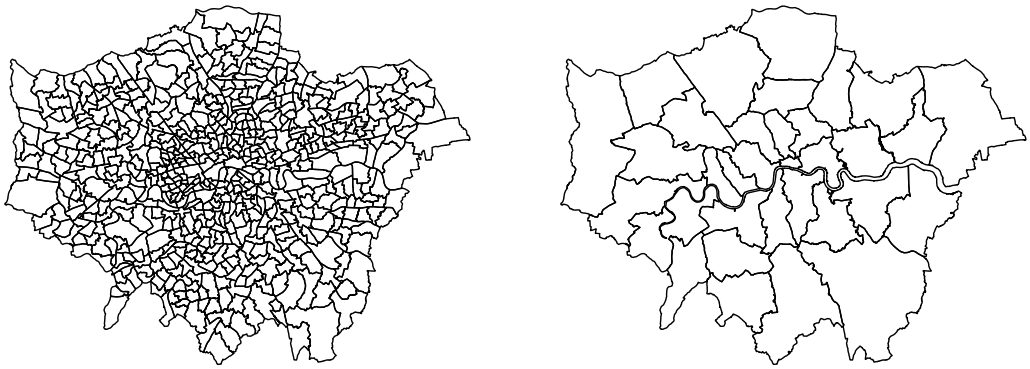

Figure 1: London wards (left) vs London boroughs (right)

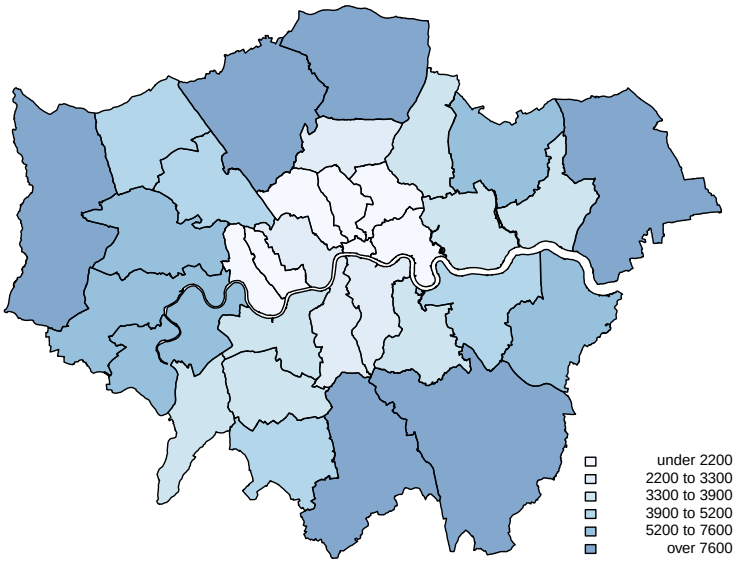

Figure 2: Hectares of London boroughs

4.2 Spatial analysis

First, I calculate the centroids of each borough, adding the name of the respective borough near the centroid.

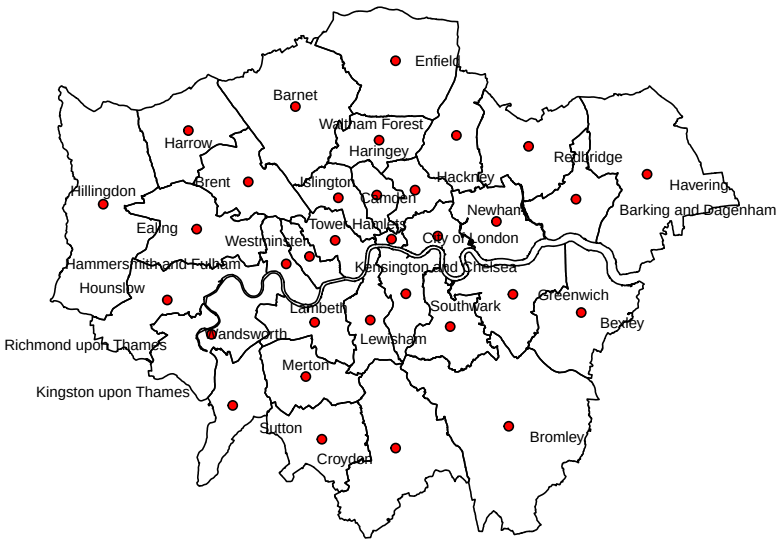

Figure 3: London boroughs with centroids

At this point, taking *City of London* as a reference, I calculated the distances between the centroids of the individual boroughs and the centroid of the central borough (i.e. City of London). I then compared the distances obtained with the area (hectares) of each borough, summarising the results in a scatter plot.

According to Figure 4, it seems evident that there is a certain relationship between the area of the borough and its distance from the city centre. Therefore, in the following section I implemented a model that could analyse such a relationship.

5 Model(s)

5.1 Estimate

The relationship in question appears to be a so called "direct effect" (the explanatory variable x, the distance, in location i, the borough, affects y, the hectares, in the same location i). Therefore, it seems sufficient and appropriate to estimate this link with a simple OLS model.

The first model I found was a simple linear dependency model between x and y without the intercept (which was removed because not significant). The model is as follows:

$$HECTARES = \beta_1 dist + \epsilon \tag{1}$$

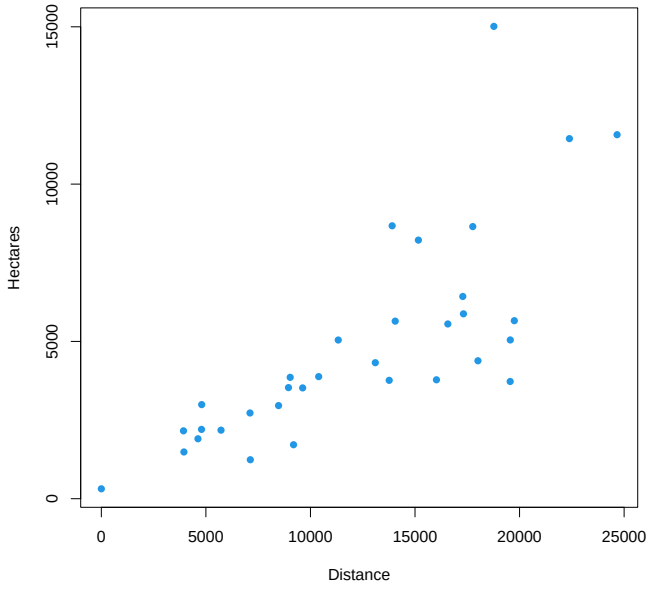

Figure 4: Distance to the city centre vs Hectares of London boroughs

This model, however, did not produce very satisfactory results (which I will discuss in the next section). Therefore, I decided to implement a second model with the logarithm of the dependent variable and the intercept, calculated as follows:

$$log(HECTARES) = \beta_0 + \beta_1 dist + \epsilon \tag{2}$$

Analysing the *summary* of these two models, it can be seen that (1) produces better results: the R^2 index is higher (87% vs. 69%) and both the p-values of the t-test and the F-statistic are lower than (2), implying that there is more evidence against the null hypothesis of non-significance.
After that, I performed diagnostic tests to understand the correctness of the models found.

5.2 Diagnostic tests

5.2.1 RESET test

The Ramsey Regression Equation Specification Error Test (RESET) test is a general specification test for the linear regression model. In our case, (1) gives a value of 0.5752. This means that the null hypothesis cannot be rejected, in other words, the model is correctly specified. Instead, (2) gives 0.0072, a value that leads to the null hypothesis being rejected. In this case, model (2) is incorrect.

5.2.2 Breusch-Pagan test

The Breusch-Pagan test is used to determine whether or not heteroscedasticity is present in a regression model. Here (1) gives an error because, in order to perform the test correctly, at least the intercept (which is missing) and a regressor are required. (2) gives a p-value of 0.1279 and, being greater than 0.05, we cannot reject the null hypothesis of homoscedasticity of the residuals.

5.2.3 Jarque-Bera test

The Jarque-Bera test, a type of Lagrange multiplier test, is a test for normality. It measures if sample data has skewness and kurtosis that are similar to a normal distribution. The best model turns out to be (2) since, with a p-value of 0.06881, it leads to the acceptance (not rejection) of the null hypothesis of normality of the residuals. (1), on the contrary, strongly rejects H_0 having a p-value of 1.008e-06.

5.2.4 Akaike Information Criterion

The Akaike information criterion (AIC) is a method for evaluating and comparing statistical models. It provides a measure of the quality of estimation of a statistical model by taking into account both the goodness of fit and the complexity of the model. Again, AIC of model (2) is significantly lower (41.34 compared to 601.62). This leads us to prefer the logarithmic model with the intercept.

6 Results and Conclusions

As already mentioned, this project aims to assess the dependency relationship between the area of London boroughs and their distance from the city centre. To this end, after visualising the data properly, two models were implemented to study this relationship in more detail.

It is not easy to determine whether the first or the second model is better. On the one hand, the model with only linear dependence offers a higher significance of the variables (lower p-values), a better goodness of fit and a correct specification of the model. On the other hand, the logarithmic one allows for normally distributed and homoscedastic residuals, with a much lower AIC value. Although not completely correct, both models can be used for further analyses, it is just a question of what matters most (the summary or the diagnostic tests).

With a view to future implementations, it would be interesting to investigate the issue of spatial correlation, thus estimating a more precise spatial model.