Understanding the spatial distribution of supermarkets in London: accounting for demographics and competition

Candidate Number: GTMT3

Course: STAT0035

word count: 12533

May 6, 2020

Abstract

This report aims to investigate the spatial distribution of supermarkets in London in the use of London supermarket locations, London census information and London station locations. Based on the combination of multitype point process models and computational application of analysis techniques such as quadrat tests, the Clark-Evans test and Monte Carlo tests of summary functions like the K-function including various edge correction approaches. Models are fitted with various covariates in order to determine which factors might influence the spatial distribution of supermarkets and their possible interpretations. Interactions between supermarkets are also tested.

Contents

1 Introduction

Retail sector refers to all companies or individuals who sell products directly to consumers, including supermarkets, internet retailers, department stores, market stalls and door-to-door sales people. In 2017, the retail sector contributed 92.8 billion in output and accounted for 5% to the UK economy. It also created 2.8 million job opportunities in the labour market, which was 9.5 percent of the UK total. Since the first supermarket opened in the USA in 1930, the supermarket has been one of the main components of the retail sector.

However, it is facing some challenges these years. Until August 2018, 28 companies with multiple stores have ceased the market, making the UK lose 2085 stores and 39000 jobs due to the change of consumer behaviour and challenging business environment. This is not only a loss for those companies but also a loss for the UK employment market. Additionally, the continuous increase in online sales in the UK squashed the market shares takes up by traditional supermarkets, making the situation of traditional supermarkets more difficult. In January 2008, online sales took up 5% of retail sales. A decade later, online sales took up 18% of all retail sales. Although internet sales show an increasing trend recently, physical store shopping is still the majority. The location of supermarkets might not play an as important role as before, but it is still a pivotal part to maximize their profit and achieve business success.(1)

Retailers might be interested in these kinds of location investigations to open a store closer to their target consumers and analysis the business strategy of their competitors. The understanding of the location choices could also potentially benefit property companies or landowners, by lowering their opportunity cost to find a tenant.

The study of making location decisions by retailers could be traced back to the work of Haig.(2) After that, many types of research studied supermarket locations from different angles. Mohammed and Bernd(3) proposed the way in which supermarkets make their location and quantity decision to minimize transportation fees and inventory fixed costs. Replenish inventory from warehouses and travel from consumers' homes to supermarkets would lead to carbon emission. Dilek(4) focused on the location of supermarkets under considerations of the carbon tax schemes. Brown(5) gave a review of retail location theory with some recent new advances and drew people's attention to not much considered areas. A more up-to-date book written by Birkin(6) focused on the UK retail location planning. This book analyse the history and current situation of UK supermarkets, geodemographics in retail planning and other techniques used. Papers or books mentioned previously are mostly focused on the theoretical while article written by Simkin(7) proposed a realistic model - Store location assessment model (SLAM), which is used by several UK's major retailers in many fields. But it is less likely for us to deduct how reliable these studies are. Supermarkets would not publish the way of the location was chosen since it is vital for them.

Spatial statistics techniques, not restrict on supermarket location, but also widely used in many other areas. Nicolis(8) applied them in forecasting the location and intensity of seismic events while Silva(9) applied them in predicting the frontier expansion of the Amazon rainforest. Besides the environment area, spatial science also contributes to biology yield. For example, analyse the interaction relationship between epidermal nerve fibers(10) and cell biology.(11) And some combine spatial statistics technique in studying sociology like crimes.(12)

2 Exploratory data analysis

Considering a large amount of data for the whole UK requires strong computing power, which laptops would not be able to complete. And the UK is a big country; it is difficult to analyse the supermarket locations in the whole country due to the different development levels and consumer behaviour. Given that the population in London is just over 10 percent of the whole UK, London census data in London is based on a larger population than other cities. London is highly urbanization, so it does not have too much mixture of the rural and urban area and this makes our analysis more accessible. Combining of these, we take a close look at the supermarkets locations in London where is the capital and the largest city of the UK instead of the UK.

2.1 The supermarket data

Retail store location data in the UK is found online from a website called Geolytix.(15) The data containing 18 categories, including names of the retailers, fascia (petrol filling station, motorway service area or else) longitude and latitude projection of stores as well as their sizes (classified into four classes), etc. in WGS84 format on June 2019. There are a total of 14014 stores in the UK and 19 main retailers like Tesco, Sainsbury's, the Co-operative Food, etc. The total area of the UK is 242,495 square kilometres so a store could be found around 17.3 square kilometres. There are 1696 supermarkets located in London, Tesco owns the most retail points (467) while the second largest retailer is Sainsbury's (361), followed by the Co-operative Group (228), Marks and Spencer (173) and Iceland (124). According to London's area, one store is occurring with approximately 1.07 square kilometres. The average density of supermarkets in London is much higher than that in the whole UK, probably because of the higher population density in London. The location of Tesco gives us the most information so Tesco would be the first to focus on the following of the analysis. We would expect that these supermarkets have their tendency. For example, Tesco is a supermarket with a variety of items while Iceland is focusing on frozen foods.(13) We would expect Tesco and Iceland have different target areas. So Iceland is also what we are going to explain in detail.

2.2 Spatial supermarkets distribution overview

Different retailers are represented by different types of points in Figure 1. The outline of Figure 1 is the outline of London, which is downloaded from online.(14) Supermarkets tend to cluster in the centre of London. Supermarkets are less concentrated on the edges compares with the centre. It might because of the higher population density in the centre leads to higher demand for supermarkets.

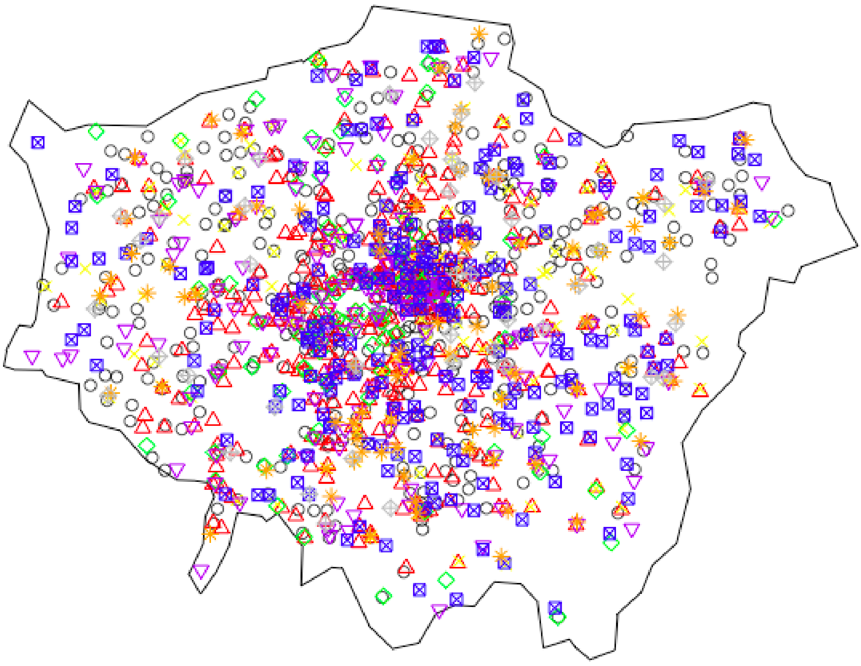
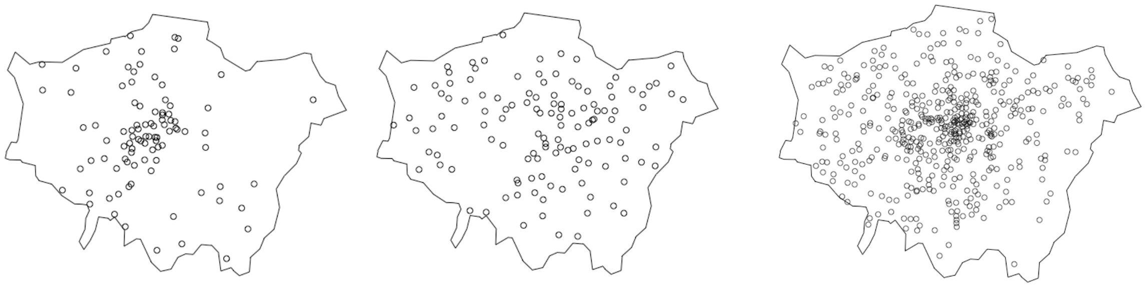


Figure 1: supermarket overview

Figure 2 shows the spatial distributions of Waitrose, Iceland and Tesco supermarkets. Waitrose is clustering at the centre area, and this might because Waitrose sells relatively expensive items and targets consumers with higher income. But Iceland is more spread out across London. Iceland does not ignore the eastern and the western areas as Waitrose does. Iceland is focusing on affordable food so it has a broader consumer group. In order to engage with its consumer, Iceland is less likely to group together. Tesco nearly covers the whole London area with clustering at the centre. This could because the property price in the centre is high. So Tesco opens many small stores in the centre but a little large store at edges.



(a) Waitrose distribution

(b) Iceland distribution

(c) Tesco distribution

Figure 2

2.3 Census information

In order to model why supermarkets may be more likely to be located in particular regions of London, the location of supermarkets is not enough. The demographic data of London is needed. The census information which summarizes the census population, housing, crime, poverty, employment, health, education, etc. in 2011 is used provided by Greater London Authority.(16) The UK Census is undertaken each decade in all parts of the UK, giving statistics support for departments of national and local governments to decide how to plan and allocate resources. The newest posted Census data is on 27 March 2011. Information like population density, income per household, number of cars per household, etc. might be useful because supermarkets could use these to determine their parking space, size of supermarkets, pricing strategy and check if their location decisions are reasonable.

The census gives the measurement values, but the region outlines are presented in a different dataset with different scales. By combining census and boundaries, maps of covariates would be produced, which is used in the computing model later. We worked on the scale of Middle Layer Super Output Areas (MSOA), which gives us enough detailed data, but not so detailed that it is difficult to analyse. There are 32 boroughs in London and each of them is divided into several middle super output area and this is what "MSOA" refers to. London is divided into 983 MSOA and boundaries of each MSOA is given.(16) The mean population of a MSOA is 7200 and the minimum population of a MSOA is 5000 (Figure

19 shows how the MSOA divide London).(17) Other scales like Borough, Ward contains far less information while the "LSOA" contains too much information. "LSOA" stands for Lower Layer Super Output Area, which partition London into a smaller scale. The mean population of a LSOA is 1500 and the minimum population of a LSOA is 1000.(18) LSOA is averaging over fewer people and computing greater amounts, perhaps trigger to less reliable data.

The use of "spatstat"(19) package in R combined the boundary data for each of MSOA to produce a map of supermarket locations across London. Datasets used are not very up-to-date but they are useful enough. Visualisation and discussion of these covariates is provided later in this report in Section 4.5.

2.4 London transportation information

Due to the expensive parking fees, congestion charges, traffic jams, lack of space and probably environmental awareness, public transport is very popularized in London. Hence, supermarkets might prefer to locate closer to these transportation links. The tube station locations are founded online(20) in latitude and longitude format. In order to transfer these points into covariate to feed into the latter modelling process, distances from every possible location in London to the nearest tube station are calculated. We could make some improvements by including not only tube stations but also bus stations and use walking distance instead of straight-line distance. To find a balance between complexity and accuracy, the use of 652 tube stations is a compromise and it is useful enough here. (Detail plot in Figure 19 (c))

2.5 Coordinate and measurement system

In order to combine all of the data sources into a single analysis, data should be transferred from different coordinate systems (WGBS84, a unit of longitude and latitude) to the unit of a kilometre, which is a ration scale and easier to interpret. Let us take longitude and

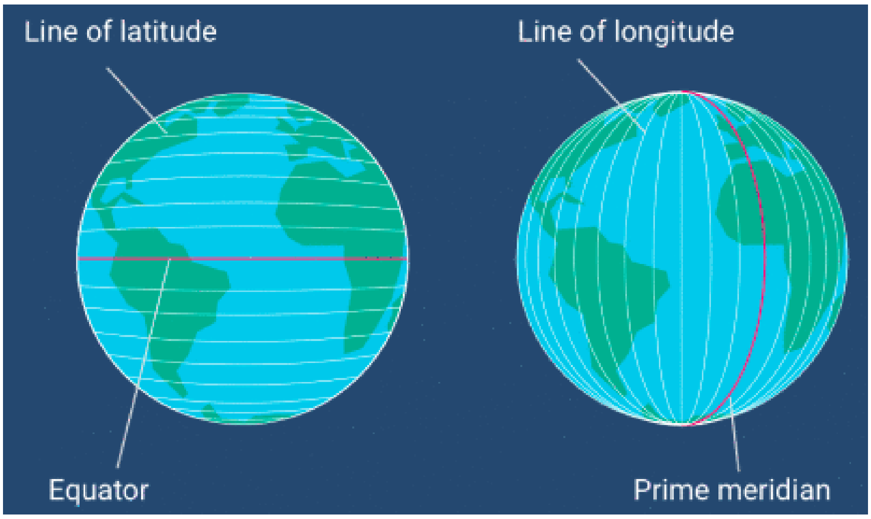


Figure 3: latitude and longitude (21)

latitude as an example. On the surface of the earth, every single point can be represented by the latitude and the longitude coordinates. Latitude is defined as the angle between the straight line of a specific point and the equatorial plane. The equator is a line of latitude with 0 degrees. The calculation of the distance between latitude is easy because it is fixed. The circumference of the earth is around 40075 kilometers. Therefore, one degree of latitude is $\frac{40075}{2\times180} = 111.3$ kilometer. Longitude is defined as the angle between the line of longitude and the prime meridian. As shown in Figure 3, lines of longitude converge to the north and south pole. The closer to the equator, the more separate the lines of longitude are. So unlike latitude, the distances between longitude lines are various. The centre of London is set to be the origin of the system to ensure computations as accurately as possible. In London, the degree of longitude is approximately to $111.3\times\cos(latitude)$ which is $111.3\times\cos(51.509865) = 69.27$ kilometer. As we know, the earth is not perfectly spherical and there is some curvature of the earth. But because London is such a small region from the earth's surface, these curvatures would not be problematic.

3 Complete Spatial Randomness

3.1 Point processes and point patterns

The occurrence of events at random intervals, which relative to the time axis or the space axis modelling by stochastic processes are called point processes. Therefore, temporal point processes and spatial point processes are the two types of point processes. Point patterns are samples from point processes, in other words, sets of points in a certain area or set.(25)

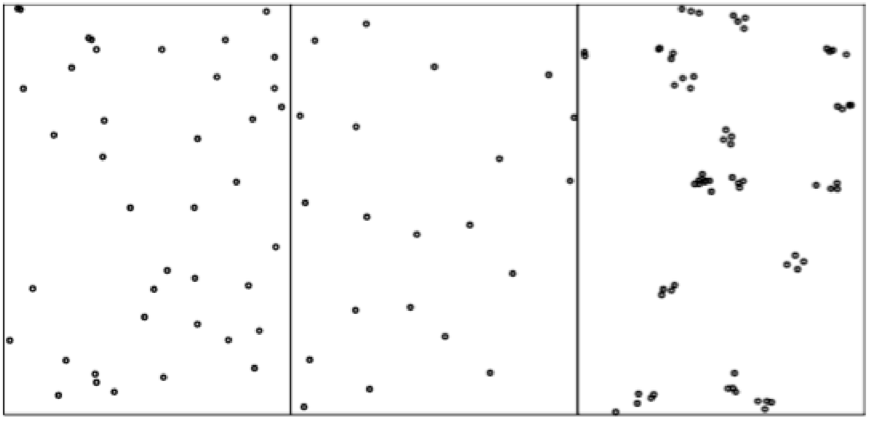


Figure 4: Point patterns example (25)

It is always useful to produce a graphical representation of point patterns to gain a rough understanding of our data. From left to right, they are random point patterns, regular point patterns and cluster point patterns. The left hand side figure could be regarded as a random point pattern because no apparent structure is showed (Figure 2(b) the Iceland distribution is also an example). The centre figure is an example of a regular point process. On the other hand, a strong cluster structure has shown by the right hand side figure, suggesting that explanation of this cluster would be required (Figure 2 (a) and (c) are also examples).

3.2 Homogeneous Poisson Point Process

Complete Spatial Randomness (CSR), also referred to as the Homogeneous Poisson Point Process, is a model under which

1. For any region C the number of points, $n(C)$, follows the Poisson distribution with expected value $\lambda \times a(C)$ for some constant intensity $\lambda > 0$ and $a(C)$ the area of the region C.

2. Point locations are independent of each other.

A consequence of these requirements is that points in the process are distributed uniformly and independently over C, hence the title of Complete Spatial Randomness. Homogeneous always relates to the uniformly proportion that the any one part of an overall dataset is the same as any other part.

The homogeneous point process has a single parameter, the intensity λ. In the 1 unit wide and 1 unit tall window C, the left picture of Figure 5 is a point pattern from a homogeneous point process with λ equals to 20 and the right picture of Figure 5 is a point pattern from a homogeneous point process with λ equals to 100. It is clear that there are less points in the left hand side pictures than the right hand side picture due to the lower intensity.

As mentioned before, point patterns are samples from point processes so point patterns do not necessarily have the same spatial distribution even they come from the same point pattern. Figure 6 shows two different sets of point patterns from the point process with intensity 50. The number of points is 63 in the left picture compares with the 38 in the right hand picture. 50 is the expected intensity but it does not require the number of points from point patterns are precisely 50.

We use the Homogeneous Poisson Point Process as the basis for building more complicated point process models. Thus, we should test if our data fits the CSR model or not. If not,

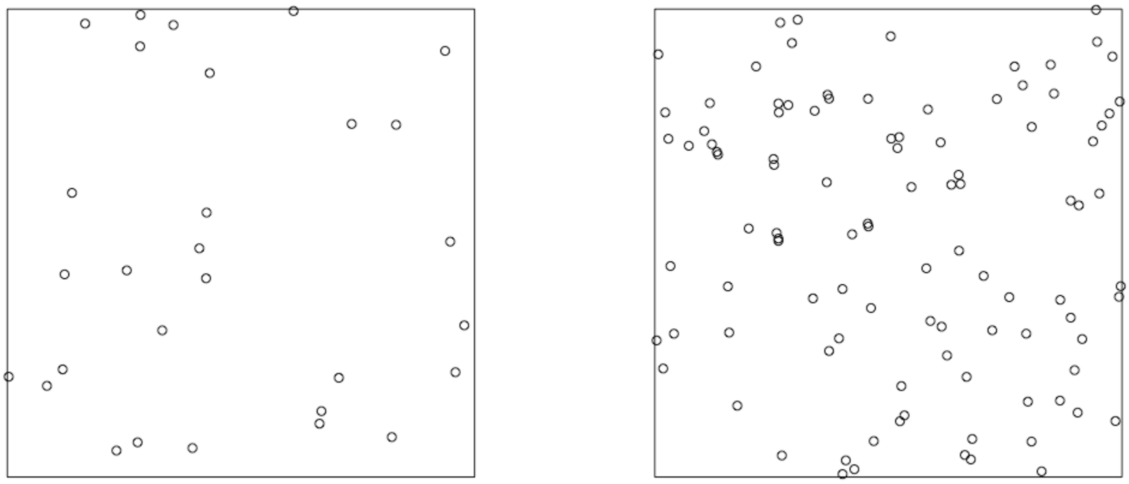


Figure 5: point patterns with different intensity

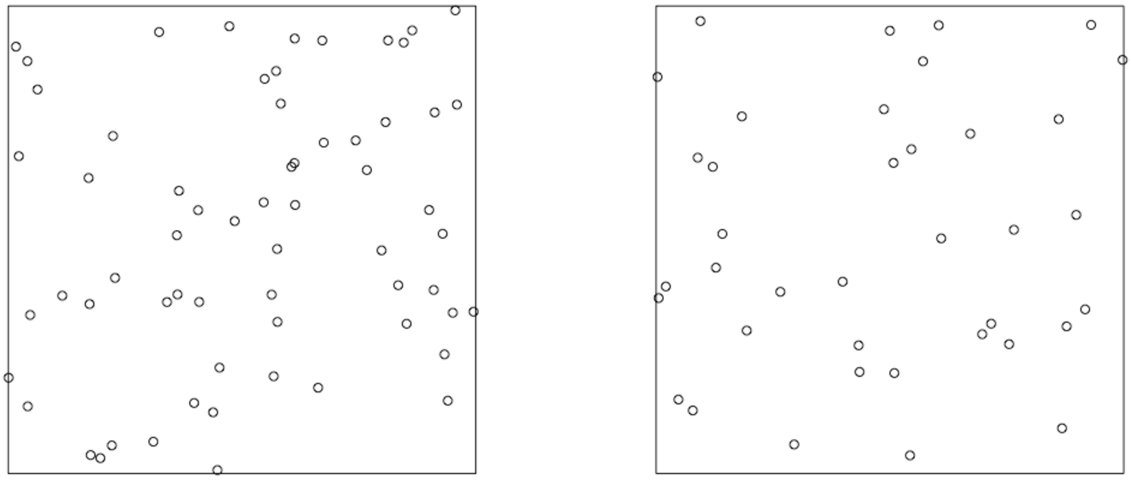


Figure 6: different point patterns from same intensity

whether it is independence inhomogeneous or dependence homogeneous Poisson process (See Section 4.1 and section 5 for more details). Although we may look at a pattern and think it looks homogeneous, our eyes may deceive us. We need to apply formal statistical approaches to test the CSR. The first such approach is the index of dispersion.

3.3 Index of dispersion

The number of points within the subregion area C_i should be Poisson distributed and its expected value should be proportional to its subregion area $a(C_i)$. The original window, C, is divided into m subregions C_i, or quadrats.

We could calculate the index of dispersion based on the counts of numbers of events in the m quadrats. Index of dispersion, as its name, is an index suggesting whether the data is acting in a “dispersion” way, a “clustering” way or randomness. The index of dispersion can be calculated by $s^2/\overline{n(C_i)}$ where s^2 is the sample variance of the $n(C_i)$ quadrat counts and the degree of freedom is m-1.

$$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(n(C_i) - \overline{n(C_i)})^2 \tag{1}$$

Because for Poisson distribution, the variance is exactly the mean. CSR suggests that the variance of the number of points is the same as its mean. If the index of dispersion is less than 1 the distribution of points might be ”dispersion” instead of randomness due to the relatively little variation. By contrast, for the index of dispersion bigger than 1, the distribution of points might be ”cluster” instead of randomness due to the relatively much variation.

The index of dispersion of the location of Tesco is larger than 1, which indicates the Tesco locations in London could be a cluster. However, the index of dispersion is not a formal test. So we would need to think about other formal tests to measure the dispersion level.

3.4 Chi squared test

The Pearson Chi-squared test is a widely used test to analyse the difference between expectations and actual observed data. We would like to find if there is a statistically significant dispersion between the observed counts n(C_i) and the expected counts. When C is divided into same area quadrats, the expected quadrats counts becomes the mean of $n(C_i)$. It also called quadrat test, with

$$\chi^2 = \sum_{i=1}^{m}\frac{(n(C_i) - \overline{n(C_i)})^2}{\overline{n(C_i)}} \tag{2}$$

When C is divided into different area subregions, the expected counts are the total number of points multiplied by the fraction of the total area made up by C_i.

$$n(C)\frac{a(C_i)}{a(C)} \tag{3}$$

This is what Chi-squared goodness-of-fit test can achieve. The two-side Chi-squared test refers to the hypothesis "there is no too large or too small deviation between the observed frequency and the expected frequency" while the one-side test refers to the hypothesis "there is no too large deviation between the observed frequency and the expected frequency". The occurrence of large deviation might be evidence for clustering and small deviations might be evidence for regularity. We are interested in any deviation so we do the two-side Chi-square test.

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(n(C_i) - n(C)\frac{a(C_i)}{a(C)}\right)^2}{n(C)\frac{a(C_i)}{a(C)}} \tag{4}$$

Then, compare the Chi-square statistic to that Chi-squared distribution. If the Chi-squared statistics is larger than tabular statistic from Chi-squared distribution, i.e., inside the rejection region, then H0 is rejected. P-value is always used according to the degree of freedom m-1 i.e., the maximum number of values that free to vary.

The London map is divided into 23 irregular windows by R automatically, which is easier for us. Because the approximation to the Chi-square distribution occurs so the expected counts should be bigger than 5, which is the case. Inside each subregions, the right top number represents the expected number of points inside this subregions, the left top number represents the actual number of points and the bottom number represents the variation level. The top right number is not the same due to the different area of each subregions.

From Figure 7, the various level is from -4.1 to 15. The number of stores is quite different in some subregions, such as 112 actual Tesco stores located in the center part of London compares with 31 expected stores. By rule of thumb, we expect the result from the Chi-squared test to say the actual frequency of Tesco location inside each subregion differs from the expectation.

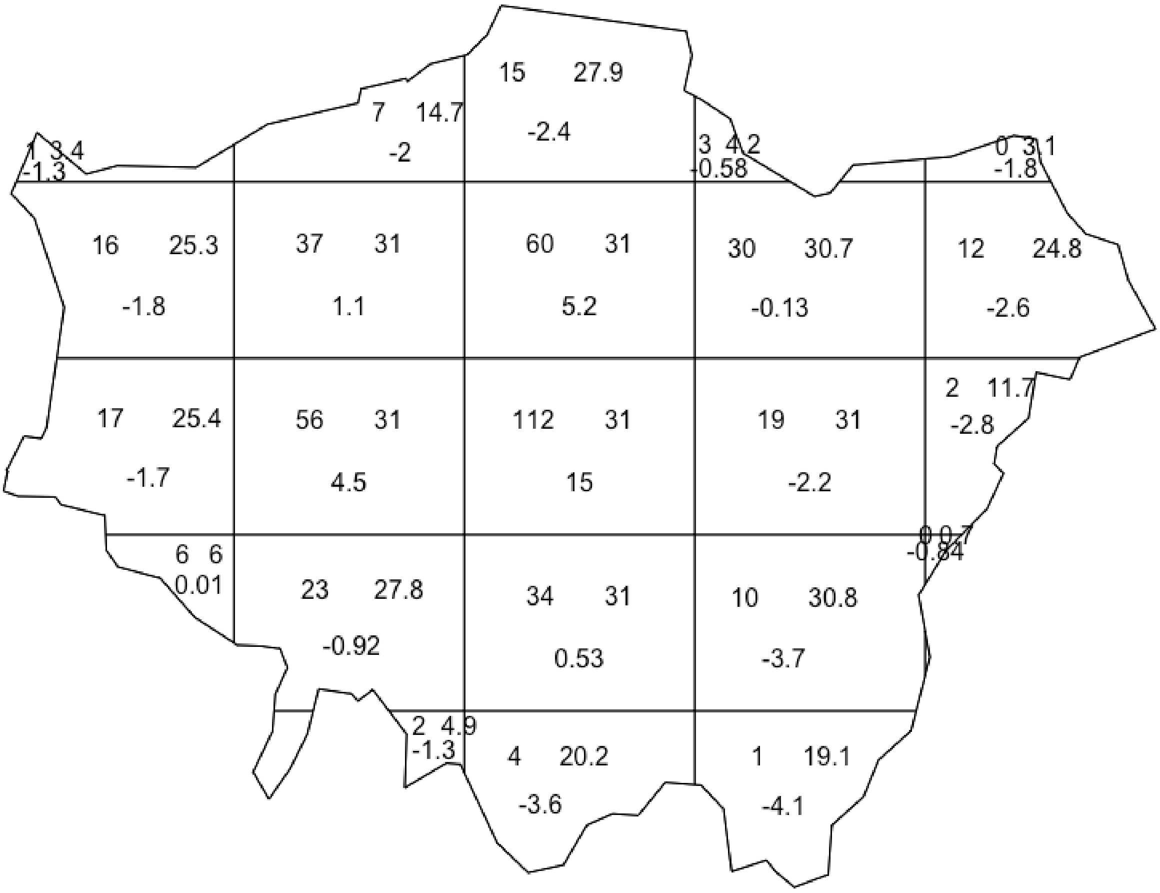


Figure 7: plot of Chi-squared test

The p-value, in this case, is extremely small (p-value $< 2.2 \times 10^{-16}$ this is the smallest p-value R can ever compute) under 22 degrees of freedom. It shows that the observed value deviates too much from the theoretical value, and we fail to accept H0. The Chi-squared statistics is too large here (15.81183× 22=347.8603) no matter under which significant level, H0 is always rejected. Thus, the locations of Tesco is not randomness as we expect.

However, the result of the Chi-squared test would be affected by many factors, such as the number of quadratic grids (m) or the different division methods of grids. Also, the Chi-squared test focuses only on the number of points in each grid but neglects the important information of inside grids. For instance, as shown in Figure 8, these two plots have an equal number of points in each quadrat. The result of the test should be the same while there is a clear cluster in the left hand side plot, which needs further consideration. Thus, it would be better if we could use a method to replace the area-based method.

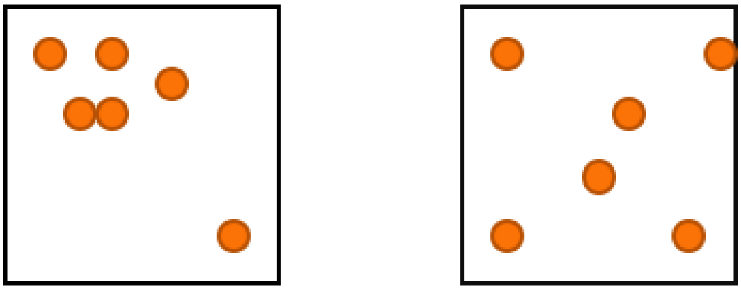


Figure 8

3.5 Clark-Evans test

The Clark-Evans test is a similar test but we use the nearest neighbour distance to replace subregions counts. The nearest neighbour distance is also known as nn-distance. For $n(C)$ events in C_r (a circle with radius r), let r_i denotes the distance from the ith event to the nearest other in C. The r_i is called the nearest neighbour distances.(28) As Figure 9 shown, r_i is the radius of the red circle. The distances from all the other points exceed r_i.

Under CSR, points follow a Poisson distribution with intensity λ and the expected number of points in a region C is $\lambda a(C)$. Fixed value r stands for the distance between points while r_i is the nearest neighbour distance. The probability that the nearest distance of point i is further than a distance r is the same thing as the probability that there are no points within a distance r of point i.

$$P(n(C_r) = 0) = exp(-\lambda \pi r^2) \tag{5}$$

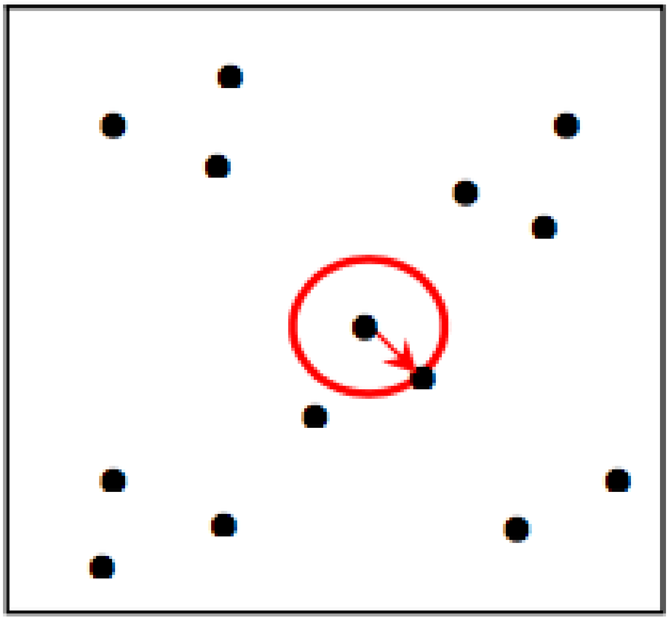


Figure 9: the nearest neighbour distance (24)

where the area of circle is πr^2 and C_r stands for a circle of radius r. Hence, the probability that there is a nearest neighbour of point i within a distance of r is

$$P(r_i < r) = 1 - P(r_i > r) = 1 - P(n(C_r) = 0) = 1 - exp(-\lambda \pi r^2) \tag{6}$$

R The mean therefore is $\frac{1}{2\sqrt{\lambda}}$ and the variance is $\frac{4-\pi}{4\pi\lambda}$. Central limit theorem refers to the sum of identically and independent variables tend toward a normal distribution. Due to the central limit theorem, we use normal approximation, and thus the mean stays unchanged while the variance should divide n(C). Therefore, we get

$$\overline{r_i} \sim N\left[\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4n(C)\pi\lambda}\right] \tag{7}$$

We use test statistics

$$Z_i = \frac{\overline{r_i} - (\frac{1}{2\sqrt{\lambda}})}{\sqrt{(4-\pi)/(4n(C)\pi\lambda)}} \tag{8}$$

compare with standard normal distribution $N(0, 1)$.

The average nearest neighbour distance for our Tesco data is 0.9046516, which is smaller than the theoretical average is 0.9290234. This suggests that Tesco is more cluster than we would expect. CSR assumes that the nearest neighbour distances are equally to be small or large across the whole region. But the following figure of Tesco distribution, colored by the value of the nearest neighbour distances, shows that the CSR assumption is not satisfied here. So this might be evidence of inhomogeneous.

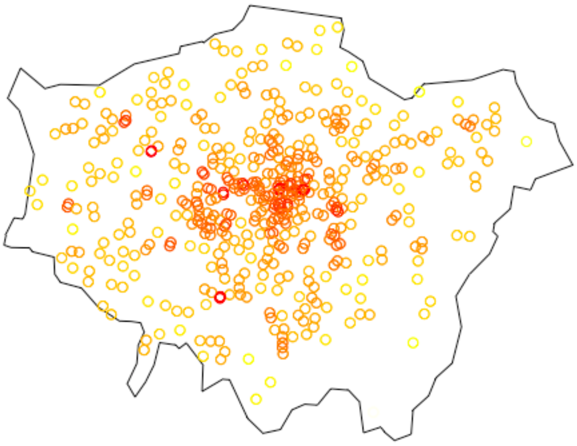


Figure 10

Similar to the Chi-squared test, the Clark-Evans test could be a two-tail test or a one-tailed test. A two-tailed test has two rejection regions (greater than and less than). We are interested in both the possibility of "is my nearest neighbour distance too big so my points are regularity or repulsion" or "is my nearest neighbour distance too small so my points are cluster". If we go back to Figure 8, we would see cluster (left) would cause smaller nearest neighbour distances while randomness or regularity would lead to greater nearest neighbour distances. A two-tailed Clark-Evans test is used.

The p-value for Tesco by doing the Clark-Evans test is 0.2781, which is relatively a large number. This suggests that we fail to reject the null hypothesis. The Clark-Evans test here suggests that the locations of London Tesco supermarkets are randomness, which is a contradiction of our results from the Chi-squared test. This might because the nearest neighbour distance should be independent so we assume the selection of a subset of the nearest neighbour distance contains no common points. In addition, the use of the central limit theorem assumes that the independent random samples are not too skewed. In Figure 11, the red line is the density function of the nearest neighbour distance of Tesco and bars are the histogram. Both of them show a positive skewness so the assumption of the central limit theorem is violent.

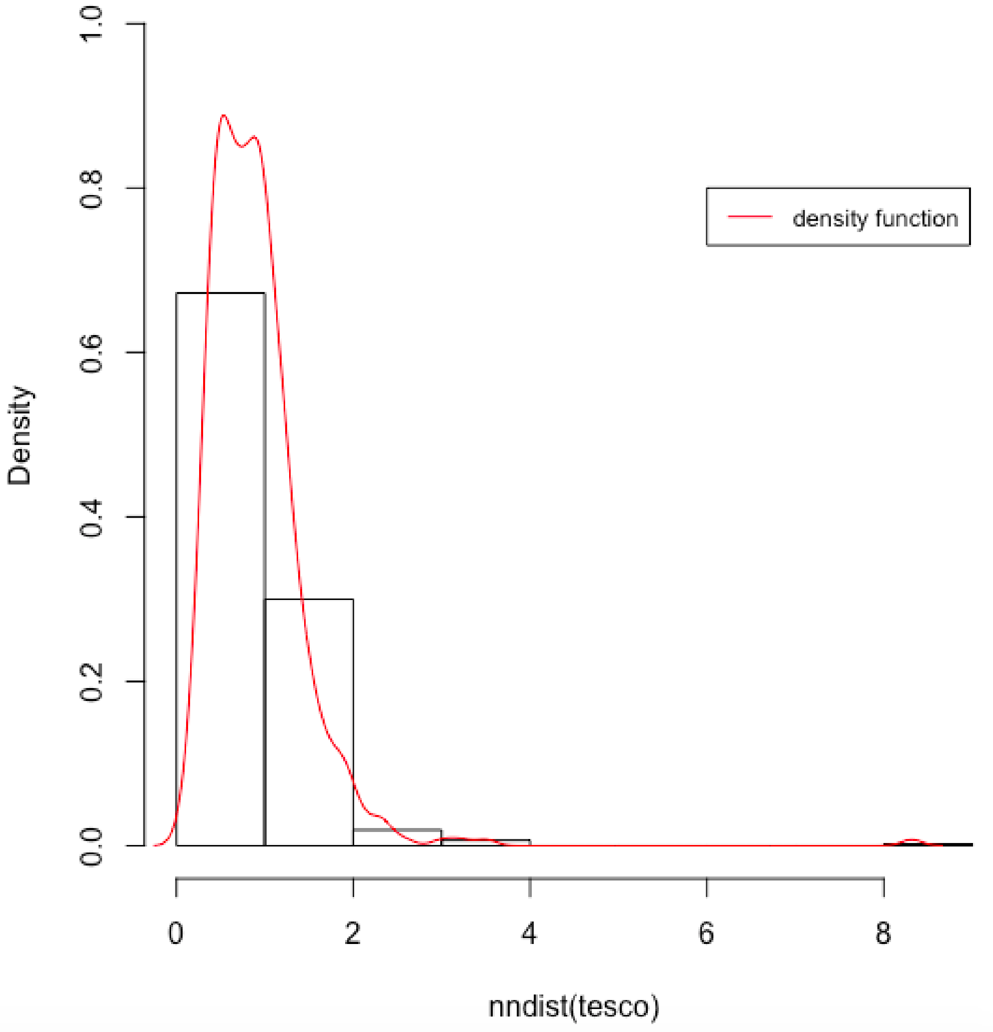


Figure 11: Histogram of Tesco

The Clark-Evans test takes the average of the nearest neighbour distance and loses some information by doing this. It might generally contain more information than the Chi-squared test but not in this case.

3.6 G-function

Rather than using mean, G-function, which is the comparison between the estimated cumulative density function of nearest neighbor distances and the theoretical cumulative distribution function of the nearest neighbour distances. In other words, G-function tells us the probability in a r radius circle, there is another point.

$$G(r) = E(number\ of\ pointswithin\ the\ circle\ of\ a\ radius\ r) = \lambda\pi r^2 \qquad (9)$$

For G-function larger than the theoretical, the actual nearest distance more positively skewed. For G-function lower than the theoretical, the actual nearest distance is more negatively skewed.

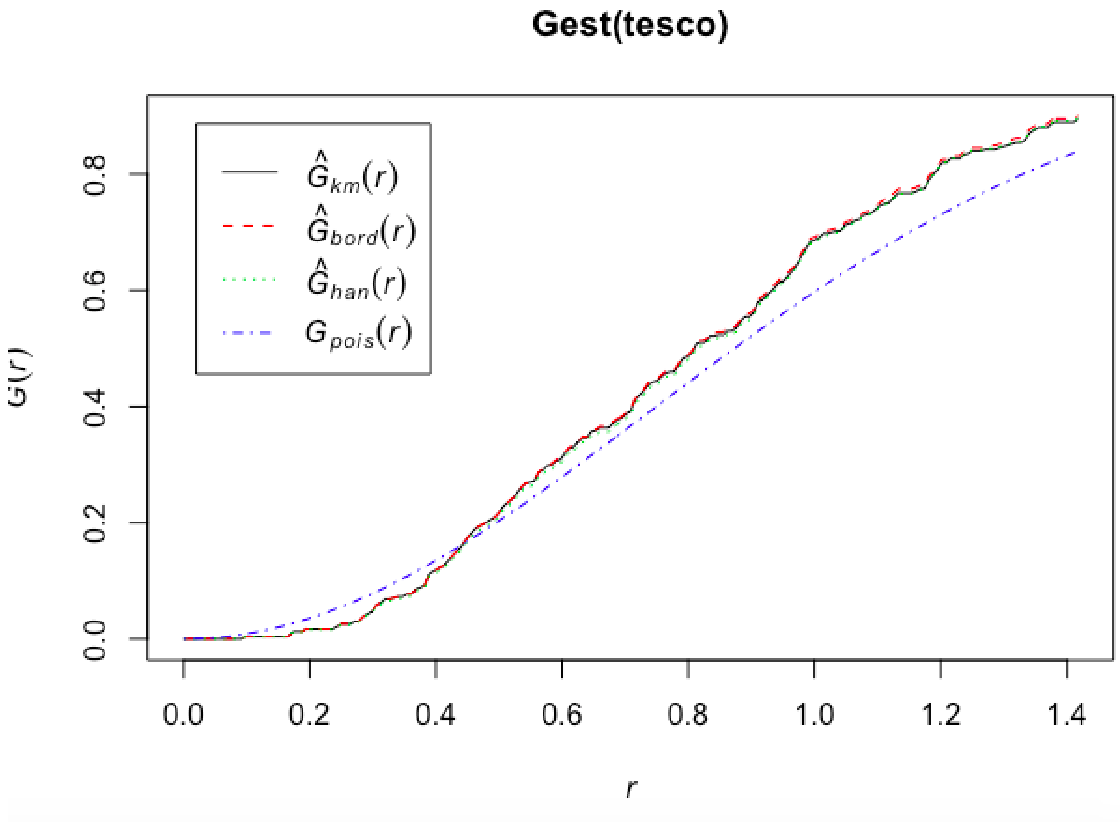


Figure 12: G-function of Tesco

From Figure 12, the blue line represents when data is simulated under the Poisson point process while the black line represents the true data. Red and green lines are nothing new but with different edge correction approaches. Edge correction is important because for points close to the edge of the window, their "true" nearest points might not be included in the window, leading to the incorrect calculation of the nearest neighbour distances. There is a 0.19 probability that other points are occurring with $r_i = 0.45$. Black line is under the blue line before the radius equals 0.45 and overtakes the blue line after 0.45. Tesco locations are more spread out before radius 0.45 and cluster after that. When the radius is between 0 and 0.1, there is no Tesco supermarket around Tesco, suggesting that supermarkets from the same brand avoid compete with each other. As the increasing of r, the $G(r)$ tends to 1.

We can see that the black line and the blue line is not coinciding, but because of the size of the window in C, it is difficult to analyse them in detail. We then apply the Monte Carlo test to construct "envelopes" for the G function.

3.7 Envelops (Monte Carlo test)

When typical hypothesis tests like the Chi-squared test and Clark-Evans test is about calculating test statistics and comparing it with a specific distribution to test the null hypothesis, spatial point patterns are mostly intractable distributed. So in spatial point patterns we could not find a tabled distribution to compare with. Instead, the Monte Carlo test is introduced. Monte Carlo test is a computational approach that simulates data by repeatedly randomising the data under nonparameter null hypothesis. Then, build up the distribution of test statistic accordingly.

Simulations with matching intensity under the homogeneous Poisson process as this is our null hypothesis. We took 39 simulations and one observation. So we have 40 different observations altogether. If things are randomly ordered, then the probability that our observation is at the highest simulation level is 1/40. The probability that our black line is at the very smallest simulation level is also 1/40. 1/40+1/40=5% which means we are testing the null hypothesis of the observed line lying outside of this shaded region at the 5 percent significant level.(23) For the pointwise graph (Figure 13), the black line (represents actual data) is not totally inside the grey region (represents the theoretical boundaries simulated under homogeneous Poisson point process). But this would not necessarily mean there is a big deviation to said data are not homogeneous Poisson process. When doing the pointwise envelopes, many tests are done with no fixed separation distance r. So the separation distance may vary from tests to tests. The very small region outsides the shaded region may not be convincing enough to against the null hypothesis.

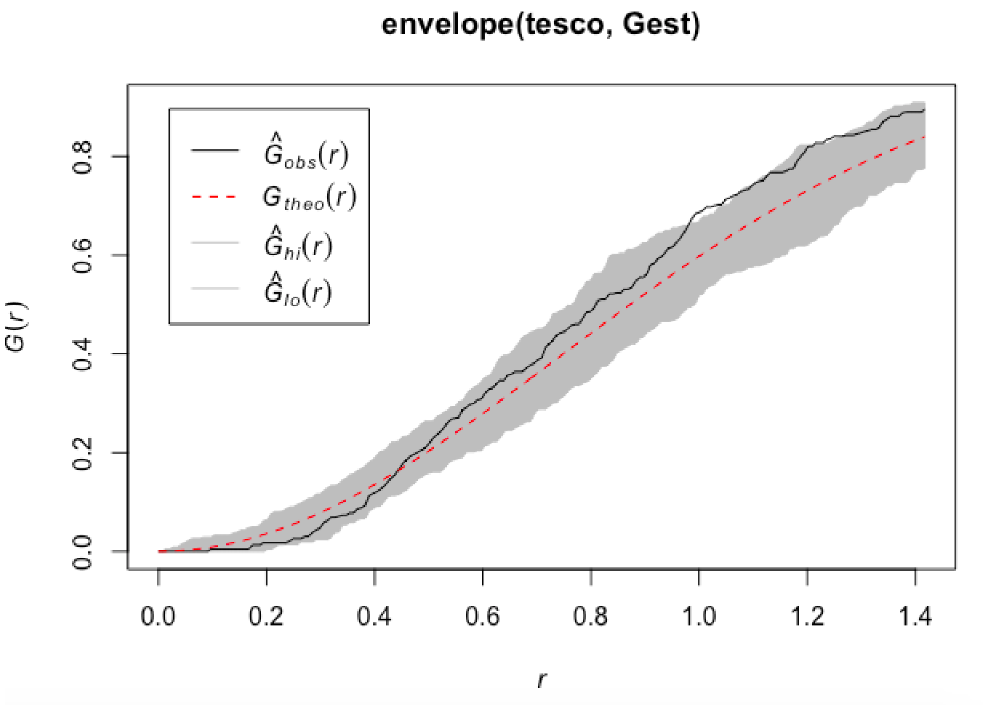


Figure 13: Tesco pointwise envelope with G-function

Alternatively, the global envelope method takes 19 simulations and has a fixed separation distance r, which is the maximum deviation from all the tests. It can be regarded as a single test, testing whether the observation data follow a homogeneous Poisson process. The null hypothesis is rejected if the observed (black) line is outside the shaded region. For our Tesco location data (Figure 14), the black line is entirely inside the grey region and there is no evidence against the null hypothesis.

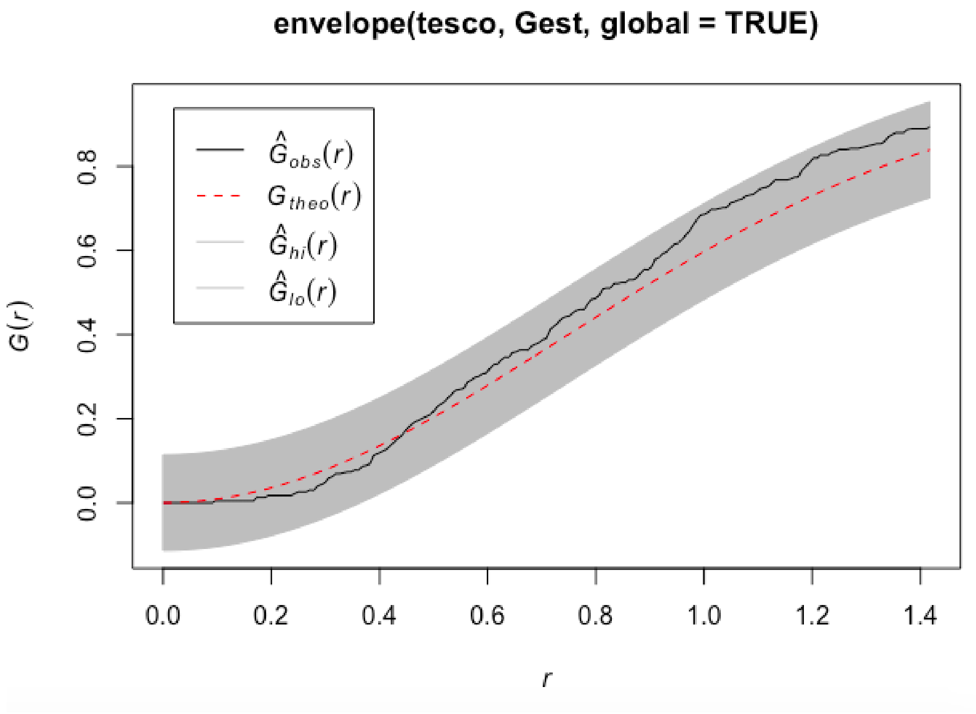


Figure 14: Tesco global envelope with G-function

This result contradicts our result of the Chi-squared test but agrees with the results of the Clark-Evans test. It may because of the loss of information by just looking at the nearest neighbour distance. Figure 10 showed that the spatial distribution of the nearest neighbour distances might not be a consistently homogeneous Poisson process. The consideration of the second, third or even more nearest neighbour distance possibly detects this discrepancy.

3.8 K-function

K-function considers distances from a point to all the other points. Recall from the nearest neighbour distance method, the nearest distances (r_i) are not independent. Under the assumption of no edge effect by using randomly sampled ordered of varying size instead of assuming independent subsets, K-function represents the expected number of points within distance d_i under fixed intensity λ. For the homogeneous Poisson point process,

$$K(r) = \frac{1}{\lambda} \times E(number\ of\ points\ within\ the\ circle\ of\ a\ radius\ r) = \frac{\lambda \pi r^2}{\lambda} = \pi r^2 \quad (10)$$

The global envelopes plot with K-function is dramatically different from G-function. It shows that the actual location of Tesco is far different from the expected Poisson point process. This suggests that the location of Tesco is inhomogeneous. The difference outcomes might cause by the shorter nearest neighbour distance are all in the middle and longer nearest neighbour distance are all over the outside. (See Figure 10)

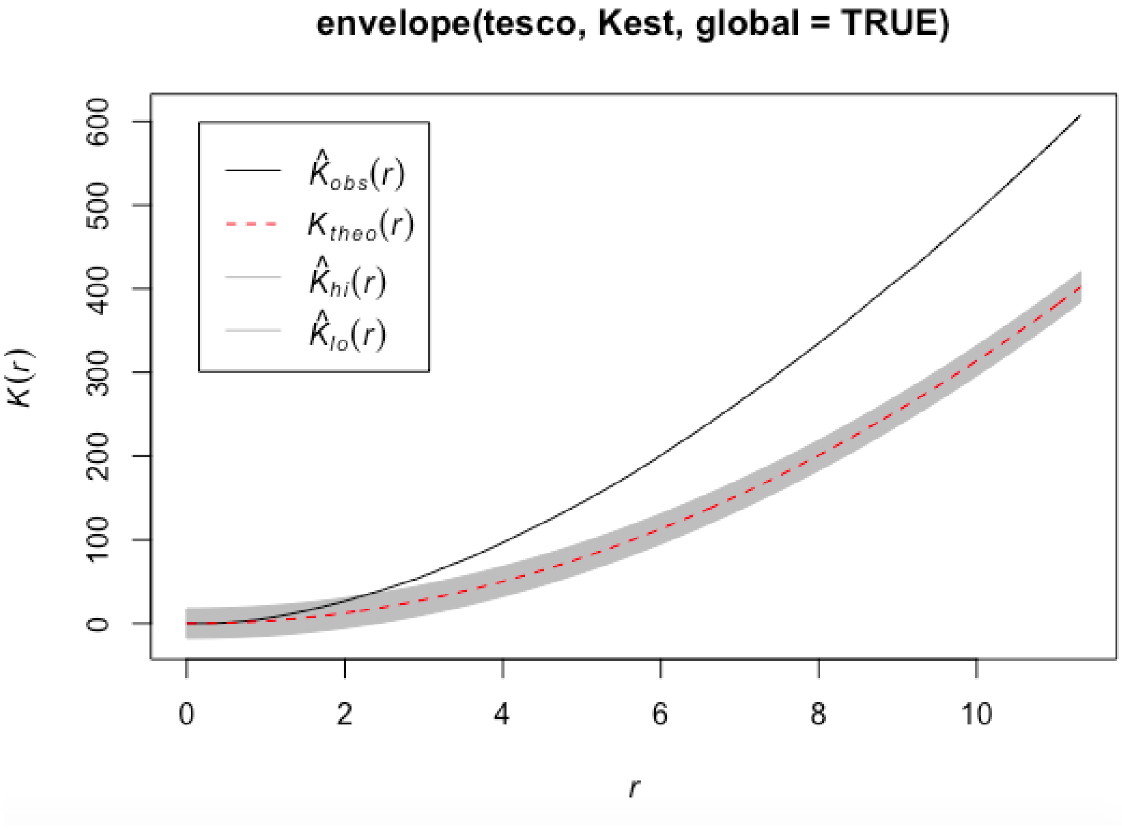


Figure 15: Tesco global envelope with K-function

The density graph (Figure 16) also shows the same trend. From the Tesco density plot, we can easily see that shops are more concentrated in the middle of London. Overall, we believe that the location of Tesco supermarket does not follow complete spatial randomness.

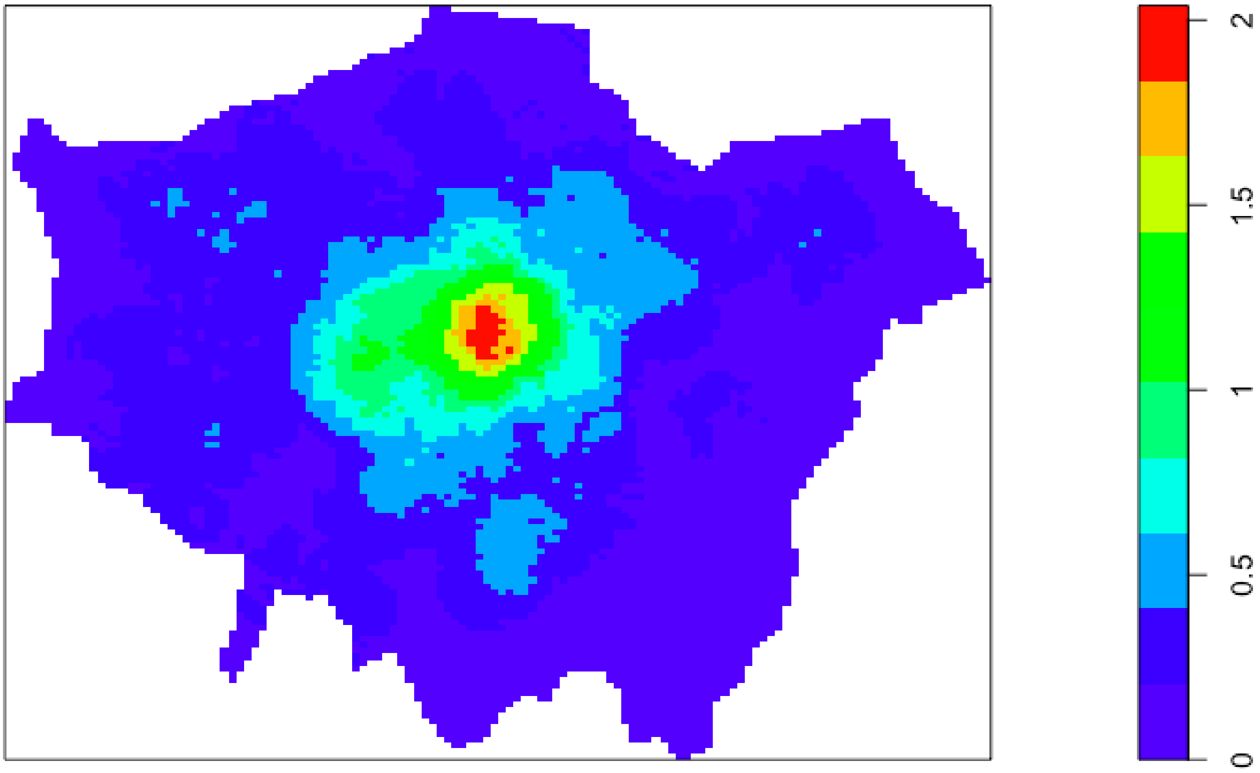


Figure 16: Tesco shop density

4 Modelling of supermarket intensity

4.1 Inhomogeneous Poisson Process

From the previous chapter, the homogeneous Poisson point process is not a suitable model for Tesco locations. As mentioned in section 4.1, now we consider the inhomogeneous Poisson process. A straightforward way of defined inhomogeneous Poisson process is introducing inhomogeneous to replace the homogeneous in the homogeneous Poisson process. This means that the Poisson process does not have a constant intensity, instead, an intensity function would vary with location. (25) Points patterns from Figure 17 are generated under the inhomogeneous Poisson point process with an intensity proportional to the bivariate normal density, showing a very different spatial distribution.

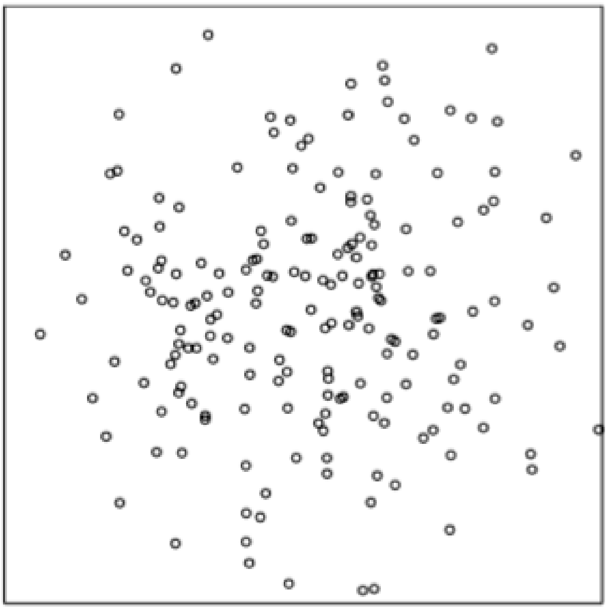


Figure 17: Inhomogeneous example (25)

4.2 Visualising inhomogeneity

In order to use the observed discrete set of points to produce an estimate of the continuous intensity across the window, Kernel density estimation is an approach which many people used. Kernel density estimation is about smoothing the intensity around each point by a distance which depends upon the bandwidth r. So, the radius, or more generally, the bandwidth r specifying the degree of smoothing and it is the tricky part when doing Kernel density estimation. A small value of r leads to a too spiky outcome (left), a large r result in a too smooth plot and loss many information (right) while an appropriately chosen r produces a nice surface (centre). A stationary Cox process, which minimises the mean squared error of the Kernel density estimator is a technique to select r.(25)

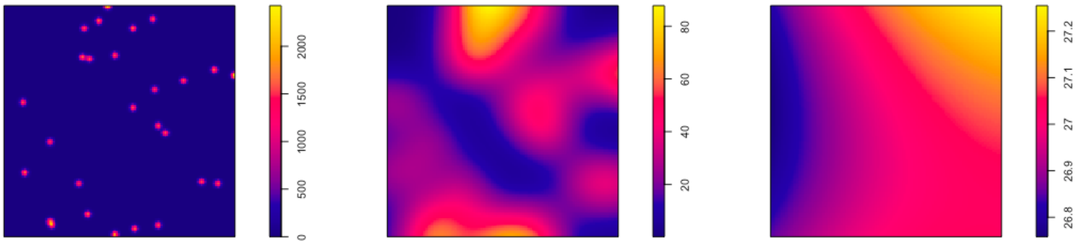


Figure 18

4.3 Inhomogeneous intensity

Now, we try to find a useful model to represent Tesco locations. Linear regression is one of the oldest and also the most frequently used statistical methods. The linear regression model is a statistical technique that applies predictor variables (X_i) to predict the outcome of a response variable (Y) and it is a special case of the general linear model (always written as GLM). A multiple linear regression model with k number of predictor variables X_1, X_2, ..., X_k and a response Y, can be written as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$. β_0 is the intercept and $\beta_1 ... \beta_k$ are regression coefficients. ϵ is the residual terms of the model. Variable can be classified into two classes, continuous or categorical. Examples of continuous variables include population density, total average monthly income and age. Examples of categorical variables include religion (Christian, Buddhist, Hindu, Jewish, Muslim, Sikh, others) or ethnic group (white, mixed, asian, black, others). If the response variable is binary, then we use a logistic regression analysis technique. If all predictor variables are categorical then we use analysis of variance technique. If some predictor variables are categorical and some are continuous then we use the covariance analysis technique. The technique used where categorical variables exist called dummy variables.

We are using a similar idea here but instead of predicting the outcome of Y, we predict the intensity at each area. We are predicting the intensity, λ, but this has to be non-negative everywhere. Similar in some ways to a Poisson GLM. The use of log function ensures non-negative. Maximum likelihood estimation is used in order to determine coefficient estimates and the uncertainties in coefficient estimates. If the mean of response variables are related to the linear predictor by a logarithmic function, $log\lambda = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ then it is described as a Poisson GLM. If we reverse this then we get $\lambda = \exp^{\beta_0 + \beta_1 x_1 + \beta_2 X_2 + ... + \beta_k X_k}$.

Z test is to test if the coefficient of a covariate is zero. Z test is accounted for the estimated value and their uncertainty and compare them to normal distribution.

4.4 Model fitting

A common method for building statistical models is to minimize variables until the most parsimonious model describing the data is found. Due to the large amount of data, automatic variable selection procedures like backward elimination, forward selection and stepwise selection are commonly used in software.

The forward selection starts with the intercept-only model. A fixed significance level of α_{enter} is set. Covariates that are not in the model are added if the calculated p-value for the Chi-squared test for this covariate is smaller than α_{enter}. After the entering of a covariate, it will stay in the model. Repeat this process until none of the covariates' p-values are smaller than α_{enter}.

The backward elimination is the reverse procedure for the forward selection. A fixed significance level of α_{remove} is set. Compare the largest p-value with the α_{remove}. Remove the covariates with the largest p-value if it is greater than α_{remove}. After the removal of a covariate, it will remain excluded. Repeat this process until none of the covariates' p-values are greater than α_{remove}. The Wald test result of a single parameter was checked. The least important effects that do not meet the model stay level are removed. After the delete an effect from the model, the effect will remain excluded. Repeat this process until no other effects in the model reach the specified removal level.

Stepwise selection is a combination of the forward selection and the backward elimination, except that it is not necessary to retain the effects already in or excluded from the model. The stepwise selection is terminating if no further covariates should be added or excluded from the model. (26)

This algorithm relies on the p-values corresponding to covariates might not enough to conduct a model. Applying all backward elimination, forward selection and the stepwise selection and compare their results might provide a start direction. After, check by ANOVA (analysis of variance), with a more theoretical basis, analyse variability to check our covariates selection. ANOVA is just an F-test to compare two general linear models.

When we do linear models, there are some techniques we normally used to check how well does our model represents the true data like R-squared, adjusted R-squared, AIC, etc. R-squared says how well does our line fits the observations while adjusted R-squared says how well does the line fits the observations minus a penalty for how complicated the model is. The more complicated the model, the closer the model to the observations. The model should be close to the observations but not too complicated. Adjusted R-squared considers both of these two so it is more preferred than R-squared. R-squared and adjusted R-squared are numbers between 0 and 1, the closer to one, the better. However, sometimes a model with a good R-squared number is not a good thing, it might trigger by overfitting. AIC is a similar measurement to adjusted R-squared. AIC used to find a balance between the conflicting demands of accuracy and simplicity. AIC does not have a scale, it can take pretty much every value. The model with a smaller AIC value is preferred.

Here, we could not find anything exactly the same as those techniques we mentioned above. But we do have something similar - log likelihood. When fitting a model in order to determinant the coefficient of this model, we use maximum likelihood estimation. Maximize the likelihood is the same thing of maximising the log likelihood. Log likelihood is a measurement of how well does our data fits the data but it does not come with a penalty on how complicated the model is. Unlike R-squared or adjusted R-squared, log likelihood does not have a useful scale. So we can not just look at the log likelihood value and said if our model is good or not. The log-likelihood on its own is not on a useful scale but it is useful to compare with model especially nested model. It is worth to mention that the comparison is only useful for the same data. Log likelihood values are usually used to calculate the deviance. Deviance is twice the difference in log likelihood. When using ANOVA to compare two models, the deviance needs to be calculated. We can compare deviance to the Chi-squared distribution and see if there is a significant improvement going from one to the another. A more complicated model will automatically fit the data better. The question is whether the improvement of the fit is worth that added complicity to the model. In order to find a better model, we have considered x, y, xy, x^2 and y^2 as covariates.

For example, model 1 and model 2 are two possible models to represent Tesco supermarket distribution. Model 2 is a model based on model 1 but have 5 more covariates x, y, xy, x^2 and y^2. The log likelihood for model 1 is -806.4544 and the log likelihood for model 2 is -801.8308. So the deviance is $2 \times (-806.4544 - (-801.8308)) = 9.2473$. Model 2 has an additional 5 parameters so there is 5 degree of freedom difference between those two models. If we compare the deviance (9.2473) with the chi-squared distribution with 5 degree of freedom, the p-value is 0.0996, which is bigger than 0.05. So there is no significant evidence against the null hypothesis. The simpler model 1 is better than model 2. In this case, the additional complexity outweighs the improvement, so the less complex one i.e. model 1 is better.

4.5 Selection of potentially relevant variables

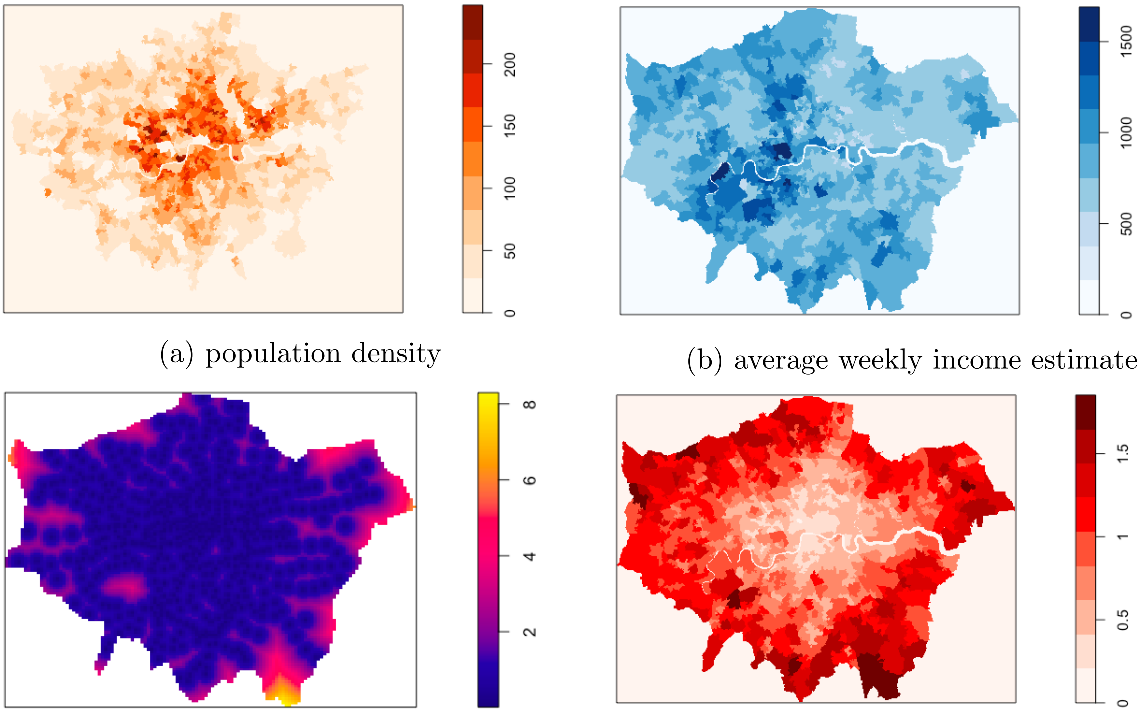
Publicly data used here are census information from Greater London Authority (16) and station location information from doogal website.(20) Predictor variables which possibly relevant to Tesco supermarket location due to past researches are listed as following table:

Variables	Unit
Location	Supermarket location information
Income	Average weekly household total income estimate
Population density	Person per square kilometer
Cars	Cars per household
Birth	percentage not United Kingdom
Stations	Distance from stations to nearest supermarket

All the variables are continuous. Population density might be useful because people have demands for supermarkets. Distance to the stations might be useful because people are more accessible to supermarkets if these supermarkets near a station. Average income might be useful because people with different income levels might have different demand for goods. The understanding of these variables could help supermarkets expand faster and earn more profits. It is also worth mentioning that the initial unit of population density from census information is person per hectare. Because all the other measurements are measured in kilometre, like distances from locations to their nearest stations. Therefore, the population density is converted to person per square kilometer.

Until 2011, the median population density is 7252.0 and this means there were 7252 people per square kilometer. The use of means might not be a good indicator here, because it is the mean of 983 different areas in London rather than the mean of the whole London area. So we use the median instead of mean to gain a rough idea about population density. The most crowded area of London, as we can see from the graph was the area around the Thames Rivers. An area from Westminster borough, where had 24720.9 people per square kilometre was the densest area. An area inside the Bromley borough was with 286.6 people per square kilometre and this was 0.012 of the most density area.

In 2007, the median value of the average weekly household total income was 790.0 pounds. From the plot, we can find that people who lived at the edge of London on average earned less than those living in the centre of London per week. People who lived in an area



(a) population density

(b) average weekly income estimate

(c) distance to station

(d) number of cars per household

Figure 19: Rough look at covariates

of Tower Hamlets had the lowest average weekly income (480 pounds) while people who lived in an area of Kensington and Chelsea had the highest average weekly income (1690 pounds), which was just over 3.5 times more than the lowest average weekly income.

The distance to stations is shorter in the centre of London than then edge. This might because at areas with higher income, the government could gain more tax income. Therefore, when planning and building infrastructure like stations, the government would primarily satisfy the needs of relatively wealthy areas. The dots on the plot are the locations of stations.

The number of cars per household approximately showed a totally different trend compared with income. Unlike the distribution of income, the closer to the center, the less likely a household would hold a car. This may be due to the expensive holding fees like parking fees and congestion charge in the centre London. The congestion charge in London is 11.50 pounds a day. If you need to pay for the congestion charge every day, it would be 80.5 pounds per week. 80.5 was around 10 percent of 790 (the median of average weekly income). The other reason is that the high public transportation coverage in the centre of London also lowers people's interest in owning a private car. The median of the number of cars per household was 0.8239. The average number of cars per household was 1.9 in an area from Havering borough, which was also the highest number. Areas from Camden borough had the least car per household (0.2).

4.6 Tesco model and its interpretation

This is our final model for Tesco supermarket. The log intensity in the left hand side equals to everything on the right hand side. Our Tesco model can therefore write as:

$$log\lambda_{Tesco} = -0.4467 + 0.0008 income + 0.00006 popdens - 1.0871 stations - 1.2183 cars \quad (11)$$

	Estimate	Ztest	Zval
(Intercept)	-0.4467		-1.8765
Income	0.0008	***	3.7355
Population Density	0.00006	***	5.3718
Stations	-1.0871	***	-8.3291
Cars	-1.2183	***	-6.2040

Since neither the population density nor the distance from a location to its nearest supermarket or other covariates is expected to be zero, the intercept does not have a good explanation in the real world. If everything else remains unchanged, the one unit increase of cars per household may lead to the average 1.22 unit decrease of the log intensity. The one unit increase of distance from stations to nearest supermarkets drives the average log intensity down 1.09 unit when the other conditions are keeping the same. The average weekly household total income and population density show the opposite trend. The rise one unit of both of these two covariates would raise the average log intensity, 0.0008 and 0.000006 unit respectively.

This model could also be written as intensity directly instead of log intensity form,

$$\lambda_{Tesco} = 0.6397e^{0.0008income}e^{0.00006popdens}e^{-1.0871stations}e^{-1.2183cars} \qquad (12)$$

The coefficients of stations or cars are less than zero, so when other things are unchanged, the exponential of negative would be smaller than one. The multiplication of a negative number makes the intensity lower, by multiplication of the intensity by 0.3372 and 0.2957. Similarly, when other covariates remain unchanged, the increase of income or stations might cause the intensity bigger because they have positive coefficients and the exponential of positive is bigger than one (the old intensity times 1.0008 and 1.00006 respectively).

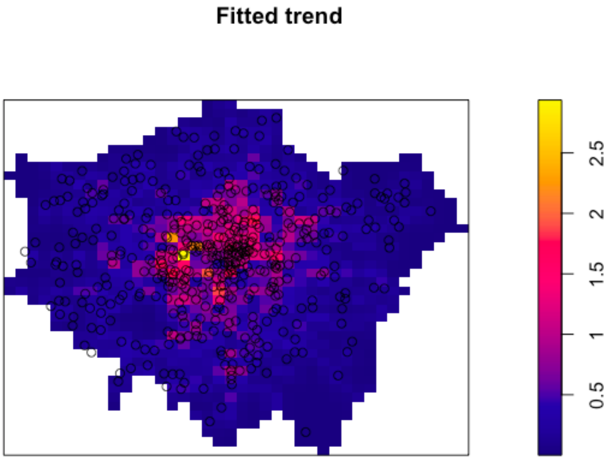


Figure 20: fitted trend

The fitted model is estimated intensity. The intensity is higher at the yellow color region and lower at the blue color region. The yellow region is where more likely to have a supermarket and the blue is where less likely to have a supermarket. If there is a new supermarket, it is more likely to be in the yellow place, moderately likely to be in the pink place and unlikely to be in the blue place.

From Figure 21, we can see that the black line is consistently inside the shaded region. It suggests that our Tesco model fits the truth at 98 confidence intervals. We made 99 simulations and include our observation so that is 100 in total. If the shaded region is the

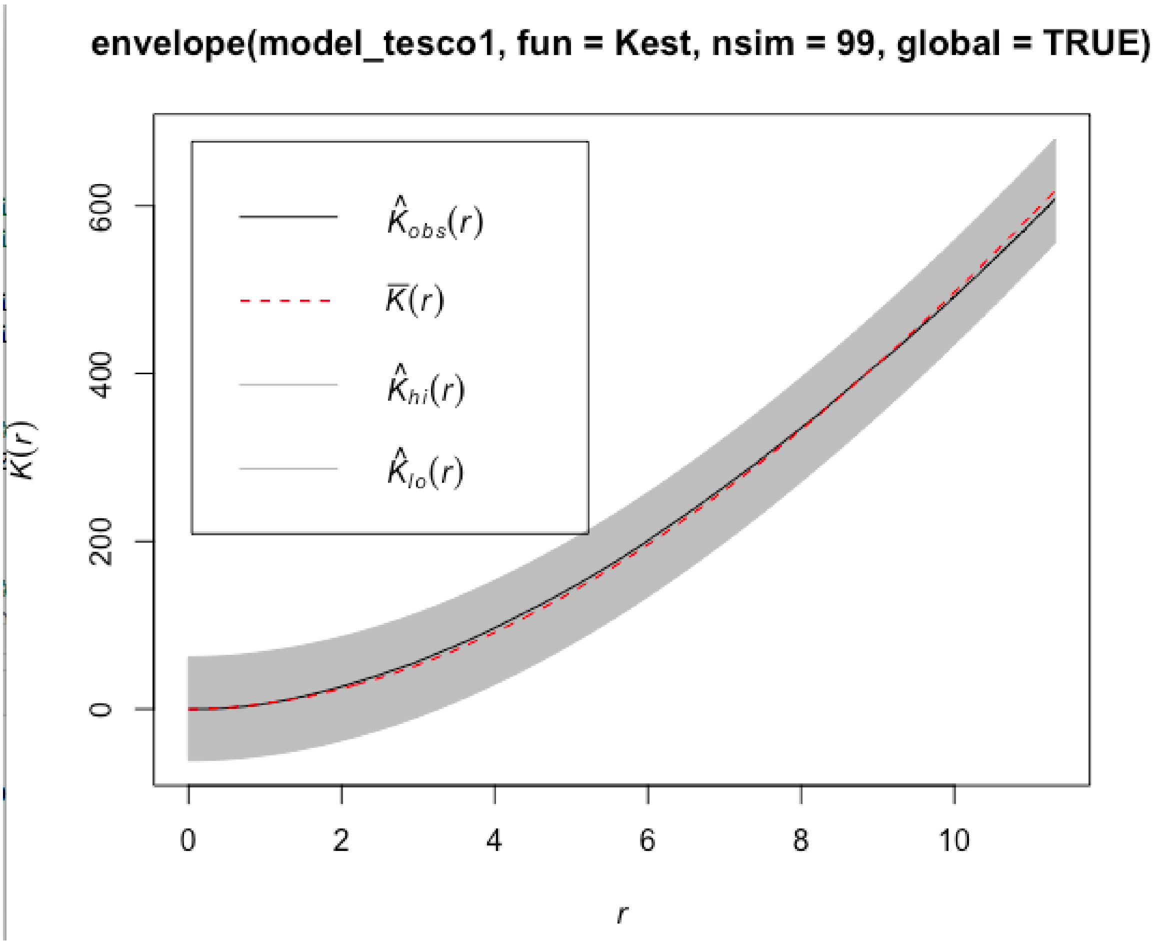


Figure 21: diagnose plot

biggest and the smallest, then we are looking at 2 out of 100. 1 in 100 chance to be very big and 1 in 100 chance to be very small. Thus, the grey shaded region maybe 98 percent confidence interval.

The other way to see if the model is a good approximation to the truth is to take a look at the cumulative sum of the raw residuals plot. The yellow region from bottom-right plot represents that our model fits in those places almost as well as it could. The red region means our model is suggested that our observed value is bigger than our fitted value. The number of points in red regions is more than we are expected to see according to our model. And the blue areas are the opposite. It said that we are predicting too many points in these areas.

So our Tesco model is not well enough predicting the number of points in the edge region, the intensity we predicted are smaller than the actual intensity at red region while the intensity is greater than that actual intensity at the blue region.

The residual plot with different colors highlighting where we make intensity prediction lower or higher than they actually are. But the significance level of that higher or lower is not taken into account. We could always over predicting some places and under predicting some other places. If the under or over predicting is not at a significant amount, then it would not be too problematic. The plots in the bottom-left and top-right are about the cumulative sum of residuals which we could take a look at. The residual is how different are the predicted values compared with the true data. Positive residuals (refers to the red color at the bottom-right) are likely to mean we have more supermarkets than we would expect to see and the negative residuals (refers to the blue color at the bottom-right) are vice versa. Dotted lines are typically taken to indicate the significant differences from what we expected to see and the observed data. So if our cumulative sum of raw residuals is totally inside the dotted lines, then there is not evidence to said our model is not a good approximation to the truth.

However, in the left-bottom plot, the cumulative sum of raw residuals is out of the dotted lines at the left hand side (inside the blue box). If we look at the top-left plot, which is the distribution of all Tesco supermarket in London, then we would see a region (inside the red box) with no supermarket. The red box is just above the blue box. By looking at the map from Google, this region might be a big park called Richmond Park with no-one lives in the park. Since there is not a MSOA region that has 0 population density, there is not a single MSOA region for this Richmond Park. A number of MSOA overlaps the park is a more possible case. So this could be one of the possible explanations of the poor of fit and the under predicting. If we look back to Figure 19(d), people usually have at least one car on the left hand side might also lead to this lack of fit.

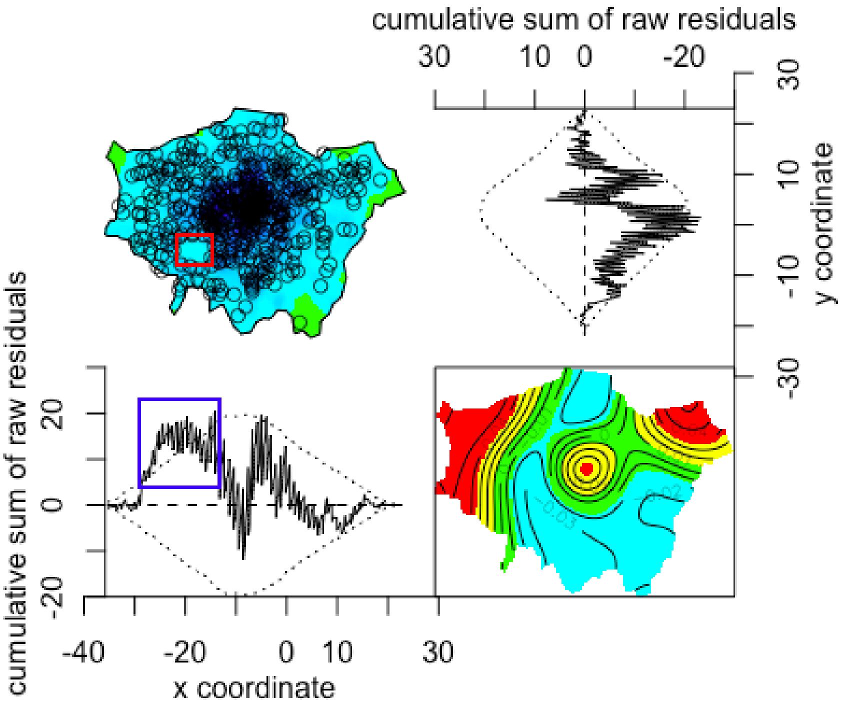


Figure 22

We could improve our model if we include some covariates which are different around edges especially at the left hand side of London. Maybe other covariates from census data which are higher in the west and east direction and are lower in the north and south direction than they are at the other region should be included. People who live in the edges might travel to other cities to work, so they might consume at supermarkets near their workplace rather than near their property. Since we are only able to consider our covariates within the window, the lack of information near edges would cause lack of fit.

4.7 Fit Iceland data into Tesco model

	Estimate	Ztest	Zval
(Intercept)	-1.1612	**	-2.7197
Income	-0.0020	***	-3.5684
Population Density	0.0001	***	5.3778
Stations	-1.3687	***	-5.4166
Cars	0.4880		1.2274

It is quite difficult to compare models with different covariates. But if we fit the same model to two different sets of data, we can compare with their coefficient. We then fit our Tesco model with the same covariates to Iceland retail points data. These are the coefficients that come out.

$$log\lambda_{Iceland\ Tesco} = -1.1612 - 0.0020income + 0.0001popdens - 1.3687stations + 0.4880cars \quad (13)$$

or

$$\lambda_{Iceland\ Tesco} = 0.3131e^{-0.0020income}e^{0.0001popdens}e^{-1.3687stations}e^{0.4880cars} \quad (14)$$

The Tesco coefficient for income is with a positive sign (0.0008) but the Iceland coefficient for income is with a negative sign (-0.0020). The intensity of Iceland would decrease to 0.998 times the origin intensity when the average weekly income per household increased by 1 and other things remain stable while the intensity of Tesco would be increase to 1.008 times the original. This would mean that Tesco is more targeting at areas where people have higher income but Iceland is more targeting at areas where people have less income. As says on its official website, Iceland aims to offer equal quality as the majority of other supermarkets at a lower price. (13) So it is unsurprisingly to see that Iceland are more interested with family living on a tight budget.

Then moving onto the population density. The population density coefficient for Iceland (0.0001) is nearly 1.5 times as it is for Tesco (0.00006). Both Tesco and Iceland have a preference for more density areas but Iceland as a stronger preference than Tesco. Maybe that is because there are overall less Iceland supermarkets. It might be sensible for Iceland to open their stores at more density areas to access to more consumers. In fact, Iceland has 124 stores only in London and that is nearly one fourth of the number of Tesco supermarkets in London (467). Unlike Iceland, Tesco has many supermarkets points and thus, abler to cover the whole London area and less focus areas' population density. Than Tesco can afford to spread them to a wide area. Another possible justification might be that most of the Iceland supermarkets are big (between 280 and 1400) while Tesco has both bigger and smaller store sizes. Tesco could manage to have small stores at low population density areas and big stores at high population density areas. But it would be more beneficial for Iceland to focus on areas with lots of people.

Both Tesco and Iceland prefer to be close to a station (they all have negative coefficient). Iceland has a stronger preference to near a station compares with Tesco. This perhaps because Iceland aims to households with lower income and they are more reliant on public transportation whereas Tesco aims to households with higher income and they are more affordable to other kinds of travel tools.

The coefficient for cars of Tesco (-1.2183) shows a totally different preference than that of Iceland (0.4880). In other words, Tesco is more likely to be found in areas where people do not own cars than own cars. Iceland is more likely to be found in areas where people own cars than not own cars. One of the explanations of this could be that Iceland is more randomly distributed but Tesco tends to cluster at the center of London. People are less likely to own a car when they live in the center of London so the coefficient for Tesco is negative.

4.8 Iceland model

	Estimate	Ztest	Zval
(Intercept)	-1.0382	**	-2.6326
Income	-0.0016	***	-3.8383
Population Density	0.0001	***	5.8204
Stations	-1.2885	***	-5.3350

We also find the best model we can compute for Iceland:

$$Log\lambda_{Iceland} = -1.0382 - 0.0016 income + 0.0001 popdens - 1.2885 stations \tag{15}$$

equivalent to

$$\lambda_{Iceland} = 0.3541e^{-0.0016income}e^{0.0001popdens}e^{-1.2885stations} \tag{16}$$

It does not very different from the previous Iceland model, which we fit is Iceland data into our Tesco model. One obvious difference between these two is that the number of cars in each area would affect the decision of Tesco supermarket selection while Iceland is not very interested in cars. Tesco values more on places where people have no car. This looks reasonable. Tesco is more likely to be found in higher income area where people are less likely to have a car. Because the car covariate does not present on the Iceland model, it may suggest that Iceland might not rely on cars to make the position selection decision.

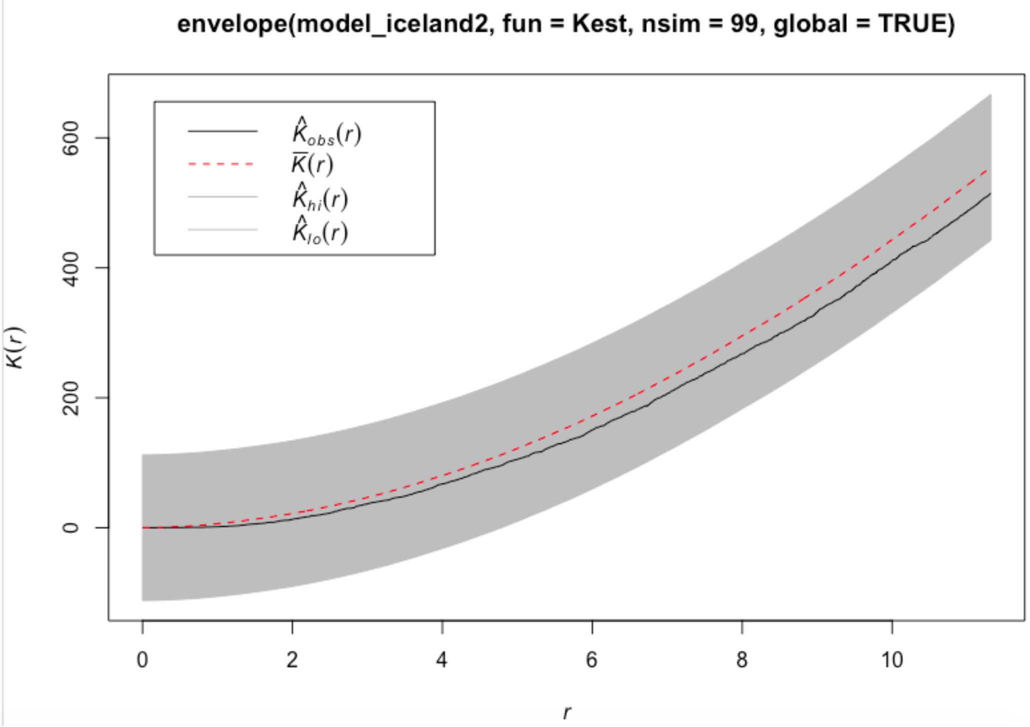


Figure 23: diagnose plot

The Iceland model fits the truth at 98 confidence interval, this suggests that our model could be used to explain the Iceland locations.

The diagnose plot (Figure 24) of the Iceland model is showing a different trend. The 1/3

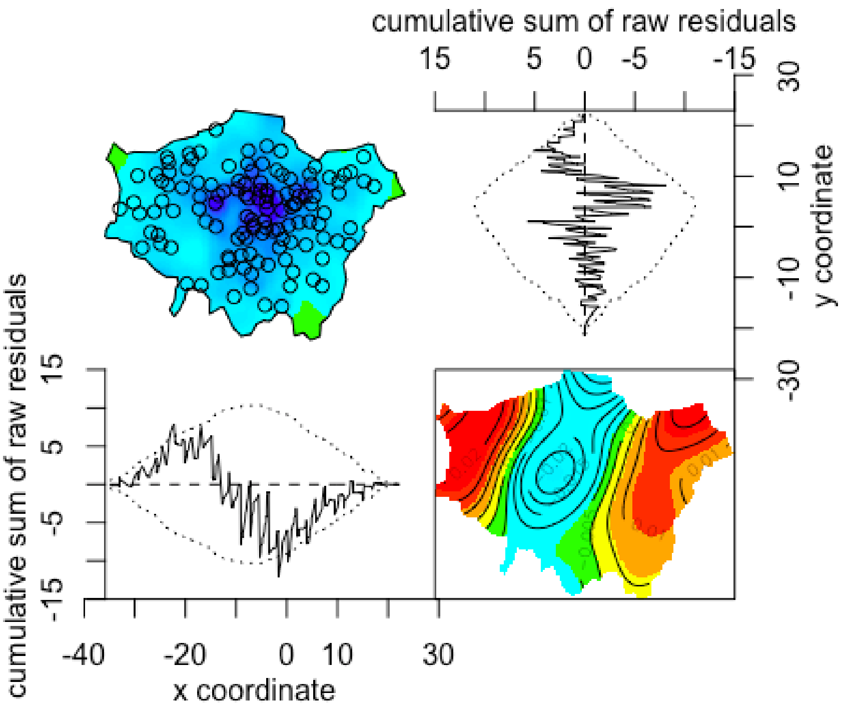


Figure 24: diagnose plot

part from the right hand side areas is relatively well presented by our Iceland model. But more points are predicted at the 1/3 of the left hand side part in comparison with the actual number of points while less point is predicted at the middle 1/3 parts.

4.9 Assumption and other concern

The constant error variances, constant variances are independent of the mean, normally distributed errors and the systematic effects combine addictively are the main assumptions of a classical general linear model. But the assumption here might not be the same. Points come from an inhomogeneous Poisson process might be our assumption.(27) The black line from the envelope for the k-function is consistently within the shaded region so it is not unreasonable to suggest that these points come from an inhomogeneous Poisson process. (Figure 21&23)

	Popdens
Cars	-0.7463

From this plot, we would see that there is a strong negative correlation (-0.746) between cars and population density. This indicates that in the very center of London, where is the most crowded area, people are less likely to have a car. When people live at the very edge of London, they are more likely to own a car. This may because the expensive holding fees and high public transportation coverage rate in the centre of London lower people's interest in owning a car and vice versa. In linear regression, we would concern if our covariates are correlated. But here, we do not need to concern too much about the dependency.

The inclusion of other useful data could improve our model. For example, include the size of family when modelling the average weekly household income. Households with the same amount of income but different sizes of family might have different consuming abilities. Supermarkets might make use of this to meet consumers' needs.

5 Interaction

So far, we only focused on the analysis of single point patterns, such as the location of Tesco supermarkets or Iceland supermarkets. But often, we also interested in the involvement of any interactions between more than one pattern. For instance, if there is a Tesco supermarket, it would compete with nearby supermarkets for consumers' loyalty and quality of items. There are two commonly used ways to test interactions between point patterns, random labelling and random shifting.(28)

5.1 Cross K-function

To gain an understanding of these two approaches, we need to introduce cross K-function first. Similar to K-function we have mentioned before. For two independent processes, event of type 1 and type 2 has the same status as arbitrary points. Type 1 and type 2 with intensity λ_1 and λ_2 respectively, then the cross K-function for population 1 with respect to population 2 denotes $K_{12}(r) = \frac{1}{\lambda_2}\mathrm{E}($expected number of points in population 2 with radius r of an arbitrary point from type 1). For example, the K-function of univariate Tesco process is calculating the expected number of points within radius r of a given Tesco supermarket location. The cross K-function for Tesco with respect to Sainsbury is calculating the expected number of Sainsbury within radius r of the Tesco supermarket.

To test whether points from type 1 event would influence points in type 2 event, the interactions between these types are investigated. We use the Monte Carlo test (details in section 4.6) and cross K-function as test statistics. If there are more type 1 events close to type 2 events under a significant level than we would expect, it suggests that attraction exists between type 1 and type 2. By contrast, if there are fewer type 1 events close to type 2 events under a significant level than we would expect, it suggests that repulsion exists between type 1 and type 2.

5.2 Random Shifting

Random shifting refers to shifted point patterns from a population by a constant value. Using random sample shifted to simulate the corresponding population of sample cross K-function. Then using the Monte Carlo test to test our null hypothesis. As shown in the following picture, random shifting method requires us to shift from a solid line to a dashed line and use the point patterns inside the dashed line to test our hypothesis. However, the involvement of the edge effects might cause some problems. In spatial point pattern analysis, edge effects occur when a region A we observed in practice is a part of a larger region and under process operates. Because we might not be able to explore the possible exist interactions between unobserved regions outside region A and region A due to the lack of observed data. We only have the information of points inside the solid line region and do not know what is going on outside this region.

If patterns are generated from a stationary point process, points are equally likely to occur anywhere inside the boundary. Thus, the use of replicating give region A is the

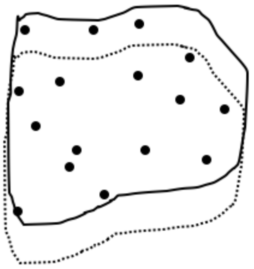


Figure 25: random shifting

same as generating point pattern simulations. As shown below, after repeating region A for 7 times, we get the left hand side region. Then we could continue our test using cross K-function.

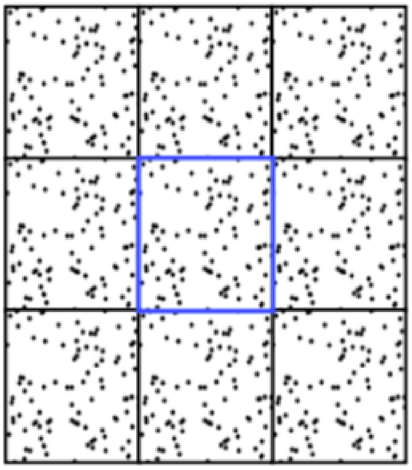


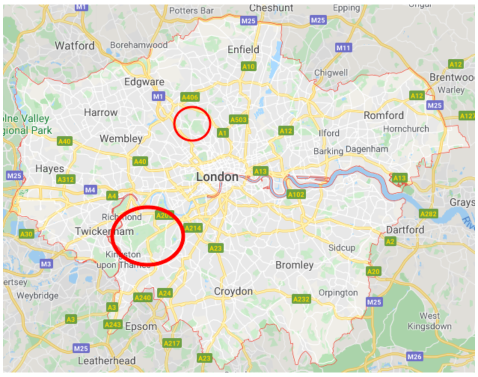
Figure 26: replication of region A(24)

However, because the previous part has already shown that our data do not follow a stationary point process and it would be an inhomogeneous Poisson point process instead. Our data, therefore, do not meet the assumptions to overcome the edge effects.

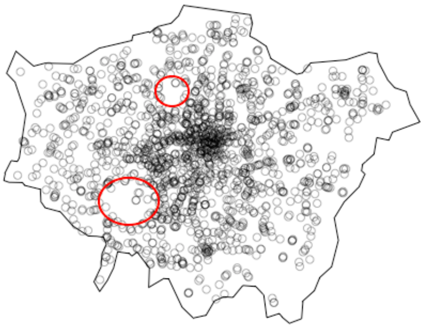
5.3 Random labelling

Random labelling, which is also a commonly used approach is more suitable here. Rather than focusing on generating a pair of independent univariate process, random labelling is, as its name, random labelling the two types of events in mutually independent trails of a univariate process. So under random labelling, any site of supermarket could be occupied by either Tesco or any other brands. Supermarkets with a different brand are not equally likely events, distributions of these brands are used. If and only if type 1 and type 2 events are both Poisson processes, the random labelling is equivalent to independence.(28)

Random labelling is particularly useful when dealing with locations with feasible with several unobserved restrictions. The distribution of supermarkets across the whole of London is plotted at the right hand side plot. We could see that some regions have no supermarket. And combine with London map, those areas with no supermarkets are parks. There are many parks in London, denoted by green on the London map picture. These parks would prevent the existence of supermarkets. Even if these parks are not observed, the observed location of supermarkets should avoid these regions. Hence we could still analyse interactions between supermarkets without model all feasible location if we condition on these park locations. Under this condition, other factors may influence the locations of supermarkets are also avoided like rivers or street networks.



(a) London map



(b) Supermarkets over London

Random shifting might be more conclusive for what we are interested in than the random labelling. There are some structures within each set of points and links between two different point patterns. Random labelling method would both break these structures, as its names, by randomly labelling these points and breaks any link between sets. By contrast, random shifting breaks only the relationship between the different point patterns so the structure of each set have remained. However, due to the violence of random shifting assumptions, we should rely on random labelling for now.

5.4 Interpretation

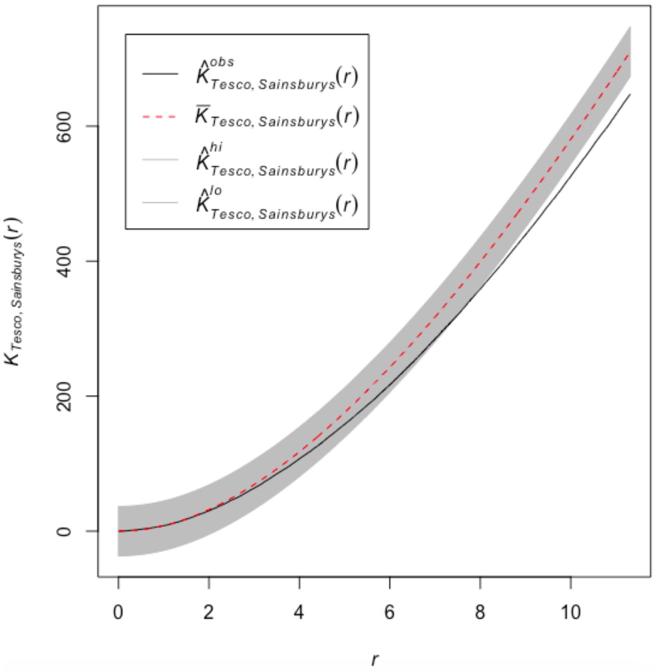


Figure 28: fitted trend

From this envelop plot, the true Tesco data is not inside the grey region at all. We could safely conclude that the rejection of the null hypothesis, which says that Tesco and Sainsbury are arranged effectively randomly. There is some dependency between Tesco and Sainsbury. Our black line is going outside the shaded region and below the grey region. K-function is lower means we have less than we normally see. This is the evidence that Sainsbury is avoiding Tesco when choosing the location.

The previous case is the comparison between Tesco and Sainsbury. What we are going to do next is to compare Tesco to all the other different types. If we look at all the supermarkets in London, we take a K-function for all of those supermarkets and compare that to the K-function when we randomly select half of the supermarkets. We would then expect those K-functions to be the same. So if we took a random selection of points from point pattern and calculated the K-function just for those points that should be the same as calculating the K-function for the whole pattern. If we randomly select a subset of points for many different times and we look at the K-function for those and we compare that for the k function for Tesco then we should be the same.

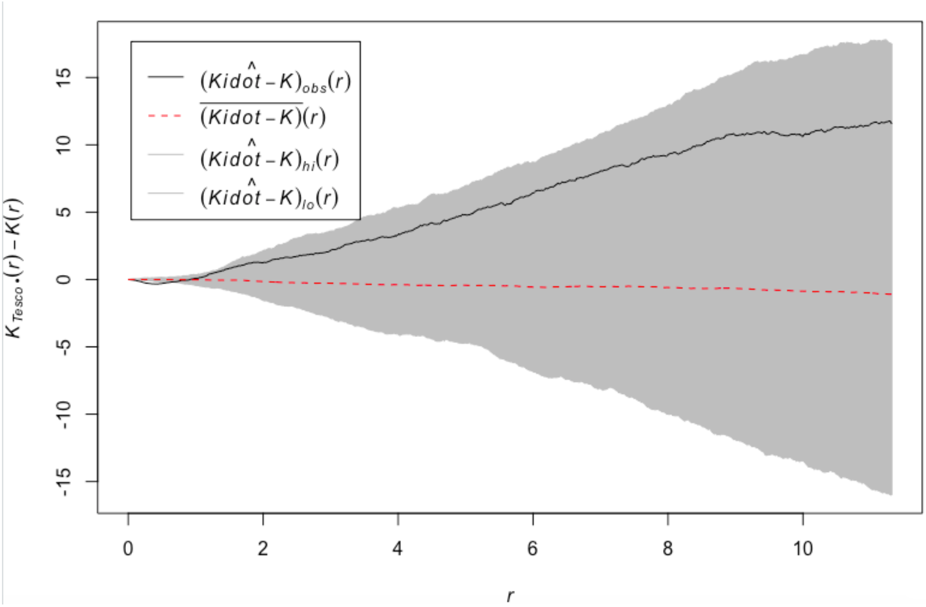


Figure 29: fitted trend

The black line represents the difference between the random subset of point K-function for all the points and its subset (Kidot and K for Tesco). This means looking around with Tesco with radius r, to look around all different types of supermarkets. And we begin over the short distances of R, which is below the average. The average is zero here and it is the horizontal line. So this suggests that the small distances Kidot is less than K. That means our Tesco supermarkets have fewer other supermarkets around than we would expect to see for the random labeling within a circle with a radius less than 1. Maybe Tesco supermarkets are on average bigger so consumers do not need other supermarkets nearby. Or maybe Tesco is one of the strongest competitors for the retail industry and as a result, fewer supermarkets are willing to open a branch nearby the Tesco supermarkets. It then swings the other way. When the black line is higher, it means that looking in a bigger circle around Tesco supermarkets, we see, on average more other supermarkets around Tesco point than around others. The reason for that might be there are more Tesco supermarkets in the center of London and the center of London is where there are lots of supermarkets quite close together. We draw big circles around those supermarkets in the center of London which will count more supermarkets. Those bigger values and bigger distances of r are not outside the grey region but they are bigger than average. There is some evidence saying that Tesco is pushing other supermarkets away from them. On the bigger value of r they are grouping together in the popular location. Supermarkets are more at a popular location than the other ones.

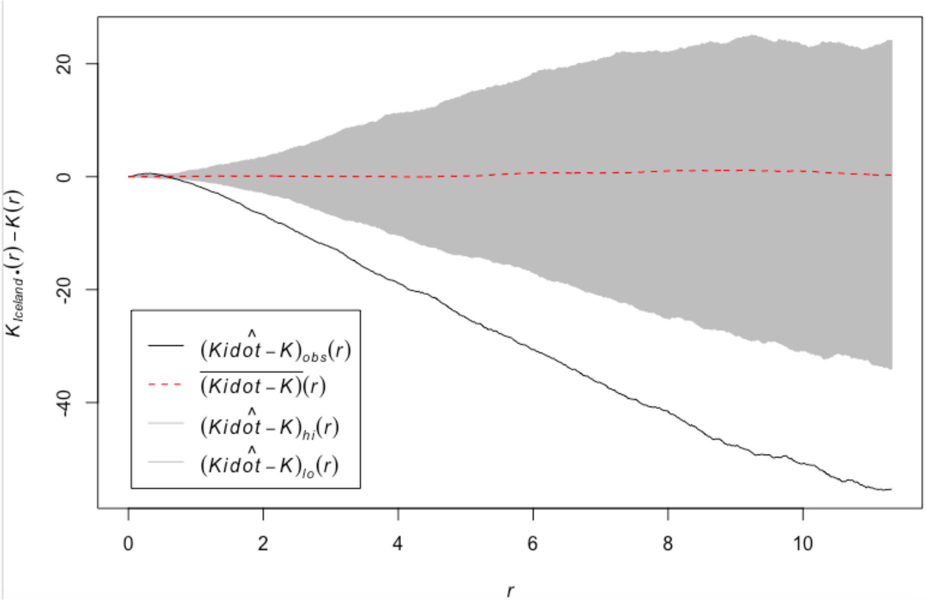


Figure 30: fitted trend

There are some types of supermarkets that are not so focused on the center of London but are more spread out at the edges. Iceland is a more out of town supermarket in which most of the items they sell are frozen food. On the very small distances, in comparison with other types of supermarkets, there are on average more supermarkets around Iceland. Perhaps there is not too necessary for other supermarkets to avoid a relatively specialist supermarket like Iceland. But in the big circle distances, there is a lot less supermarket. This might because Iceland is the type of supermarkets which are away from the middle of London.

6 Conclusion

One obvious result is that retailers have considered the location selection carefully before they make a decision. The locations for each retailer are not randomly distributed and each of them may have different preferences. For Tesco supermarket, managers might rely more on household income, population density, number of cars per household and the distance to a tube station to make decisions. Iceland managers nearly have the same things to focus on but the number of cars per household is not interested to them. It is interesting to see that Tesco is focusing more on wealthier areas while Iceland is focusing more on less developed areas.

From the investigation of the relationship between supermarkets, there is supportive evidence for saying that interactions occur between different brands supermarket. Other supermarkets are avoiding being too close to Tesco while showing a degree of willingness to be close to Iceland.

6.1 Limitations

First, the sources of our data are not all up-to-date and not all from the same year. The census data we have used is from 2012 while the supermarket locations data is from 2019. These differences might influence the accuracy of our outcomes. Second, we did not account for all types of public transport but just the tube and train stations. And the distances from supermarkets to stations we considered are straight line stations which might have some discrepancy compares with the real situations. Furthermore, the dependency is not included when considering inhomogeneous Poisson Process models. Supermarkets for different brands with similar target consumers might not be willing to be too close to each other to avoid unnecessary competition. Supermarkets from the same brand would either be too close to their own supermarket. So there might be repulsion between supermarkets, which we did not take into account when building models.

6.2 Future works

For future thinking, we are perhaps taking account of more information. The sizes of supermarkets are given in our supermarket location data but we have not made use of it. Maybe smaller supermarkets could be closer together whereas larger supermarkets are avoiding each other. We could also look at different cities and see if our supermarket models remain useful, or other cities have different covariates to take into account for retailers. In addition, we could apply different data using a similar idea of this report. For instance, investigate where crime is more likely to happens. Furthermore, we could also forecast the population density, income, etc. and in the combination of this prediction to make suggestions about where supermarkets would open a new cite.

References

[1] Chris Rhodes. *Retail sector in the UK*. House of Commons Library [Internet]. 2018. [Assessed 25 April 2020]; 1:15. Available from: http://researchbriefings.files.parliament.uk/documents/SN06186/SN06186.pdf.

[2] Robert Murray Haig. *Toward an Understanding of the Metropolis*. Quarterly Journal of Economics. 1926; Vol.40: 402-434.

[3] Alnahhal, Mohammed; Noche, Bernd. *A genetic algorithm for supermarket location problem*. Assembly Automation. 2015; Vol. 35(1): 122-127.

[4] Hande Dilek, Özgen Karaer, Emre Nadar. *Retail location competition under carbon penalty*. European Journal of Operational Research. 2018; Vol. 269(1): 146-158.

[5] Stephen Brown. *Retail location theory: evolution and evaluation*. The International Review of Retail. 2006; Vol. 3(2): 185-229.

[6] Mark Birkin, Graham Clarke and Martin Clarke. *Retail Location Planning in an Era of Multi- Channel Growth*. Oxon: Routledge; 2017.

[7] LP Simkin. *SLAM: Store location assessment model—Theory and practice*. Omega. 1989; Vol. 17(1): 53-58.

[8] Orietta Nicolis, Francisco Plaza, Rodrigo Salas. *Prediction of intensity and location of seismic events using deep learning*. Spatial Statistics. 2020; 100442.

[9] Alexsandro C.O.Silva, Leila M.G.Fonseca, Thales S.Körting, Maria Isabel S.Escada. *A spatio-temporal Bayesian Network approach for deforestation prediction in an Amazon rainforest expansion frontier*. Spatial Statistics. 2019; Vol. 35: 100393.

[10] Nancy L.Garcia, Peter Guttorp, Guilherme Ludwig. *Interacting cluster point process model for epidermal nerve fibers*. Spatial Statistics. 2020; Vol. 35: 100442.

[11] Emma Lundberg, Georg H. H. Borner. *Spatial proteomics: a powerful discovery tool for cell biology*. Nature Reviews Molecular Cell Biology. 2019; Vol. 20: 285-302.

[12] Matthew Quick, Guangquan Li, Ian Brunton-Smith. *Crime-general and crime-specific spatial patterns: A multivariate spatial analysis of four crime types at the small-area scale*. Journal of Criminal Justice. 2018; Vol. 58: 22-32.

[13] Iceland. *Our Strategy [Internet]*. 2018. [Assessed 25 April 2020]; Available from: https://about.iceland.co.uk/our-strategy/.

[14] GADM. *Download GADM data (version 3.6) [Internet]*. 2019. [Assessed 10 October 2019]; Available from: https://gadm.org/download_country_v3.html.

[15] Geolytix. LTD. *GeoData [Internet]*. 2019. [Assessed 10 October 2019]; Available from: https://www.geolytix.co.uk/#!mapp.

[16] Mayor of London. *MSOA Atlas [Internet]*. 2012. [Assessed 6 November 2019]; Available from: https://data.london.gov.uk/dataset/43ae680d-12fe-4d23-a253-3bcb2cccfd07.

[17] Wikipedia. *Middle Layer Super Output Area [Internet]*. 3 May 2020. [Assessed 1 May 2020]; Available from: https://en.wikipedia.org/wiki/Middle_Layer_Super_Output_Area.

[18] Wikipedia. *Lower Layer Super Output Area [Internet]*. 21 October 2019. [Assessed 1 May 2020]; Available from: https://en.wikipedia.org/wiki/Lower_Layer_Super_Output_Area.

[19] Adrian Baddeley, Rolf Turner and Ege Rubak. *spatstat analysing spatial point patterns [Internet]*. 2014. [Assessed 10 October 2019]; Available from: https://spatstat.org.

[20] Chris Bell. *London stations [Internet]*. 2000. [Assessed 10 October 2019]; Available from: https://www.doogal.co.uk/london_stations.php.

[21] Aparna Kher. *What Are Longitudes and Latitudes? [Internet]*. 2020. [Assessed 1 May 2020]; Available from: https://www.timeanddate.com/geography/longitude-latitude.html.

[22] Oliver C. Ibe. *Markov Point Processes.* Source Code for Biology and Medicine. 2008;3: 17. in Markov Processes for Stochastic Modeling (Second Edition), 2013;

[23] SAMPRIT CHATTEFUEE and ALI S. HAD1. *Regression Analysis by Example.* Fourth ed. Canada: John Wiley & Sons, Inc.; 2006.

[24] Tony E. Smith. *ESE 502 Home Page.* 2020. [Assessed 1 October 2019]; 1:15. Available from: https://www.seas.upenn.edu/ ese502/.

[25] Janine Illian, Antti Penttinen, Helga Stoyan, Dietrich Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns.* England: John Wiley & Sons Ltd; 2008.

[26] Zoran Bursac, C Heath Gauss, David Keith Williams and David W Hosmer. *Purposeful selection of variables in logistic regression.* Source Code for Biology and Medicine. 2008;3: 17.

[27] Henrik Madsen and Poul Thyregod. *Introduction to General and Generalized Linear Models.* Boca Raton, FL: CRC Press; 2011.

[28] Peter J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* Third Edition. England: Taylor & Francis Group, LLC; 2014.