Spatial dependency between a linear network and a point pattern

by

Thembinkosi Kunene

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Department of Statistics

In the Faculty of Natural and Agricultural Sciences

Univeristy of Pretoria

October 2020

I, *Thembinkosi Kunene,* declare that this mini-dissertation (100 credits), which I hereby submit for the degree Magister Scientiae in Advanced Data Analytics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date:

Summary

In this mini-dissertation we discuss the spatial relationship between point processes and a linear network. As a starting point, we discuss basic spatial point processes and tests for first-order homogeneity. Following that, we discuss second-order properties of point processes in the form of Ripley's K-function for unmarked point patterns and the cross-K function for marked point patterns. We then get to the main focus of this mini-dissertation, that is, the spatial relationship between points and linear structures, particularly linear networks. Recently developed is a method to characterise the spatial relationship between points and linear networks by Comas *et al.* [13], similar to Ripley's K-function for point-to-point relationships. The non-stationarity of a linear network is of particular interest in how it affects the measurement of this spatial relationship, which has not been explicitly investigated in the literature before. To investigate this we consider the Poisson line process and how one might simulate a non-stationary line process. Furthermore, we discuss a mechanism to extend tests of first-order homogeneity of point patterns to line patterns. The non-stationary line process is used to model linear networks in the simulations conducted to determine the effect of this non-stationarity on the developed method, which was not covered in the original article [13]. The methodology is developed and tested on a real data set.

Acknowledgements

I would like to thank Dr Inger Fabris-Rotelli for encouraging me to pursue this degree and her subsequent support and availability for the duration of this journey. Your feedback was never half-hearted despite having a number of other students fighting for your attention. Our meetings always put me at ease and gave me the belief that I could succeed in writing this mini-dissertation. I am forever grateful.

I would also like to thank ESRI South Africa for the financial support, through the Statistics HUB in the University of Pretoria, for both my years in the programme. I did not have to stress about my university fees at all and thus only focused on my studies and research. I am truly thankful for the financial support.

Finally, I would like to thank my family, in particular, my parents Dr J. Kunene and Mrs T. Kunene, my siblings Nkosingiphile, Ncobile and Lindinkosi and my friends for their support. It has been a tough journey and I am grateful for all your support and encouragement all the way through.

Contents

1 Introduction **15**

2 Spatial point processes **21**

2.1 Introduction 21

2.2 Intensity 21

2.2.1 First-order intensity 21

2.2.2 Second-order intensity 25

2.3 Poisson spatial point process 26

2.4 Other spatial point processes 29

2.5 Multitype point patterns 32

2.6 Line processes 35

2.6.1 Theory 35

2.6.2 Simulating a line pattern 36

2.7 Conclusion 44

3 Tests for first-order homogeneity **45**

3.1 Tests for first-order homogeneity 45

3.2 Modified test for first order homogeneity 48

3.2.1 Simulations 48

3.3 Conclusion 52

4 Second order properties **53**

4.1 Introduction 53

4.2 The K-function 53

4.2.1 Theoretical definition 53

4.2.2 Estimation of the K-function 55

4.2.3 Edge effects 57

4.2.4 Simulations 59

4.2.5 Confidence intervals 60

4.2.5.1 Block bootstrap 60

4.2.5.2 Loh's bootstrap 63

4.3 The cross-K function 64

4.3.1 Theoretical definition 64

4.3.2 Estimation of the cross-K function 65

4.4 Points and linear structures 67

4.4.1 Introduction 67

4.4.2 Definitions and estimation 67

4.4.3 Confidence intervals: Loh's bootstrap 71

4.5 Hypothesis Tests and Simulation Envelopes 72

4.5.1 Simulation envelopes 72

4.5.2 Maximum Absolute Deviation Test 73

4.5.3 Second deviation test 74

4.6 Conclusion 75

5 Simulations **77**

5.1 Simulations and testing of the empirical estimate 77
5.1.1 Independence 79
5.1.2 Attraction 79
5.1.3 Repulsion 80
5.2 Discussion 90
5.3 Power of the MAD test 92
5.3.1 Setup 92
5.3.2 One-sided alternative 92
5.3.3 Two sided alternative 94
5.4 Power of the second deviation test 100
5.5 Conclusion 105

6 Application **106**

6.1 Data 106
6.2 Conclusion 116

7 Conclusion **117**

List of Figures

2.1 Gorilla nesting sites point pattern obtained from the `gorillas` dataset in the `spatstat` package. There are 647 gorilla nesting sites on this polygonal window. 23

2.2 Realisations of spatially inhomogeneous point patterns with quadrat counts and density plots. The point patterns were generated using the `rpoispp` function, the quadrat counts using the `quadratcount` function and the density plot using the `density` function, of the `spatstat` package. 24

2.3 Realisations of spatially homogeneous point patterns on a 10×10 window with corresponding quadrat counts and density plots. The point patterns were generated using the `rpoispp` function, the quadrat counts using the `quadratcount` function and the density plot using the `density` function of the `spatstat` package. 25

2.4 Realisations of CSR simulated on a unit square window using the `rpoisspp` function of `spatstat` with $\lambda = 100$. 29

2.5 Different realisations of the Matérn cluster process using the `rMatClust` function on a $[0,1] \times [0,1]$ window. 30

2.6 Realisations of Matérn's Model I generated using the `rMaternI` function in `spatstat` with Poisson parameter 2 on a $[0,10] \times [0,10]$ window. 31

2.7 Realisations of the SSI model generated using the `rSSI` function in `spatstat` on a unit square window. 31

2.8 Locations of Japanese pines from the `japanesepines` dataset in `spatstat`. 32

2.9 Two realisations of the multitype Poisson point process generated using the `rmpoispp` function in `spastat` on a [0,1]× [0,1] window. 33

2.10 Plot of locations on fires from `nbfires` dataset in `spatstat` for the year 1998. 34

2.11 Locations of fire types for the year 1998 from `nbfires` dataset in `spatstat` 34

2.12 An equilateral triangle with side length 6.9cm inscribed in a circle of radius 4. 36

2.13 Illustration of the third approach. The chord is the perpendicular line through point C which lies on the radius, and was generated uniformly on $(0, r)$. 37

2.14 Illustration of the important angles α, γ and β and their relationship to the chord $\overrightarrow{DE}$. α is the angle the perpendicular bisector of DE, i.e., $\overrightarrow{AC}$ makes with the x-axis. β is the angle $\overrightarrow{DF}$ makes with the x-axis in anti-clockwise fashion, and $\gamma = \pi - \beta$. 38

2.15 Different line configurations simulated on a unit square window. The non-stationary parameters are simulated using the positively skewed Beta(2.7,6.3) distribution. 41

2.16 Positively skewed Beta(2.7,6.3) distribution. 42

2.17 Negatively skewed Beta(6.3,2.7) distribution. 42

2.18 Different line configurations simulated on a unit square window. The non-stationary parameters are simulated using the negatively skewed Beta(6.3,2.7) distribution. 43

3.1 Realisation of Poisson point patterns on a unit square window generated using the `rpoispp` function with the intensity functions as arguments. 50

3.2 Line patterns in a [0,1]×[0,1] window $\subset \mathbb{R}^2$ with the corresponding representation on $C = [0, 2\pi] \times [0, r_{max}]$. The horizontal axis corresponds to $[0, 2\pi]$ and the vertical axis corresponds to $[0, r_{max}.]$ 51

4.1 Illustration of increasing the radius r. 56

4.2 Regular point patterns with corresponding K-function estimates, generated using `rSSI` and `Kest` respectively on a unit square window. 60

4.3 Clustered point patterns with corresponding K-function estimates, generated using `rMatClust` and `Kest` respectively on a unit square window. 61

4.4 K-function estimates with 95% confidence intervals. These were calculated using the `varblock` function in `spatstat`. **lo** $\hat{K}_{iso}$ and **hi** $\hat{K}_{iso}$ are the lower and upper confidence limits respectively. **mean** $\hat{K}_{iso}$ is the estimate of the K-function. 62

4.5 K-function estimates with 95% confidence intervals. These were calculated using the `lohboot` function in `spatstat`. $\hat{K}_{loCl}$ and $\hat{K}_{hiCl}$ are the lower and upper confidence limits respectively. $\hat{K}_{iso}$ is the estimate of the K-function. The number of resamples was 999. . 64

4.6 Clustered point patterns with corresponding K-function estimates, generated using `rMatClust` and `Kcross` respectively. Displayed in red are the 95% confidence intervals calculated using Loh's bootstrap with 999 resamples. 66

4.7 A simple linear network from the `spatstat` package. 67

4.8 A simple linear network and a point pattern clustered around it. 68

4.9 A simple linear network and a point pattern being repelled by the linear network. 69

4.10 A simple linear network and a point pattern clustered around it. 70

4.11 Pointwise simulation envelopes with the K-function for different point configurations generated using the `envelope` function in `spatstat`. $\hat{K}_{obs}(r)$ is the K-function estimate for the originally simulated dataset, K_{theo} is the expected curve under independence, $\hat{K}_{hi}(r)$ and $\hat{K}_{lo}(r)$ represent the under and lower simulation envelopes respectively. 73

5.1 Poisson point patterns superimposed on linear network with corresponding K_{LX} estimates. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 81

5.2 Non-stationary Poisson point patterns superimposed on linear network with corresponding K_{LX} estimates. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 82

5.3 Poisson point patterns superimposed on a non-stationary linear network with corresponding K_{LX} estimates. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 83

5.4 Illustration of shifting of points initially on a linear network by different values of d to reduce the degree of dependence on the linear network. The linear network is simulated using `rpoisline` function with parameter 5. 84

5.5 K_{LX} estimates for Case 1 for varying values of d, for $n = 100$ points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 84

5.6 K_{LX} estimates for Case 1 for different values of d and n. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 85

5.7 K_{LX} estimates for 3 cases and different values of d, $n = 1000$ points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 86

5.8 Illustration of repulsion of points by linear network for different values of the repulsion distance r. The linear network is simulated using the `rpoisline` function with parameter 5. 87

5.9 K_{LX} estimates for different values of r, for n=100 points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 87

5.10 K_{LX} estimates for Case 1 for different values of r and n. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. 88

5.11 K_{LX} estimates for cases 2-4 and different values of r, for n=1000 points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. The parameter used for the line process was 5 in all cases. 89

5.12 K_{LX} and stabilised K_{LX} estimates for different values of d. The point patterns used in each case had 1500 points and were on a unit square window. The curve in black is the estimate and the curve in blue is the estimate expected under CSR. The vertical red line indicates the point at which the estimate deviates the most from the independence curve, i.e. the MAD value. 97

6.1 The linear network of the Gugulethu area. The total length of roads is 139871.2m. The area of the observation widow is 8487924m^2. The total number of crime points is 16 074, and these were recorded from 2006 to 2016. 107

6.2 Locations of different types of crimes in the Gugulethu area. 108

6.3 K_{LX} estimates for the different crime categories, the empirical estimate in black, the expected curve under complete spatial randomness in blue and 95% confidence intervals. (Note that category 1 is Assault with the purpose to inflict grievous bodily harm.) 110

6.4 K_{LX} estimates for the different crime categories, the empirical estimate in black, the expected curve under complete spatial randomness in blue and simulation envelopes for values of d of interest in shaded grey. (Note that category 1 is Assault with the purpose to inflict grievous bodily harm.) 114

6.5 K_{LX} estimates for the different crime categories, the empirical estimate in black, the expected curve under complete spatial randomness in blue and simulation envelopes for values of d of interest in shaded grey. 115

List of Tables

2.1 Kernel estimators 23

3.1 p-values for test of homogeneity for different point patterns on a unit square window, where the intensity function is $\lambda(x)$ and the Label column corresponds to the labels given in Figure ??. 49

3.2 p-values for test of homogeneity for different line configurations. 50

5.1 Power values for the MAD test when the null hypothesis is CSR and the alternative is the points shifted by a value of d from the linear network. 93

5.2 Power values for the MAD test when the null hypothesis is that the points are within 0.01 of the linear network and the alternative is the points shifted by a value of d from the linear network. 94

5.3 Percentiles for the test statistics calculated for point-to-line configurations for different values of d. 1000 point patterns were simulated for each case, each with 500 points on a unit square window. 95

5.4 Power values for the two-sided MAD test when the null hypothesis is that the points are within 0.01 units of the linear network and the alternative is the points shifted by a value of d from the linear network. 96

5.5 Power values for the two-sided MAD test when the null hypothesis is that the points are within 0.05 units of the linear network and the alternative is the points shifted by a value of d from the linear network. 98

5.6 Power values for the two-sided MAD test when the null hypothesis is that the points are within 0.10 units of the linear network and the alternative is the points shifted by a value of d from the linear network. 99

5.7 Power values for the second deviation test when the null hypothesis is $d = d_1$ and the alternative is $d = d_2$, calculated for n=100, 500 and 1500 on a unit square window. The linear network was simulated using a stationary Poisson line process. 101

5.8 Power values for the second deviation test when the null hypothesis is $d = d_1$ and the alternative is $d = d_2$, calculated for $n = 100$, 500 and 1500 on a unit square window. The linear network was simulated using a non-stationary Poisson line process. 102

5.9 The variance of the test statistic under different null hypothesis values of d and number of points n. The Quantiles (Q) are shown, E representing the empirical quantiles and T representing the theoretical quantiles. The linear network used was simulated from a stationary Poisson line process. 103

5.10 The variance of the test statistic under different null hypothesis values of d and number of points n. The Quantiles (Q) are shown, E representing the empirical quantiles and T representing the theoretical quantiles. The linear network used was simulated from a non-stationary Poisson line process. 104

6.1 p-values obtained after applying the MAD test using the variance-stablised estimate. As can be seen all the p-values are less than or equal to $\alpha = 0.05$. The null hypothesis of complete spatial randomness can be rejected at a 5% level. 109

6.2 Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Assault with the purpose to inflict grievous bodily harm' crime category for different values of d under H_0.111

6.3 Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Common assault' crime category for different values of d under H_0. 111

6.4 Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Robbery with firearm' crime category for different values of d under H_0. 111

6.5 Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Attempted murder' crime category for different values of d under H_0. 112

6.6 Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Car jacking' crime category for different values of d under H_0. 112

6.7 Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Business robbery' crime category for different values of d under H_0. 112

6.8 Test statistics obtained for the different crime categories assuming the value of d shown under H_0 and the corresponding p-value. 113

Chapter 1

Introduction

Spatial statistics is a large and rich field that is still expanding. It is a widely studied field that is not short of contributors building on past work, proposing new methodology and even providing a different angle to work that has been done already. Some authors that have done work in spatial statistics include Cressie in [14], Diggle in [18], Baddeley, Rubak and Turner [5], Chiu, Stoyan, Kendall, and Mecke [10], Daley and Vere-Jones [15], to name a few. Cressie gives a brief history of spatial data in [14]. In his book, Cressie discusses in detail the three types of spatial data namely, geostatistical data, lattice data and point pattern data, of which the last of these is of particular interest in this mini-dissertation.

As defined by Cressie in [14], a spatial point process $\boldsymbol{X}$ is a model governing locations of events on a plane and a spatial point pattern is one realisation of such a process. Cressie writes extensively about spatial data and statistical analyses in [14]. A point pattern is observed on an observation window which we shall refer to as D. In the introduction to spatial point patterns, locations of longleaf pines in a 400 metre square window are shown and the question of if these locations appear to be completely spatially random is posed. This is one of the more important questions to ask about point patterns. Indeed many writers have gone on to say that testing for complete spatial randomness should form part of the preliminary analysis of a point pattern [14, 18]. Complete spatial randomness is the benchmark against which point patterns are compared to and it is synonymous with the homogeneous Poisson process. There are usually two more point configurations of interest, that of positive association (also called clustered points or aggregated points) and that of negative association (also called regularity) [5, 14, 18]. In addition to the locations, Cressie further discusses the case where there are marks associated with the locations of the longleaf pines, that is, extra information about the trees such as their size. This raises even further worthwhile questions about the interaction between large and small trees.

One statistic that we can consider is the number of points in an area. In the early work involving spatial

point patterns, quadrat counts were an important part of the analysis. These were usually compared to a Poisson distribution because, as shall be seen, it plays a very pivotal role in point pattern analysis [14]. Pearson's χ^2 test can be used to test if the quadrat counts do indeed follow a Poisson distribution and hence conclude if the point pattern is completely spatially random. If the null hypothesis of CSR is rejected, the researcher may then attempt to quantify the deviation from CSR. Cressie then discusses such quadrat-based methods including those that use the Poisson distribution property of equality of mean and variance of random variables from this distribution. Attention is brought to the fact that reducing point patterns to quadrat counts may result in the loss of spatial information, so perhaps this should not be the only analysis performed. Furthermore, it assumes independence of quadrats which is to be considered carefully when dealing with spatial data.

Second-order analysis of point patterns is quite popular in the field of spatial statistics. Most commonly known is Ripley's K-function which is used for assessing spatial dependency between points [32]. This was limited to points of the same type i.e. a single point pattern. Briefly, the K-function is cumulative in nature and it counts the number of points within a given distance r of another point. The result may be higher than a certain benchmark (indicating some form of clustering), or lower than that benchmark (indicating a repellent relationship). This benchmark described as complete spatial randomness [18].

There has been no shortage of the use of and commentary on the K-function because of its usefulness. Haase in [22] highlights its growing use in ecological studies and predicts its continued rise in popularity. The author raises concerns about the difficulty for researchers with minimal mathematical knowledge in understanding the K-function. An intuitive and thought-out statistical background of the function and its uses are given in an attempt to fill in the knowledge. The ever-present problem of edge effects is also discussed along with ways to overcome this problem so that the researcher gets sound statistical results.

Dixon in [20] also discusses Ripley's K-function in an ecological setting. The ability of the K-function to describe point processes at multiple distance scales is highlighted as a property that other methods like nearest neighborer methods do not possess. Dixon discusses these nearest neighbor methods in [21]. Authors in [17, 21] highlight the main issues with the Clark and Evans test statistic proposed in [12] based on the average nearest neighbour distance. The first of these issues is that of reflexive neighbours, i.e. two points being each other's nearest neighbours which inflates the variance of the average nearest neighbour distance. The second problem is that the test did not account for edge effects which, according to [21] leads to over-estimation of the average mean nearest neighour distance. Dixon in [21] further discusses alternative methods such as considering the whole distribution of the nearest neighbours, not just the mean, also considered in [17] and commonly called the G-function. The nearest neighbour function G is usually considered together with the empty space function F. These two and their application in the `R` software are covered in [5].

The cross-K function is an extension of the K-function to multitype point patterns discussed by Lotwick and Silverman in [29] to enable the analysis of two types of points in a point pattern. The idea behind the working of the cross-K function is similar to that of the K-function. The main difference is that we now study the number of one type of points within a given distance r of another type of points. It is clear that the researcher must limit their attention to two types of points at a time.

The cross-K function has been widely used in practice including investigating the clustering of fast food restaurants around schools by Day and Pearce in [16], arguing that the proximity of fast food outlets to schools may have a negative impact on the diet of children. Their study included 406 schools, which were divided into primary, middle school and secondary schools, and 998 fast food outlets in New Zealand. Their findings were that there was a tendency of fast food outlets to cluster around schools, particularly at short distances and it was more apparent for secondary schools and schools in socially deprived areas.

Another application of the cross-K function is found in [3] where the authors were also concerned with the exposure of school-going children to unhealthy foods. They investigated the clustering of fast-food restaurants around schools in Chicago. They found that the clustering was statistically significant. Kwate and Loh in [27] had a similar statistical question about the clustering of fast-food locations around schools, with also an interest in the differences in public schools with predominantly Black students in comparison to those with mostly White students. The former was found to have more significant clustering than the latter.

An important application of the cross-K function is in the analysis of crime locations, particularly in relation to schools as investigated in [8]. The authors make a case about its relevancy especially in South Africa where sexual crimes are all too common. It is also argued that schools should be a relatively safe place for learners and any indication to the contrary should raise concerns. It is not an ideal situation for schools to be crime generators. As defined in [23], crime generators are areas that pull large groups of people towards them and within these groups may be opportunists likely to commit crimes. Crime attractors are areas that pull in people with high levels of criminal intent [23]. The analysis of the authors in [8] included 55 schools and 1518 crimes and their conclusion was that there was clustering of sexual crimes around schools, even for the separately analysed schools i.e. primary and secondary schools.

Line processes are considered in [10] and are defined as a random collection of lines. The discussion in [10] includes that of the representation space which is a way to represent lines on a plane as points on the so-called representation space. This involves transforming the orientation of the line and its perpendicular distance from the origin to a set of co-ordinates on this new space. In this way, line patterns may be treated as point processes and the analyses involving point patterns may be extended to line processes. Moving on from lines we have linear networks that also now feature prominently in point pattern analysis. The construction of a linear network starts with a line segment, l_i, on a plane with end points a and b.

The line segment is defined as $[a, b] = \{ta + (1 - t)b : 0 \leq t \leq 1\}$. A linear network is then defined as the union of line segments i.e. $L = \bigcup_{i=1}^{n_l} l_i$, where n_l is the total number of line segments in the linear network [2]. Alternatively, as described in [5], a linear network has line segments which end at the nodes.

An important extension to the K and cross-K function was by Okabe and Yamada in [30]. The Network K and cross-K function were developed for application of points lying on a network. The need for this extension was clear. Some problems do not lend themselves to the ordinary K-function because of the setting. The assumption of a continuous infinite plane may be inappropriate and with it the use of Euclidean distance between the points, therefore a different approach is needed. The approach proposed by the authors involves the use of the shortest path distance, a concept from graph theory, called the network distance. The ideas of clustering and repulsion of points are still applicable but instead of Euclidean distance, the network distance is used. The developed functions can be used to test if points are uniformly distributed on a street network for instance. The binomial point process is used in the hypothesis of uniformity and in defining the network K-function. If there is non-uniformity, it could be indicative of spatial clustering or repulsion.

The use of these extensions are shown in [35] in yet another ecological application. Their goal was to analyse the spatial distribution of roadside populations of Acacia species in South-Eastern Australia. The argument of the inappropriateness of the usual K-function for point patterns on networks is made, citing the problem of the Euclidean distance as already alluded to above. The results they obtained showed clustering of these populations at multiple scales, and this clustering was deemed significant through the use of confidence intervals.

Yamada and Thill in [36] compare the network K-function and the ordinary K-function. The authors investigate the tendency of the ordinary K-function to over detect clustering if it is applied to points that are constrained to lie on a network. This is done through the use of simulation envelopes under the null hypothesis of independence between the points. The network K-function can identify a random pattern while the ordinary K-function suggests clustering. In doing this study the authors show the advantages of using the network K-function over the K-function for points on a network. They further apply the network K-function to accident data in New York to study the spatial clustering of accidents.

Ang *et al.* in [2] modify the network K-function proposed in [30]. They bring attention to the fact that this network K function depends on the network geometry and hence it is difficult to compare network K-functions for different networks. Furthermore, there is no 'one size fits all' benchmark for tests of randomness, like there is for the ordinary K-function. That is, the result expected under complete spatial randomness is dependent on the network as well. This is also extended to deal with inhomogeneous points on a linear network in the form of the inhomogeneous network K-function.

In [34] Shiode proposes an extension of quadrat methods that assume a homogeneous Euclidean plane.

The problem of using Euclidean distance to measure distance between two points in a network setting is again raised, and the author argues the distance may be smaller if Euclidean distance is used compared to if network distance is used. That is, points may be 'close' using Euclidean distance but actually far if network distance is used. The assumption under the method Shiode proposes is that of a network space that is homogeneous in the sense that the probability distribution of points on the network is uniform. The author also details an algorithm for obtaining network quadrats because, as expected, these are different from planar quadrats and are constrained by the network geometry. A comparative analysis between the two methods using Pearson's χ^2 test resulted in larger test statistics for the planar quadrats compared to the network quadrats when done for random points on the linear network. Further discussion about the two approaches, including about Moran's I and an application can be found in the article by Shiode [34].

Monte Carlo methods also feature prominently in spatial statistics such as in [4, 5, 19, 24]. Monte Carlo tests involve calculating the summary function of interest first, like the K, G or F function for the observed data and comparing these to the functions computed from simulations done under a certain null hypothesis, such as that of complete spatial randomness. The authors in [4] argue very strongly about the validity of simulation envelopes after pointing about some misconceptions in the literature concerning these. They also discuss pointwise envelopes and global envelopes, the former which has been the source of confusion concerning their interpretation. The authors caution on the interpretation and give the valid interpretation of these while arguing for their statistical validity.

A lot of work has been done on the Euclidean plane and the network space, as we have discussed above. However, Comas *et al.* in [13] argue for another angle on the issue of point patterns and linear networks that is quite relevant. In [13] the authors present a setting where the points do not lie exactly on the linear network. It is argued that in such cases the dependence between points and the linear network is not as apparent as the case where the points lie on the network, as in the applications given earlier on. They then propose another extension of Ripley's K-function where now instead of considering point to point distances, they consider point to linear network distances, that is a second order characterisation of the spatial relationship between points and a linear network. This was then applied to a data set consisting of human causes of fires in Asturias, in the north of Spain. Their results suggest that the fires tend to be within 500 metres of the road network.

It is on the work of the authors in [13] that we base this mini-dissertation. We have previously discussed the application of the cross-K function presented in [8] for crime locations relative to school locations. We investigate the spatial relationship between the crime locations and the linear network, that is, the road network. The area considered is the Gugulethu area in the Western Cape province of South Africa. This problem fits well with the technique developed in [13] because the crime locations do not lie exactly on the linear network as well, thus the dependence on the linear network, if any, may not be as apparent.

The structure of this mini-dissertation will be as follows. In Chapter 2 we consider spatial point processes and their various types. Furthermore, we consider line process theory and simulation because we shall be using this for simulating a linear network. In Chapter 3 we briefly consider tests for first order homogeneity for point patterns, and also how they may be applied to line patterns. Chapter 4 discusses the well known K-function and the cross-K function. We build on these discussions and also discuss the function proposed in [13] and derive an empirical estimate for it. We also discuss how confidence intervals might be calculated for this estimate. Furthermore, a discussion on Monte Carlo methods for spatial point patterns is presented, along with simulation envelopes for summary functions and formal hypothesis tests. Chapter 5 has a number of simulation studies where the ability of the function to pick up attraction of the points to the linear network and repulsion from the linear network is tested. We consider various point to line configurations and orientations, and non-stationary line patterns to investigate if this affects the function in any way. In Chapter 6 we apply the methods discussed largely in Chapter 4 on the aforementioned data set of crime locations in an attempt to understand the nature of the spatial relationship between crimes and the linear network in the Gugulethu area. Chapter 7 ends with a discussion of limitations and drawbacks, as well as future research in this area.

In summary, in this mini-dissertation we shall:

- discuss point processes and tests for first-order homogeneity
- discuss the K- and cross-K functions for unmarked and marked point patterns respectively,
- discuss the extension to these proposed in [13] along with confidence intervals for the estimate,
- discuss Monte Carlo methods including simulation envelopes and the maximum absolute deviation test and the deviation test considered in [19],
- investigate the powers of the tests considered including the case where the non-stationary Poisson line process was used to simulate a linear network,
- apply the methods discussed to a data set of crime locations in Gugulethu, South Africa.

A great contribution to spatial statistics is the book by Baddeley, Rubak and Turner [5]. This book is an importance reference for researchers keen on spatial point patterns. It includes an introduction to the software language `R` and their very own package `spatstat`. The package includes a plethora of methods for spatial point pattern analysis including summary functions which are part of descriptive statistics, all the way to statistical hypotheses and validation for fitted models which are part of inferential statistics. Indeed this book and the `spatstat` package will make multiple appearances in this mini-dissertation.

Chapter 2

Spatial point processes

2.1 Introduction

In this chapter some general theory of spatial point processes is discussed. The Poisson spatial point process is discussed at first because of the role it plays in most analyses involving point patterns. We also look at other mechanisms for generation of points such as the Matérn cluster process. The concepts of first-order homogeneity and second-order homogeneity are also explored. These two concepts, as will be seen, relate in some way to first and second-order moments of spatial point processes. The first-order moment relates to the mean number of points per unit area, and this can be constant or a spatially-varying function. The second-order moment relates to inter-point interaction, that is if the points influence the locations of each other. Lastly, we briefly introduce line processes, in particular Poisson line processes.

2.2 Intensity

2.2.1 First-order intensity

The idea of intensity is an important one when dealing with spatial point patterns. One simple definition of intensity is the average number of points per unit area [5]. Just like a random variable, X, in traditional statistics has first-order properties that relate to the mean, we have an analogous representation in spatial point processes. We have what is called the intensity function, λ, which is explained in [18] as describing the first order properties of a spatial point process,

$$\lambda(x) = \lim_{|dx| \to 0} \left(\frac{E\big[N(dx)\big]}{|dx|} \right) \tag{2.1}$$

where x is a spatial location on plane, dx is an infinitesimally small region and $N(dx)$ is the number of points in the region dx. This already has a familiar feel because if we ignore the limit we are essentially looking at the expected number of points divided by the area, denoted $|dx|$.

Alternatively we could drop the limit and multiply $|dx|$ by $\lambda(x)$,

$$|dx|\lambda(x) = E[N(dx)]$$

so that we have the expected number of points in region dx equal to $|dx|\lambda(x)$. We know that the region surrounding x is just a small part of the larger observation window D. We can therefore imagine working through all such small (and disjoint) partitioned areas of D, and multiply these by an appropriate $\lambda(x)$. These products can then be summed. If we increase the number of these disjoint regions, or equivalently reducing their individual areas, we get an approximation of the expected number of points of the point process $\boldsymbol{X}$, in D,

$$E[N(\boldsymbol{X} \cap D)] \approx \sum_{x \in D} |dx|\lambda(x)$$

which is essentially an integral,

$$E[N(\boldsymbol{X} \cap D)] = \int_D \lambda(x)dx \tag{2.2}$$

as argued in [5].

When the intensity function as shown in (2.1) is a function of the spatial location x, the intensity is said to vary spatially and is called inhomogeneous intensity. For example if one is studying crop distribution, it is possible that some areas of our observational window have favourable soil conditions meaning that crops are more likely to grow there. There will be a higher number of crops in such areas than other areas with less favourable conditions, hence the intensity function will vary spatially. An example of spatially varying intensity is seen in Figure 2.1 and is one of the example datasets from `spatstat(gorillas` dataset [5]). The figure shows 647 gorilla nesting sites on a polygonal window, and it is clear that we have inhomogeneous intensity. There are a number of tests for testing of inhomogeneous intensity which will be discussed in Section 3.1.

A more intuitive way of understanding an intensity function is found in [5]. It is argued that, for the 2-dimensional case, a plot of the intensity function $\lambda(x, y)$ over values of x and y results in a $3D$-plot whose height at a point (x_1, y_1) corresponds to the intensity. Therefore, following from (2.2) the integral over area D of the intensity function, i.e. the volume under the intensity function over area D gives the expected number of points falling in region D.

Estimation of the intensity function can be done non-parametrically using kernel estimation. Three commonly used kernel estimators namely, uncorrected, uniformly corrected and Diggle's correction are given in [5] and shown in Table 2.1, where u is a spatial location on the observation window.

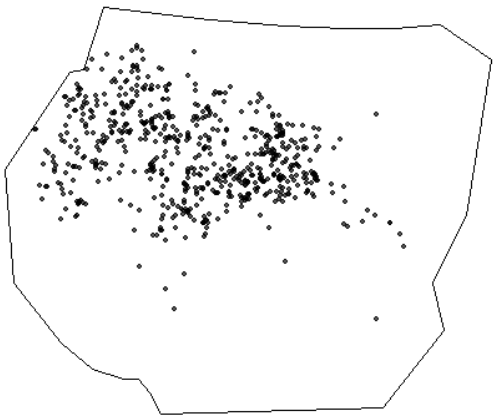


Figure 2.1: Gorilla nesting sites point pattern obtained from the `gorillas` dataset in the `spatstat` package. There are 647 gorilla nesting sites on this polygonal window.

Uncorrected	Uniformly corrected	Diggle's correction
$\lambda(u) = \sum_{i=1}^{n} \kappa(u - x_i)$	$\lambda(u) = \frac{1}{e(u)} \sum_{i=1}^{n} \kappa(u - x_i)$	$\lambda(u) = \sum_{i=1}^{n} \frac{1}{e(x_i)} \kappa(u - x_i)$

Table 2.1: Kernel estimators

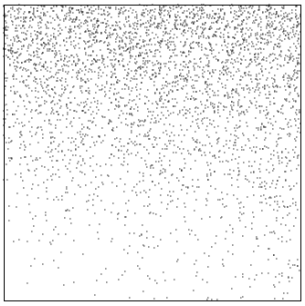
The term $e(u)$ is a correction for edge effects which will be discussed in Section 4.2.2 when the K-function is discussed. The kernel function, κ, used is a probability density, and the one usually used is the Gaussian density given by

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

where μ and σ are the mean and standard deviation respectively, although many other densities are possible. The smoothing bandwidth is the standard deviation of the Gaussian kernel, and controls how much smoothing is applied.

Figures 2.2 (a) and (d) show two realisations of a point process with spatially varying intensity functions. We can see clearly that the intensity is a function of location. This is more evident when we study the quadrat plots (b) and (e). There are more points in the upper quadrats in (b) than the lower quadrats for example. The same is true for (e) where the quadrats to the right have more points than those on the left. The density plots are shown for completeness where no edge correction was used and a bandwidth of 1 was used in the Gaussian kernel.

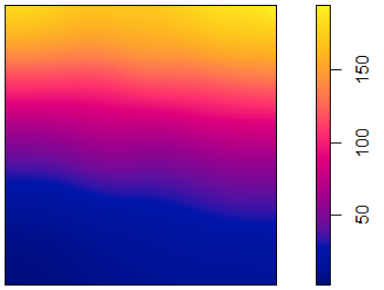
This leads us to the important idea of homogeneous intensity. A point process $\boldsymbol{X}$ is said to have homogeneous intensity if for a region $D \subset \mathbb{R}^2$ the expected number of points of $\boldsymbol{X}$ in D is proportional to the area of D, i.e. $\lambda|D|$ [5]. We note that the intensity function λ defined in (2.1) is independent of x, i.e. a constant. A homogeneous intensity of a point pattern means that the average number of points per unit area does not change as we move around our window of observation. Assuming that the point process is



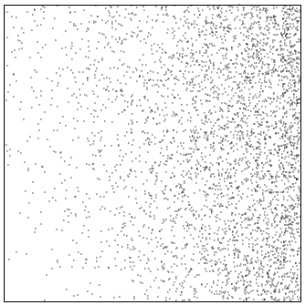
(a) $\lambda = (x + y^2)$

335	343	349	342	364
198	196	224	232	209
108	109	122	127	131
35	46	55	56	84
11	10	28	25	45

(b) Quadrat count



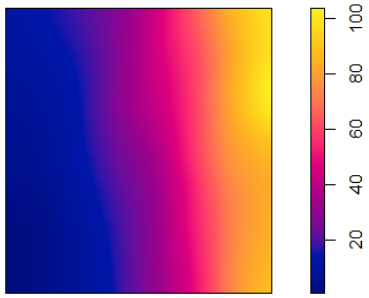
(c) Density plot



(d) $\lambda = (x^2 + y)$

41	84	100	227	387
36	81	136	214	337
30	55	116	239	361
17	52	112	202	308
10	35	102	192	336

(e) Quadrat count



(f) Density plot

Figure 2.2: Realisations of spatially inhomogeneous point patterns with quadrat counts and density plots. The point patterns were generated using the `rpoispp` function, the quadrat counts using the `quadratcount` function and the density plot using the `density` function, of the `spatstat` package.

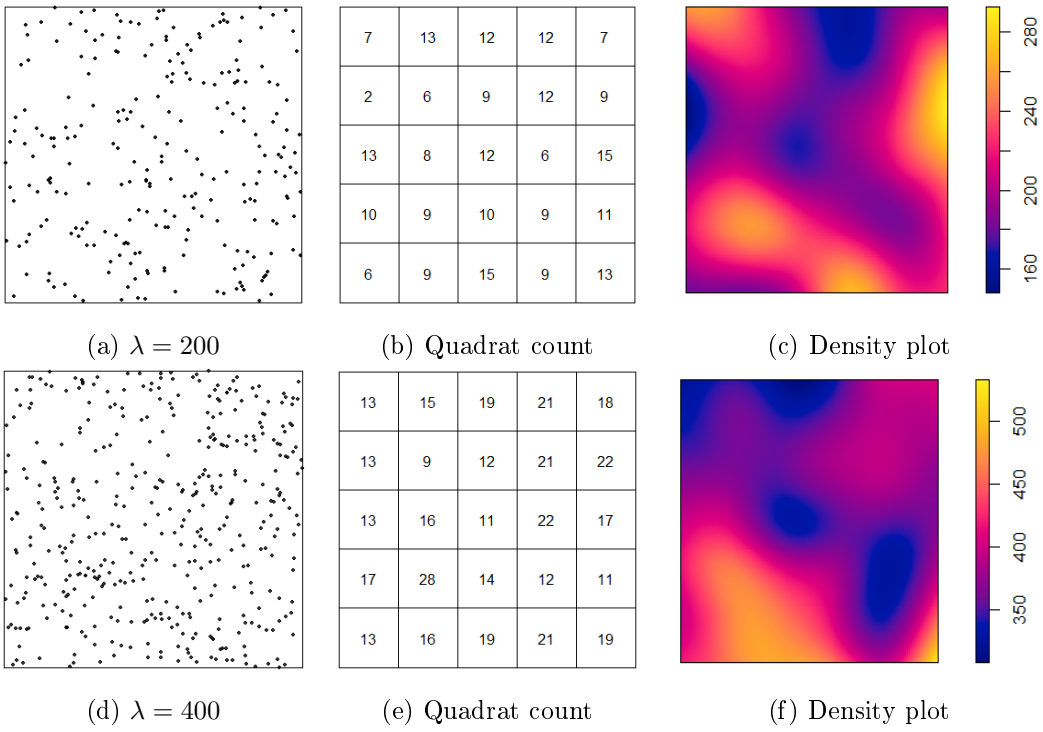
homogeneous λ can be estimated simply by

$$\hat{\lambda} = \frac{n(\boldsymbol{x})}{|D|} \tag{2.3}$$

where $\boldsymbol{x}$ is the observed point pattern, $n(\boldsymbol{x})$ the number of points observed and $|D|$ the area of the observational window [18]. We see that this is just the average number of points per unit area. Realisations of a point process with homogeneous intensity are shown in Figures 2.3 (a) and (d). The quadrat counts also reflect this homogeneity, with the counts not differing wildly as they did in the spatially varying intensity plot in Figure 2.2.

It is important to determine if a realised point pattern is spatially homogeneous. One reason for this is if one blindly applies tests for detection of inter-point interactions, like regularity and clustering, the tests may conclude that the points cluster around each other but the underlying reason of this distribution of points is spatial inhomogeneity [5].

A concept related to homogeneity is that of stationarity. These two concepts are sometimes used interchangeably but the definitions slightly differ. Stationarity is more concerned with how the statistical properties of a point process change as we move around the window. A point process $\boldsymbol{X}$ is called stationary if shifting the points of the point pattern does not change the statistical properties. That is $\boldsymbol{X}$ and $\boldsymbol{X} + u$



(a) $\lambda = 200$ (b) Quadrat count (c) Density plot

(d) $\lambda = 400$ (e) Quadrat count (f) Density plot

Figure 2.3: Realisations of spatially homogeneous point patterns on a 10×10 window with corresponding quadrat counts and density plots. The point patterns were generated using the `rpoispp` function, the quadrat counts using the `quadratcount` function and the density plot using the `density` function of the `spatstat` package.

are identical for any vector u [5]. It follows then that stationarity of $\boldsymbol{X}$ implies that $\boldsymbol{X}$ has homogeneous intensity [5]. This is easy to argue because homogeneity implies that the average number of points (a statistical property) over the observational window does not change as we move over the window.

Another concept that is usually mentioned with stationarity is that of isotropy. A point process is said to be isotropic if the statistical properties of the process do not change when the process is rotated about the origin [10]. A motion-invariant process is one that is stationary and isotropic [10]. It is worth noting that the homogeneous Poisson point process is motion-invariant.

2.2.2 Second-order intensity

Similar to the first order properties we have what is called a second-order intensity function as defined in [18]

$$\lambda_2(x,y) = \lim_{|dx|,|dy|\to 0} \left(\frac{E\big[N(dx)\big]E\big[N(dy)\big]}{|dx||dy|} \right), \tag{2.4}$$

where x and y are spatial locations on a plane. This definition is analogous to traditional statistics in that we may have two random variables X_1 and X_2 and may want to determine if they are related somehow

i.e. what their covariance or correlation is. In our case we are concerned with determining if the location of one point or event influences the position of the other, that is, are the distributions of these points dependent on each other? This is generally referred to as interpoint interaction [5]. This interaction can either be positive (which manifests as clustering), or negative (which manifests as regularity) [5]. As we have seen from the properties of CSR and from Section 2.4, patterns in which points do not show spatial dependency can be modelled with a Poisson process.

When we speak of second-order stationarity this implies that $\lambda_2(x, y) = \lambda_2(||x - y||)$ [18]. That is, a point process is second-order stationary if the second-order intensity function is a function only of the distance between the points. The idea of conditional intensity is also an important one particularly when we study the mathematical relationship between second-order moments and the K-function in subsequent sections. This conditional intensity is given by

$$\lambda_c(x|y) = \frac{\lambda_2(x, y)}{\lambda(y)}$$

which maybe be thought of as the intensity at the point x given that there is a point at location y [18].

2.3 Poisson spatial point process

Point process theory has been covered at great depth in various writings. Writers have different formulations of what a point process is, but all are largely similar. Cressie defines a spatial point process $\boldsymbol{X}$ as a model governing locations of events or points in $\mathbb{R}^d$ [14]. Note that $d = 2$ for our discussion. Diggle defines it as a stochastic mechanism that generates a set of countable events $x_1, x_2, x_3, ...$ on a plane [18]. We note, however, that we only observe a finite number of points so that this set has n points and not an infinite number. From these two definitions, any discussion in this topic will centre on *how* these points are distributed on the plane and, amongst other things, how they are located relative to each other. A point pattern is then defined as a realisation of a spatial point process. To aid understanding it may be helpful to think of the point process as a random variable X with a particular distribution in the traditional statistical sense. The realisation of X or the observed value x is then the same as our observed point pattern $\boldsymbol{x}$. In [10] the authors give two ways of defining a point process. The first manner is similar to what have have mentioned above, that i.e., it is thought of as a random sequence or a set of points $x_1, x_2, x_3, ...$ in $\mathbb{R}^2$. The second of these defines point processes as counting measures. That is, $\boldsymbol{X}(D)$ is the number of points of $\boldsymbol{X}$ that lie in the region $D \subset \mathbb{R}^2$. If we think of $\boldsymbol{X}$ as a set of points we can use the intersection operator between set $\boldsymbol{X}$ and D, so that $\boldsymbol{X} \cap D$ is the number of points of $\boldsymbol{X}$ that also belong to D.

The most trivial example of a point process is that containing one point. Let (x_1, y_1) be the cartesian coordinates of a point in $\mathbb{R}^2$. To generate this point one needs to generate two random numbers x_1 and

y_1. If the joint probability density of x_1 and y_1 is $f(x_1, y_1)$ then the probability that this point is in $D \subset \mathbb{R}^2$ is given by

$$P((x_1, y_1) \in D) = \int_D f(x_1, y_1) dx_1 dy_1$$

Now, when we say a point is uniformly distributed on D, this means that the coordinates x_1 and y_1 have a joint probability density that is constant inside the spatial window D and zero everywhere else. This density is given by

$$f(x_1, y_1) = \begin{cases} \frac{1}{|D|} & \text{if } (x_1, y_1) \in D \\ 0 & \text{otherwise} \end{cases}$$

where D is the area of the observational window D. Note that this holds only if $|D| \neq 0$ and if $|D| < \infty$ i.e. $|D|$ is non-zero and finite. The case when D is a rectangle is again a simple case, but it aids discussion. If $D = [0, x_{max}] \times [0, y_{max}]$ then $(x_1, y_1) \in D$ if $x_1 \in [0, x_{max}]$ and $y_1 \in [0, y_{max}]$. Thus if x_1 is uniformly distributed on $[0, x_{max}]$ and y_1 is uniformly distributed on $[0, y_{max}]$ and x_1 and y_1 are independently distributed, then (x_1, y_1) is uniformly distributed point on D.

As mentioned above, in investigations involving points, the main question posed is about the distribution of the points on the plane. There are generally three possibilities or configurations of point patterns [14]. Given a point pattern on an observational window D:

1. The points could be distributed randomly.
2. The points could be in clusters across the window.
3. There could be some spacing between the points such that the distances from point to point are always greater than a certain distance.

Of particular interest is the first possibility listed above related to the property of *complete spatial randomness*. The properties associated with a point process that is completely spatially random are given in both [5] and [18]. One should note that in some texts CSR is synonymous with the homogeneous Poisson process and we will keep with this convention. Note also that the intensity of a homogeneous Poisson process is given by λ. The properties that are given by Baddeley *et al.* in [5] are as follows. Suppose $\boldsymbol{X}$ is the point process and D is the observational window, then the following properties hold concerning CSR:

1. The number of points of the process contained in a region D, i.e. $N(X \cap D)$ has a Poisson distribution;
2. The expected number of points in D is proportional to the area of D, i.e. $|D|$, and is given by $E[N(\boldsymbol{X} \cap D)] = \lambda \times |D|$;
3. The number of points falling in disjoint areas of D i.e. $N(\boldsymbol{X} \cap D_1)$, $N(\boldsymbol{X} \cap D_2)$, $N(\boldsymbol{X} \cap D_3)$, ...,$N(\boldsymbol{X} \cap D_Q)$, are independent random variables;

4. Given n i.e., the number of points on D, these points constitute a random sample and are uniformly distributed.

While (perfect) CSR is near improbable when studying a real-life dataset as discussed in [18], writers in the literature suggest that testing for CSR should be the first step before we continue with other statistical tests such as those of spatial clustering. Put differently, testing for complete spatial randomness should form part of the researcher's preliminary descriptive analysis [14, 18]. This is an important step because if we test for CSR and do not reject the null hypothesis of CSR, there is hardly a reason to continue with any other spatial analyses [18].

The spatial distribution of points may look different if viewing subjectively. One person can claim randomness while another claims some spatial clsutering. Take for example Figure 2.4. This shows nine completely random spatial point patterns generated in the `spatstat` package in `R` using the `rpoispp` function with $\lambda = 100$, on a $[0,1] \times [0,1]$ window. It is not an easy visual task to determine if these patterns are CSR, especially since in some areas the points look to be clustering together. One should remember however that these points are independently and uniformly distributed despite this deceiving layout of apparent clustering. Just because there is a point at location x for example does not in any way affect the probability that there will be a point in location $x + \epsilon$, where $\epsilon \in \mathbb{R}^2$. The distributions of their locations are still independent [5].

Having said the above it is therefore important to have a concrete statistical way of determining whether or not points are indeed completely spatially random. There are various methods for testing for CSR outlined in the literature such as in [14]. Since the number of points in CSR is important (seen clearly in points 2 and 3 in the properties of CSR), Cressie [14] claims that in the early days of point pattern study the focus was the quadrat counts $N(\boldsymbol{X} \cap D_1)$, $N(\boldsymbol{X} \cap D_2)$, $N(\boldsymbol{X} \cap D_3)$, ..., as is still the case even today. These quadrat counts are then compared to the Poisson distribution. If these counts deviate significantly one could conclude that the pattern is not completely spatially random. So we can see that counts play an important role in studying point patterns. This then leads us to the well-known χ^2 goodness-of-fit test which will be discussed briefly in Section 3.1. There are other tests outlined in the literature such as those that utilise the Poisson distribution property of the equality of the mean and variance of the distribution to quantify the departure from CSR [14].

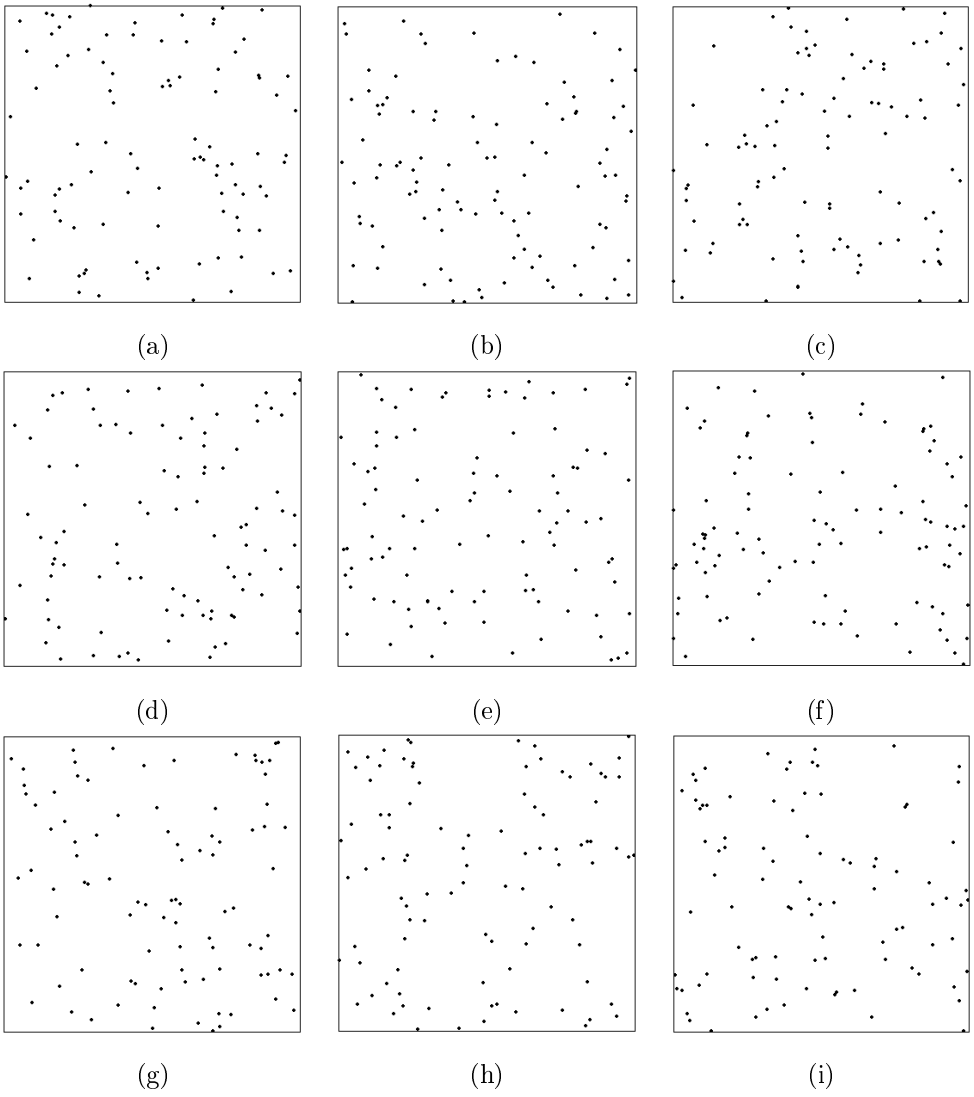


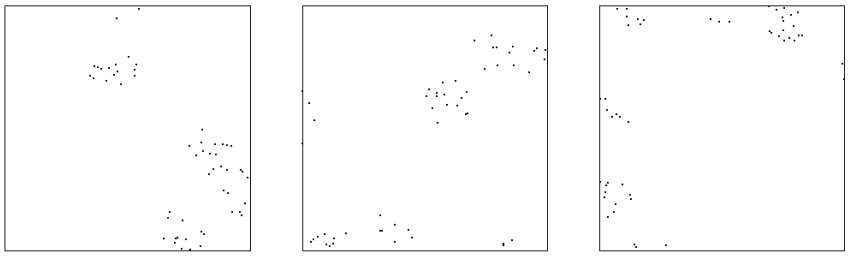
Figure 2.4: Realisations of CSR simulated on a unit square window using the `rpoisspp` function of `spatstat` with $\lambda = 100$.

2.4 Other spatial point processes

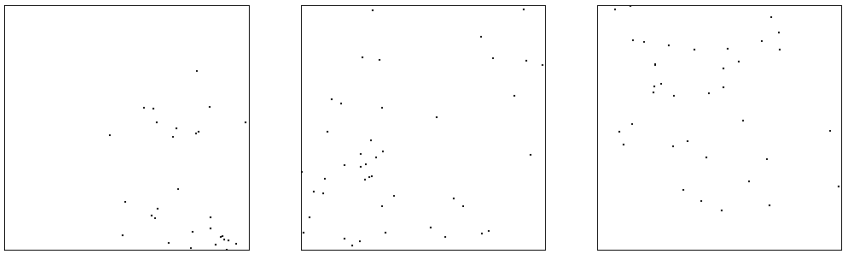
The Poisson process model is just one of the many spatial point processes. Recall that a key property of the homogeneous Poisson process is that the points are independent, implying further that there are no interactions between the points. Therefore if we have some interaction or spatial dependency between the points then the point process is not a Poisson process [5].

Recall that we mentioned that there are three possible configurations of point patterns, i.e. CSR, clustering and regularity. One mechanism of generating clustering points is the Matérn cluster process [5]. To generate points from this mechanism three parameters are needed. The first is the intensity of a homogeneous Poisson process κ. So-called 'parent' points are first generated from a homogeneous Poisson

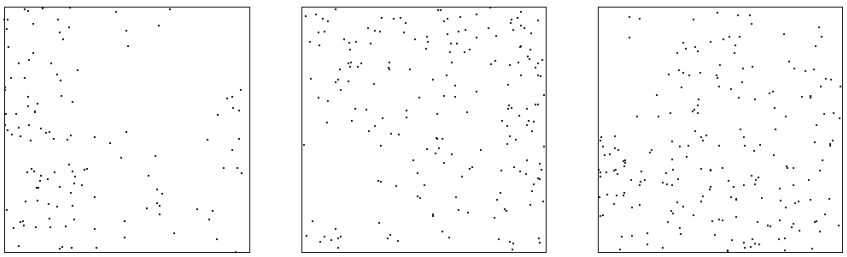
process with parameter κ. Each parent then has a number of offspring, and that number has a Poisson distribution with mean μ. The offspring are then independently and uniformly distributed over a disc of radius r, centered at the original parent point. It is clear that there is some dependency between the points formed from one parent. The point process then has an intensity of $\kappa\mu$. Figure 2.5 shows different realisations of this process for different parameter values, generated using the `rMatClust` function. It is clear, especially for the patterns with a smaller number of points that there is some clustering. For the other point patterns where it is less apparent there are statistical tests in place to test for clustering of point patterns which will be discussed in a later section. The next point pattern we consider is a regular



(a) $\kappa = 5$, r=0.10, μ=10



(b) $\kappa = 5$, r=0.25, μ=10

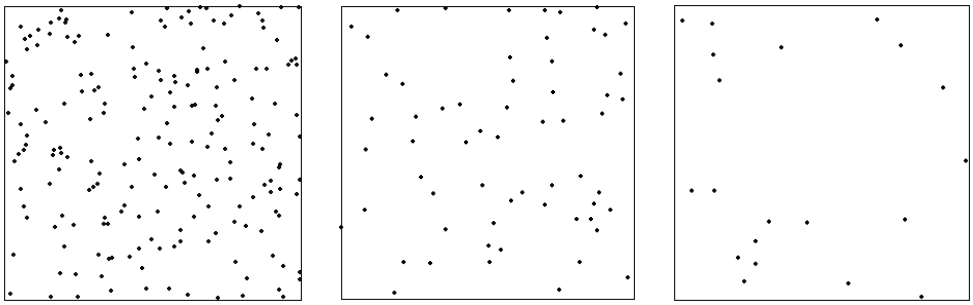


(c) $\kappa = 10$, r=0.25, μ=20

Figure 2.5: Different realisations of the Matérn cluster process using the `rMatClust` function on a $[0,1] \times [0,1]$ window.

point pattern. A mechanism for generating such points is Matérn's Model I which uses dependent thinning [5]. We first generate a homogeneous Poisson process. Secondly each point in the generated point pattern that is closer than a specified distance r from its nearest neighbour is removed. Point patterns realised from this process are shown in Figure 2.6. A homogeneous Poisson process with parameter 2 was used to

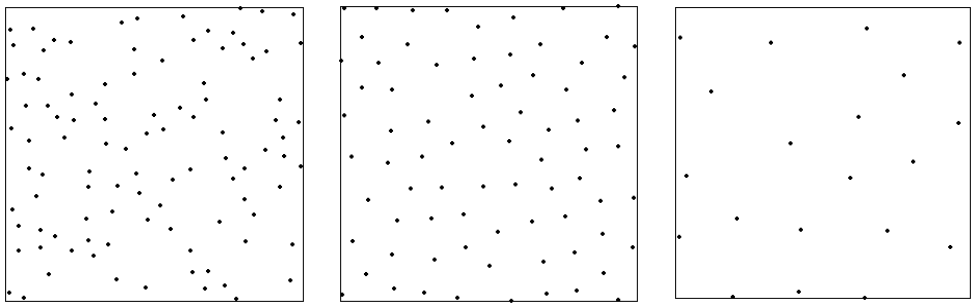
generate these patterns, with values of r shown below the images.



(a) $r = 0.10$ (b) $r = 0.4$ (c) $r = 0.6$

Figure 2.6: Realisations of Matérn's Model I generated using the `rMaternI` function in `spatstat` with Poisson parameter 2 on a $[0, 10] \times [0, 10]$ window.

There is also the Simple Sequential Inhibition (SSI) process [5]. In this mechanism we start with a window D and we generate a point uniformly on the window, independent of the previously generated points. A point is then rejected if it lies within r units of another point on the window. This is continued until we reach a predetermined number of points or once the algorithm reaches a certain maximum number of attempts. Realisations of this process are shown in Figure 2.7. The points were generated on a unit square with the specified minimum distance and a request of 100 points, using the `rSSI` function. As can be seen in Figures 2.7(b) and (c) the algorithm could not fit the desired 100 points because of the increased inhibition distance r and the constraint of the window.



(a) $r = 0.05$, $n = 100$ (b) $r = 0.10$, $n = 76$ (c) $r = 0.2$, $n = 20$

Figure 2.7: Realisations of the SSI model generated using the `rSSI` function in `spatstat` on a unit square window.

A real-world example of a point pattern that may exhibit regularity is seen in Figure 2.8. This dataset is obtained from the `R spatstat` package and shows the locations of Japanese pines on a 5.7×5.7 metre square window [5]. The authors in [5] mention that one possible reason for this distribution of points is the competition of resources. Trees will tend to avoid growing near each other to increase the probability of getting enough nutrients for themselves, essentially creating a barrier between an individual tree and other trees.

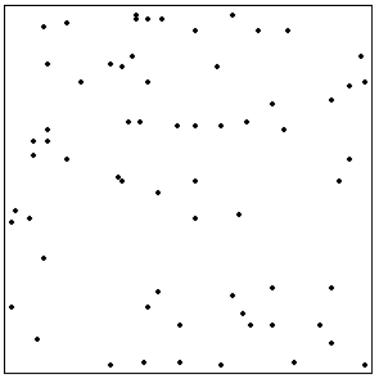


Figure 2.8: Locations of Japanese pines from the `japanesepines` dataset in `spatstat`.

2.5 Multitype point patterns

In this section we consider a multitype point process which consists of points of different types. In multitype point patterns we have the usual location of events $x_1, x_2, ..., x_n$, and corresponding marks $m_1, m_2, ..., m_n$ for each point. There is a myriad of ways points can be categorised depending on the context of the study.

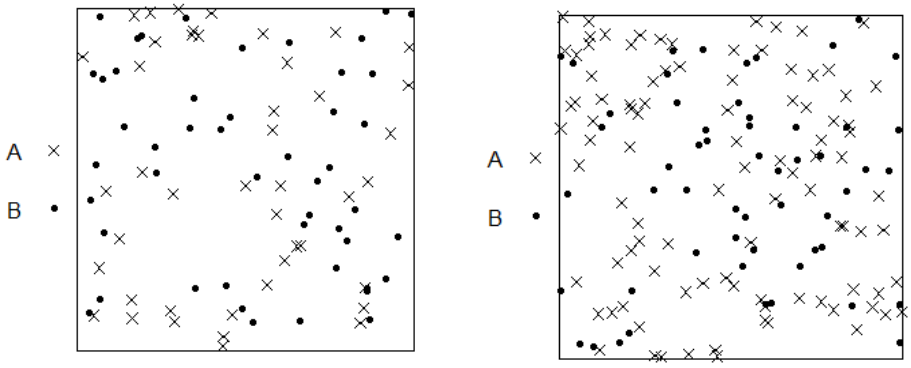
The most elementary process to consider is again associated with the Poisson process. In the construction of multitype Poisson process with N marks, we assume points of type i form a homogeneous Poisson process as discussed in Section 2.3, with intensity λ_i. More concretely, there are N homogeneous Poisson processes $\boldsymbol{X}_1$,$\boldsymbol{X}_2$,...,$\boldsymbol{X}_N$, with corresponding intensities $\lambda_1, \lambda_2, ..., \lambda_N$. It is further assumed that points of different types are independent. Then we have a homogeneous multitype Poisson process [5]. An alternative construction is to suppose that the unmarked points form a homogeneous Poisson point process with intensity λ^*. If we now randomly assign each point to mark i with probability p_i independently of other points, then we also have a homogeneous Poisson process, where type i points have intensity λ_i.

The CSR counterpart for multitype points is called *complete spatial randomness and independence* (CSRI) [5]. The properties of this characterisation are given in [5] as:

1. The unmarked locations of points constitute a homogeneous Poisson process;
2. The labels m_i are completely random, i.e. they are independent and identically distributed;
3. The sub-point process $\boldsymbol{X}_i$ of type i points is a homogeneous Poisson process, for $i = 1, .., N$;
4. The sub-point processes $\boldsymbol{X}_1$,$\boldsymbol{X}_2$,...,$\boldsymbol{X}_N$ are independent point processes.

The intensity referred to as λ^* is the overall intensity of the unmarked point process, i.e. $\lambda^* = \sum_{i=1}^{N} \lambda_i$. Furthermore, $p_i = \lambda_i/\lambda^*$ is the probability that a point is of type i [5]. Another important note is that

while the multitype point process is homogeneous, the mark distribution is not uniform, i.e. we can have varying probabilities p_i. Two simulated multitype point patterns are shown in Figure 2.9. In Figure 2.9 (a) both the sub point processes are homogeneous Poisson point processes with intensity 50. As expected, the number of points from each type is nearly the same: there are 44 points of type A and 48 points of type B. In Figure 2.9 (b) type A points have an intensity of 100 (with 97 points), and type B have intensity 50 (with 55 points).



(a)$\lambda_A = \lambda_B = 50$ (b) $\lambda_A = 100$, $\lambda_B = 50$

Figure 2.9: Two realisations of the multitype Poisson point process generated using the `rmpoispp` function in `spastat` on a [0,1]× [0,1] window.

A point pattern can also have multiple marks associated with each location. The dataset locations of fires in New Brunswick, Canada called `nbfires` found in `spatstat` is one such example and it has nine marks. Two of the marks include the type of fire (forest, grass, dump or other) and the year in which the fire was recorded (1987 through 2003). The data is on a polygonal window with an enclosing rectangle of [0, 1000] × [0, 958.9142] units, where one unit is 0.403716 km. A plot of locations of fires for the year 1998 with associated marks of fire types is shown in Figure 2.10.

The concepts of stationarity still apply to multitype point process. A multitype point process is stationary if the processes that make it up are stationary as well [5]. For example with the fire locations shown in Figure 2.10 would be considered stationary if the four point patterns shown in Figure 2.11 are each stationary. It also follows from the definition that a multitype point process is stationary if the unmarked point pattern is stationary [5]. Estimation of the intensity function, use of kernels and other ideas discussed in Section 2.2.1 are all applicable to multitype point processes. It is up to the user to decide if the analyses of intensity should be done for each type, or if the marks should be dropped and focus on the unmarked point pattern.

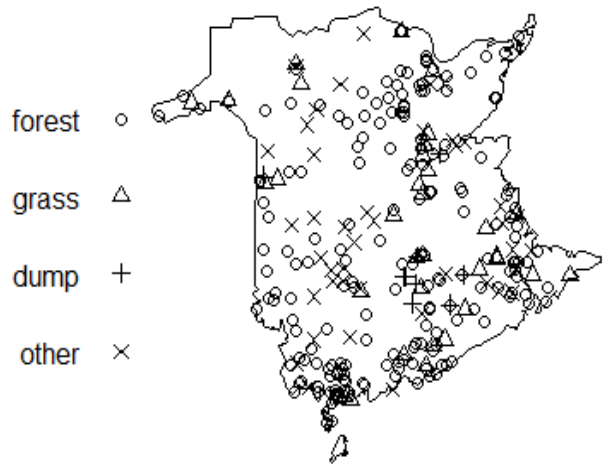
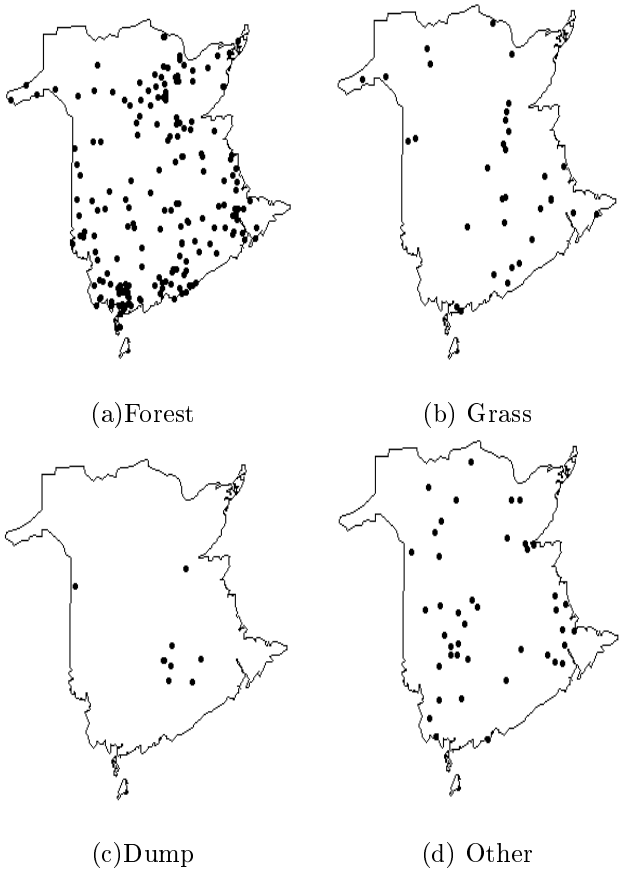


Figure 2.10: Plot of locations on fires from `nbfires` dataset in `spatstat` for the year 1998.



(a)Forest (b) Grass

(c)Dump (d) Other

Figure 2.11: Locations of fire types for the year 1998 from `nbfires` dataset in `spatstat`

An analogous definition of second-order intensity discussed in Section 2.2.2 for bivariate point pattern is

defined in [18] as

$$\lambda_{ij}(x,y) = \lim_{|dx|,|dy|\to 0} \left(\frac{E\big[N_i(dx)N_j(dy)\big]}{|dx||dy|} \right)$$

where $N_i(dx)$ and $N_j(dx)$ are the number of type i points in region dx and the number of type j points in region dy respectively.

2.6 Line processes

2.6.1 Theory

The natural progression from the distribution of points is to consider lines and their distributions. This subject is dealt with in various sources such as [10], on which we base the following discussions. A line process is defined as a random collection of lines on a plane. A single line can then be described by two parameters:

1. The perpendicular distance p from the origin
2. The angle θ between the perpendicular line p and the x-axis measured in the anti-clockwise direction.

The perpendicular distance p may take any positive real number i.e. $p \in [0, \infty]$, while the angle $\theta \in [0, 2\pi]$. Thus these two parameters can be seen as coordinates on a cylindrical surface $C = [0, 2\pi] \times [0, \infty]$. This C is called the representation space in the literature. It is then easy to see that if we have points on C, we can get a corresponding set of lines in $\mathbb{R}^2$, showing that there is a one-to-one relationship between the lines and points on C. Thus an alternative definition of a line process is a random collection of points in the representation space C.

As was the case for point patterns, line processes have the concept of stationarity. Denote a line process in $\mathbb{R}^2$ by $\boldsymbol{X}_l$. Let the perpendicular distance of line l_i from the origin be p_{l_i}, and the angle this line makes with the x-axis be θ_{l_i}. A line process is stationary if the translated line process has the same statistical (or distributional) properties as the original line process [9]. Stationarity can also be thought of from the representation space C as shown in [10]. That is, the line process $\boldsymbol{X}_l$ is stationary if the point process on C is stationary as well. More concretely, consider a point process $\boldsymbol{X}$ on C: $\{\big((p_{l_1}, \theta_{l_1}), (p_{l_2}, \theta_{l_2}), ...\big)\}$. The translated point process obtained by shifting the line process through distance t parallel to the x-axis results in a new set of coordinates for the original point process: $T(\boldsymbol{X}_l) : \{\big((p_{l_1} + t\sin(\theta_{l_1} + \alpha), \theta_{l_1}\big), \big((p_{l_1} + t\sin(\theta_{l_2} + \alpha), \theta_{l_2}\big), ...\}$ on C. If the original point process $\boldsymbol{X}$ has the same statistical properties as the translated point process $T(\boldsymbol{X}_l)$ for all translations T, then $\boldsymbol{X}_l$ is stationary [9].

A concept that is specific to line processes is that of line density which is the mean line length per unit

area [9]. If the line process $\boldsymbol{X}_l$ is motion-invariant then the line density of the process is $\mu = \lambda 2\pi$, where λ is the intensity of the point process on C.

It follows naturally that a Poisson line process is directly linked to a Poisson point process on C. A Poisson line process is a line process generated by a Poisson process on C [10]. A Poisson line process that is stationary, or motion-invariant, is such that the angle θ is uniformly distributed on $[0, 2\pi]$, [10].

2.6.2 Simulating a line pattern

The classical way of simulating random lines is one of the solutions to the well-known Bertrand's paradox and various authors have written about this problem, such as in [33]. Briefly, the following question is posed: if a random chord is thrown at a circle with an inscribed equilateral triangle, what is the probability that the length of the chord is greater than the length of the side of the equilateral triangle. Figure 2.12 shows the setting of the problem (constructed using the GeoGebra tool[1]). The circle drawn is of radius 4 and BCD is an equilateral triangle, where the side length is 6.9cm. The chord in this case is line EF and has length 6.7cm. The question then, in this context, is what is the probability that EF is greater than BD (equivalently BC and CD).

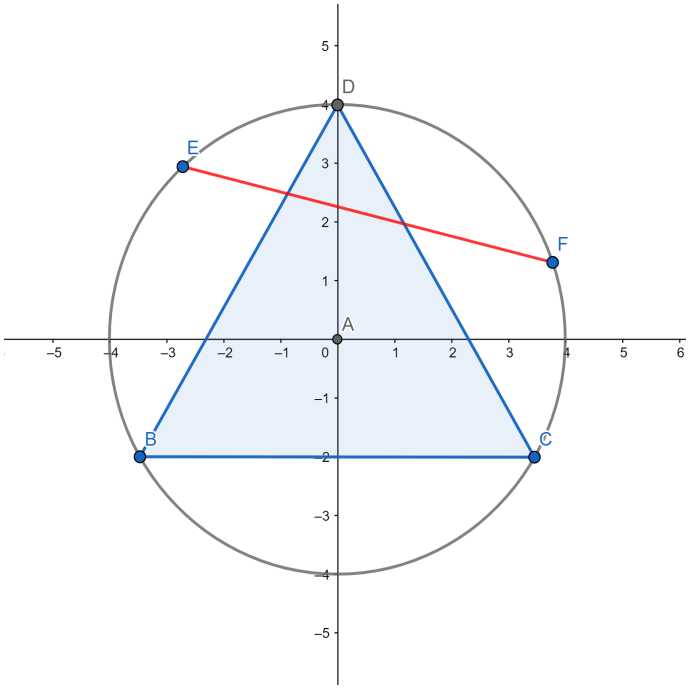


Figure 2.12: An equilateral triangle with side length 6.9cm inscribed in a circle of radius 4.

Three solutions to the problem were derived by Bertrand depending on how the word 'random' was

[1] https://www.geogebra.org

interpreted.

First approach First we consider two points on the circumference of the circle. We assume these two points are uniformly distributed on the circumference. These two points are joined to form a chord. The resulting probability as calculated by Bertrand and many others was $\frac{1}{3}$.

Second approach In this approach the midpoint of the chord is considered random. Two points are generated uniformly and independently on the circle and this point is the midpoint of the chord. In this case the probability is calculated to be $\frac{1}{4}$.

Third approach Finally, we consider a point on the circumference which will represent the angle. Then we generate uniform random variable between 0 and r which will be the center of the chord. The chord is then the perpendicular line through this point on the line. This is shown in Figure 2.13. B is the point selected on the circumference, and so the angle is FAB. Point C is the uniformly chosen point in $(0, r)$. The chord DE is then constructed perpendicular to the radius AB, through point C. The probability in this case is then $\frac{1}{2}$.

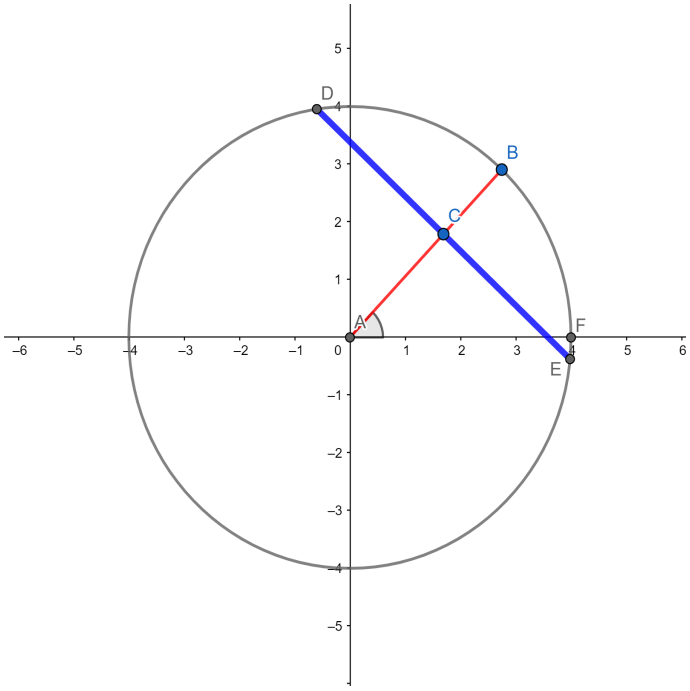


Figure 2.13: Illustration of the third approach. The chord is the perpendicular line through point C which lies on the radius, and was generated uniformly on $(0, r)$.

This third approach is the method used to generate a realisation of a Poisson line process in the `spatstat` package in `R` using the `rpoisline` function. Having done the above, i.e. obtain the chord, what remains is to find the coordinates of the end points of the chord, which will be the starting and ending points of the

simulated line. To get the coordinates then we consider Figure 2.14. Let the magnitude of $\overrightarrow{AC}$ be P and the magnitude of $\overrightarrow{CD}$ be Q. Note that by properties of chords, $|\overrightarrow{CD}|=|\overrightarrow{CE}|$. The coordinates of interest

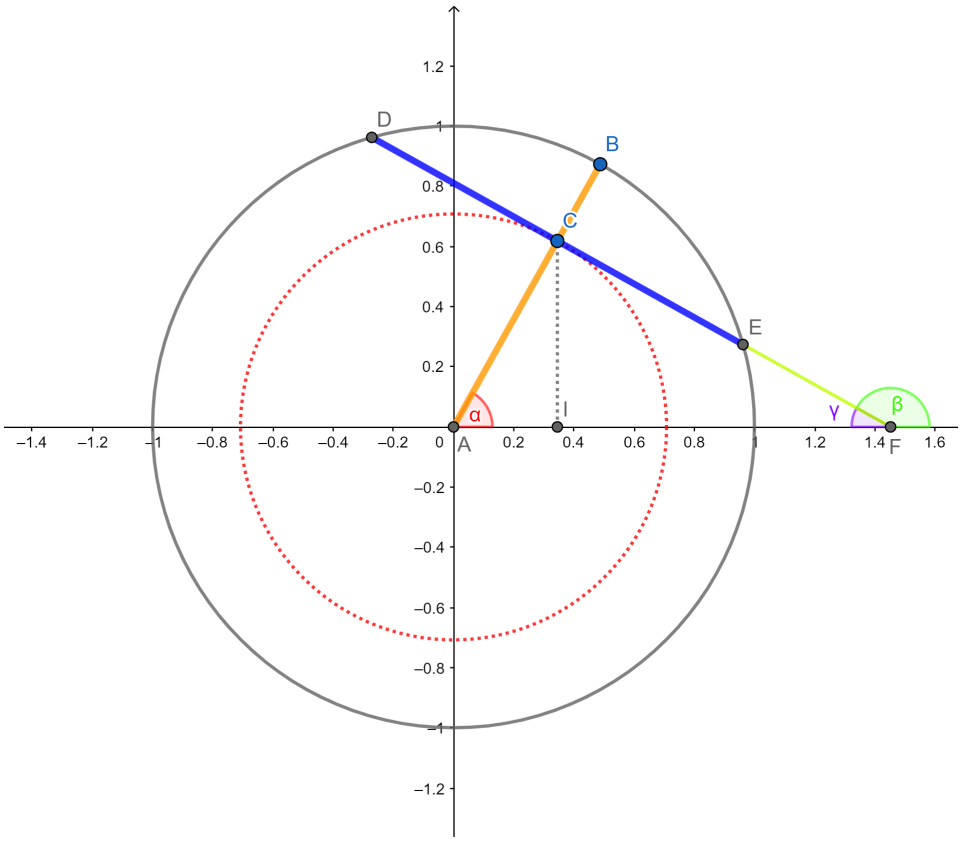


Figure 2.14: Illustration of the important angles α, γ and β and their relationship to the chord $\overrightarrow{DE}$. α is the angle the perpendicular bisector of DE, i.e., $\overrightarrow{AC}$ makes with the x-axis. β is the angle $\overrightarrow{DF}$ makes with the x-axis in anti-clockwise fashion, and $\gamma = \pi - \beta$.

are D and E. Using trigonometry rules we can find the vector $\overrightarrow{AC}$. We first determine the coordinates of the point C, (x_1, y_1). By trigonometry rules

$$\begin{aligned} \cos(\alpha) &= \frac{x_1}{P} \\ \sin(\alpha) &= \frac{y_1}{P} \end{aligned}$$

so that $x_1 = P\cos(\alpha)$ and $y_1 = P\sin(\alpha)$. Therefore $\overrightarrow{AC} = (P\cos(\alpha), P\sin(\alpha))$ in component form.

To find the vector $\overrightarrow{EC}$ we use standard results from trigonometry again. That is, the vector in component form is the magnitude multiplied by $(\cos(\beta), \sin(\beta))$ where β is the angle it makes with the x-axis in the counter-clockwise direction. In our case this magnitude is Q, thus the vector $\overrightarrow{EC}$ is of the form $(Q\cos(\beta), Q\sin(\beta))$. Note that $\gamma + \beta = \pi$ and $\alpha + \gamma = \frac{\pi}{2}$ and since $\cos(\beta) = -\cos(\pi - \beta) = -\cos(\gamma)$ and $\sin(\beta) = \sin(\pi - \beta) = \sin(\gamma)$, the component form then becomes

$$\overrightarrow{EC} = (-Q\cos(\gamma), Q\sin(\gamma)).$$

Finally, since $\cos(\gamma) = \sin(\frac{\pi}{2} - \gamma) = \sin(\alpha)$ and $\sin(\gamma) = \cos(\frac{\pi}{2} - \gamma) = \cos(\alpha)$, the final component form in terms of the angle α is

$$\overrightarrow{EC} = (-Q\sin(\alpha), Q\cos(\alpha)).$$

Therefore to get E we add vectors $\overrightarrow{AC}$ and $\overrightarrow{CE}$ (or - $\overrightarrow{EC}$, note the sign change because of direction change)

$$\begin{aligned} \overrightarrow{AC} + \overrightarrow{CE} &= (P\cos(\alpha), P\sin(\alpha)) + (Q\sin(\alpha), -Q\cos(\alpha)) \\ &= (P\cos(\alpha) + Q\sin(\alpha), P\sin(\alpha) - Q\cos(\alpha)). \end{aligned}$$

To get D, we use the fact that vectors $\overrightarrow{EC}$ and $\overrightarrow{CD}$ have the same magnitude and direction, so that $\overrightarrow{CD} = (-Q\sin(\alpha), Q\cos(\alpha))$, giving

$$\begin{aligned} \overrightarrow{AC} + \overrightarrow{CD} &= (P\cos(\alpha), P\sin(\alpha)) + (-Q\sin(\alpha), Q\cos(\alpha)) \\ &= (P\cos(\alpha) - Q\sin(\alpha), P\sin(\alpha) + Q\cos(\alpha)). \end{aligned}$$

Therefore in summary, to generate random Poisson lines:

1. Generate an angle α uniformly on $[0, 2\pi]$,
2. Generate a point P uniformly from $[0, r]$,
3. Calculate $Q = \sqrt{r^2 - P^2}$,
4. (a) Calculate starting coordinates of the line $x_0 = P\cos(\alpha) + Q\sin(\alpha)$, $y_0 = P\sin(\alpha) - Q\cos(\alpha)$
   (b) Calculate end coordinates of the line $x_1 = P\cos(\alpha) - Q\sin(\alpha)$, $y_2 = P\sin(\alpha) + Q\cos(\alpha)$.

A final note about how this is implemented in `spatstat`, the user enters a parameter λ and a window. A maximum r value is calculated from the dimensions of the window, and this is just the diagonal of the window i.e. $\sqrt{\text{height}^2 + \text{width}^2}$. This value will be used in step 3 in the summary above. The number of lines generated is from a Poisson$(2\pi\lambda r_{max})$ distribution. This is because $2\pi r_{max}$ is the circumference of the circle with radius r_{max}, so it is scaled by the parameter λ like it is done when working with Poisson. The angle and P are generated as above and the coordinates calculated as shown. Note that the coordinates in steps $4(a)$ and (b) are shifted by the midpoint of the window range in the x and y directions respectively: $x_0 = x_{mid} + P\cos(\alpha) + Q\sin(\alpha)$, $y_0 = y_{mid} + P\sin(\alpha) - Q\cos(\alpha)$, and the same for the other two. This is because all the calculations are done assuming a circle with the origin for the centre so you have to shift, if for example the window is a $[5, 10] \times [5, 10]$.

Now, there are interesting alterations towards non-stationarity that can be made to the way the line pattern is simulated. These alterations include:

1. Use a non-uniform probability distribution for the angle α, while using a uniform distribution for the distance P.

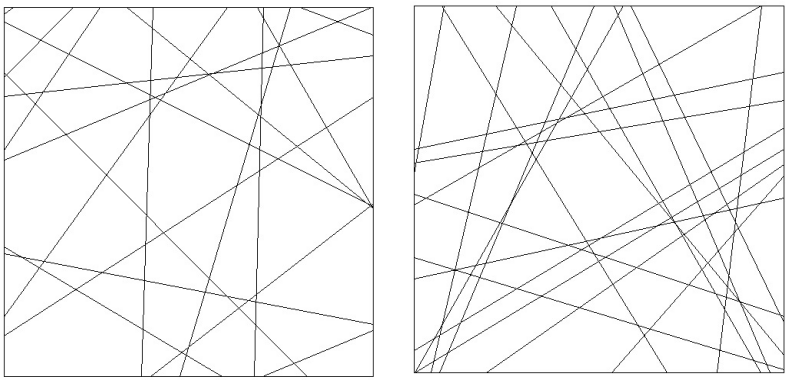
2. Use a non-uniform probability distribution for the distance P, while using a uniform distribution for the angle α.

3. Use non-uniform probability distributions for both α and P.

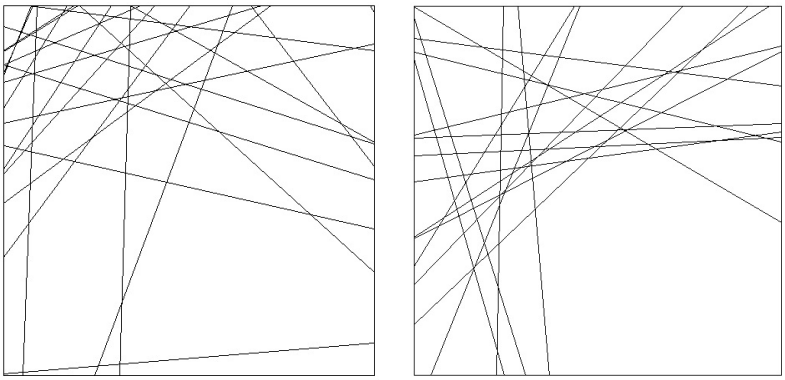
We illustrate some of the visible differences between these approaches. Figure 2.15 shows four results of simulations. The simulations are done on a $[0,1] \times [0,1]$ window. Figure 2.15(a) shows a realisation of a line process using the standard `rpoisline` function without any alterations. Figure 2.15(b) shows a realisation of a line process if p is taken to be non-stationary, specifically from a Beta distribution while the angle is taken from a uniform distribution. The choice of the distribution is not important, but it was chosen particularly because it is defined on [0,1] so we can easily scale the value of p as the size of the window changes. The Beta distribution used is shown in Figure 2.16, with parameters 2.7 and 6.3. We see that this is positively skewed with a mean of 0.3 and median of 0.284. The effect of this on the resultant line pattern is that the perpendicular distance p is shorter than when the standard `rpoisline` function was used. That is, the magnitude of $\overrightarrow{AC}$ in Figure 2.14 will be shorter and consequently, the Q (the magnitude of $\overrightarrow{CE}$) will be larger. The latter point can also be seen easily from the way Q is calculated i.e. $Q = \sqrt{r^2 - P^2}$.

Figure 2.15(c) shows a realisation of a line process if the perpendicular distance is generated from a uniform distribution and the angle taken from a Beta(2.7,6.3) distribution. The effect of this alteration is very apparent in the resultant point pattern because the angle affects the orientation of the line. Looking at Figure 2.16 again we see that the positive skewness of the distribution means that smaller angles will be chosen more often. The mean of this distribution, scaled by 2π to ensure we get values between $[0,2\pi]$, is $1.885rad$ compared to the mean of the uniform $3.142rad$. The median of the angles generated using the Beta distribution is 1.784 compared with the median of 3.142 when using the uniform distribution. The effect of using the Beta distribution on Figure 2.14 is that the angle α will be smaller than if it were generated from a uniform distribution. Finally, Figure 2.15(d) shows a realisation of the process when both p and α are non-stationary. The effect that is more apparent is that of the Beta distribution imposed on the angle.

For completeness we also consider the case where the Beta distribution is negatively skewed as in Figure 2.17, with parameters 6.3 and 2.7. The results of this change are seen in Figure 2.18. Again, Figure 2.18(a) shows the standard case. Figure 2.18(b) shows the case where p is generated from the distribution in Figure 2.17, while the angle is generated from a uniform distribution. The mean of this distribution is 0.7 and the median is 0.716. The effect of this is that the magnitude of $\overrightarrow{AC}$ in Figure 2.14 will be larger and consequently, the Q (the magnitude of $\overrightarrow{CE}$) will be smaller than if the uniform distribution was used.



(a) Stationary p and α (b) Non-stationary p



(c) Non-stationary α (d) Non-stationary p and α

Figure 2.15: Different line configurations simulated on a unit square window. The non-stationary parameters are simulated using the positively skewed Beta(2.7,6.3) distribution.

In the same thinking as before, if a non-stationary negatively skewed distribution is imposed on the angle, the angles will be larger on average than in the case of a uniform distribution. The mean of the angles from this distribution is 4.398 and the median is 4.499. A line pattern generated from using a stationary p and non-stationary angle is shown in Figure 2.18(c). Then finally, the case where both p and the angle are non-stationary and both from a negatively skewed distribution is shown in Figure 2.18(d).

We shall return to the importance of these alterations in Section 5.1 where a number of simulations are conducted.

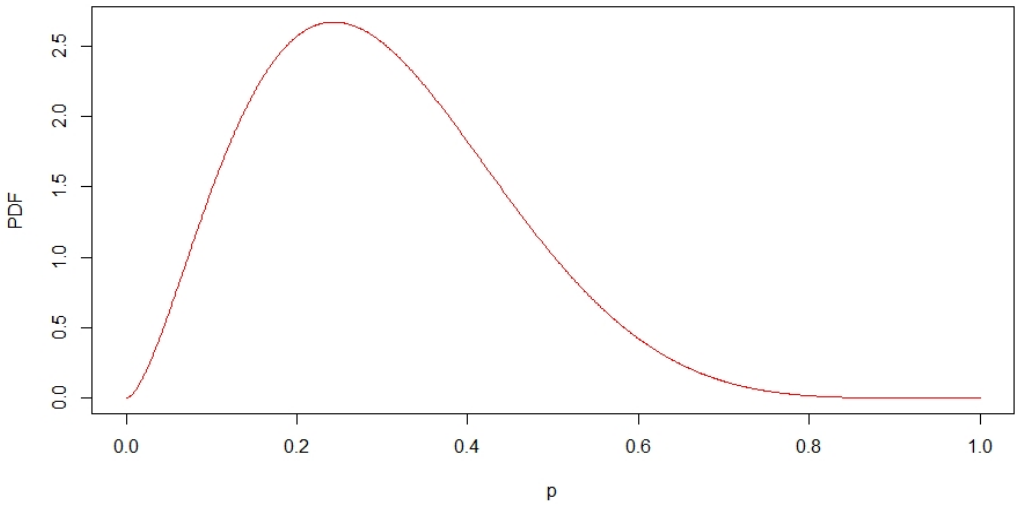


Figure 2.16: Positively skewed Beta(2.7,6.3) distribution.

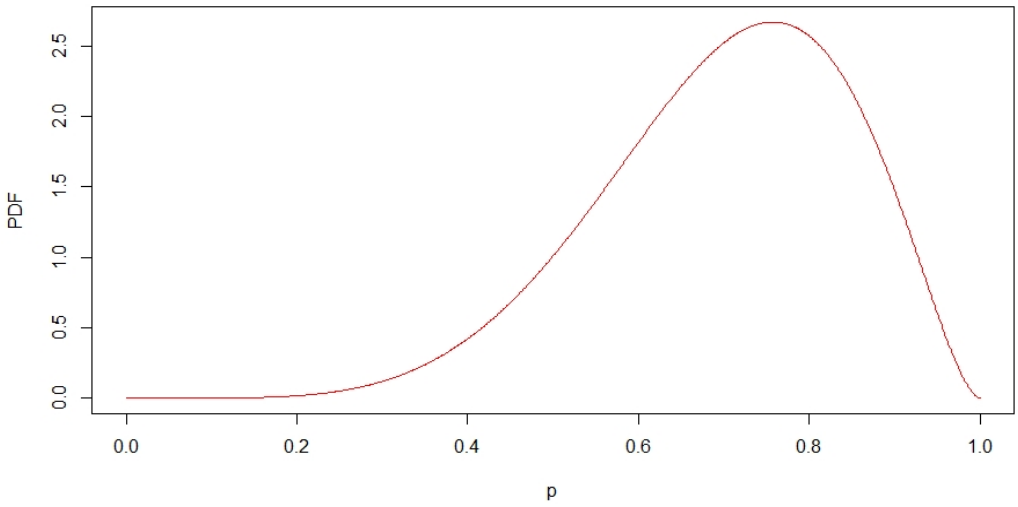
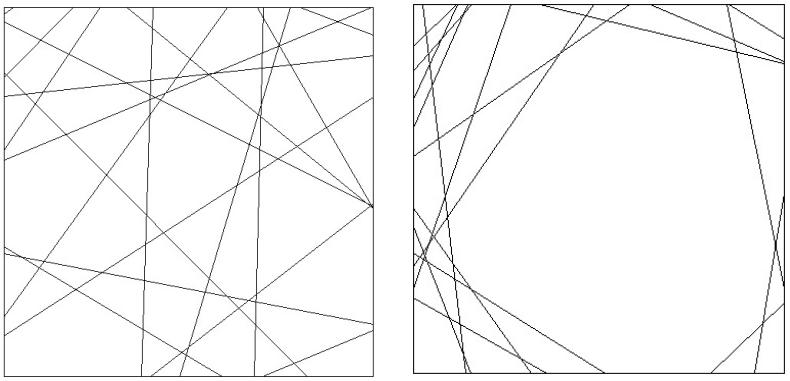
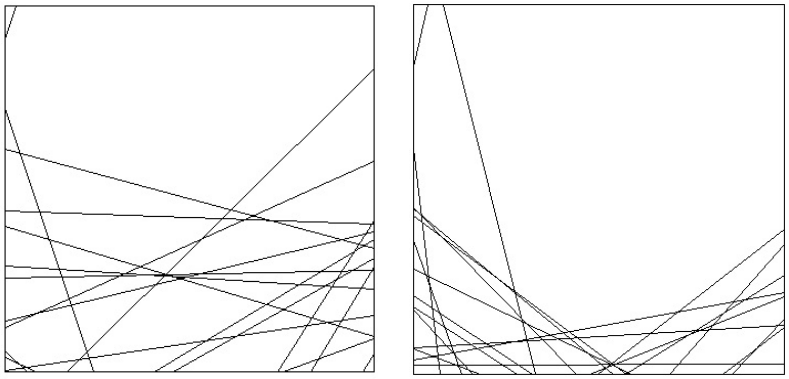


Figure 2.17: Negatively skewed Beta(6.3,2.7) distribution.



(a) Stationary p and α (b) Non-stationary p



(c) Non-stationary α (d) Non-stationary p and α

Figure 2.18: Different line configurations simulated on a unit square window. The non-stationary parameters are simulated using the negatively skewed Beta(6.3,2.7) distribution.

2.7 Conclusion

In this chapter we considered a number of spatial point processes, including the well-known Poisson point process. Multitype point processes, the case where we have marks associated with the points, were also discussed. For both univariate and multitype point processes, the intensity functions were discussed and the conditions under which we have stationarity and non-stationarity of the point patterns. The line process which is an extension of previously discussed point processes was presented along with the analogous ideas of stationarity. Non-stationary line processes were generated assuming non-stationary angles or non-stationary perpendicular distances, or both. The non-stationarity of the Poisson line process will play a role when we conduct a number of simulations in Section 5.1. We next consider briefly tests for first order homogeneity for point processes and how these might be extended for line processes.

Chapter 3

Tests for first-order homogeneity

In this chapter we consider tests for first-order homogeneity. These tests are briefly summarised in [25, 26], where references for the main contributors in these tests are included. The various tests are given in Section 3.1 below briefly, and more information can be found in the references mentioned along with them. In Section 3.2 we discuss a modified test for first order homogeneity proposed in [25, 26]. A few simulations are also conducted to test this proposed method, and we also consider how these tests may be extended for line processes.

3.1 Tests for first-order homogeneity

To set the scene we suppose that we have an observed point pattern. The observational window D is divided into say, Q quadrats, $D_1, D_2, ..., D_Q$. For each quadrat we count the number of points that fall within that quadrat. For brevity, we write $n_i = n(\mathbf{x} \cap D_i)$ i.e. the number of points of the point pattern falling in quadrat region D_i. The area of D_i is given by $|D_i|$. We recall that for the Poisson homogeneous patterns, the number of points, N_i, falling in region i is Poisson distributed with mean $\lambda|D_i|$. Furthermore, under H_0 the number of points falling in these disjoint subregions are independent random variables, essentially implying that we have Q random variables $N_1, N_2, ..., N_Q$. Now, let $\lambda_1, \lambda_2, ..., \lambda_Q$ be the Poisson parameters for each subregion. Under the null hypothesis of spatial homogeneity these values are equal i.e. $H_0 : \lambda_1 = \lambda_2 = ... = \lambda_Q$. This means that each N_i follows a Poisson distribution with $\lambda_i = \lambda$. The alternative hypothesis is that the intensity is inhomogeneous, i.e. $H_A : \lambda_i \neq \lambda_j$ for at least one pair of $i, j \in \{1, 2, ..., Q\}$ [5, 25, 26].

The writers in [25] and [26] propose a different method of testing for homogeneity which we will get to shortly. We now state these tests following a similar manner to [25] and [26].

1. The Pearson χ^2 test The test statistic for this test is given by:

$$\chi^2 = \sum_{i=1}^{Q} \frac{(n_i - \hat{\lambda}|D_i|)^2}{\hat{\lambda}|D_i|}. \tag{3.1}$$

Under H_0, the test statistic given by (3.1) has an asymptotic χ^2 distribution with $Q-1$ degrees of freedom. One should note that this test could be used either to test for the goodness-of-fit for the Poisson distribution assuming homogeneous intensity, or to test homogeneity having assumed independence [5]. The focus for our case is the latter.

One of the drawbacks of this test is discussed in [5] and it is the claim that the alternative hypothesis is too broad. That is, H_A is that the process is not a homogeneous Poisson process. Therefore it is possible that H_0 is rejected because the process does not have homogeneous intensity or because of the violation of independence between the points. Hence the rejection of the null hypothesis leaves a large range of possibilities. This holds for the tests 2-5 below as well.

2. Potthof and Whittinghill's test

Found in [31] is the first of the two tests by these authors. First we define:

$$V = \left(\sum_{i=1}^{Q} |D_i|\right) \sum_{i=1}^{Q} \frac{n_i(n_i-1)}{|D_i|}.$$

The test statistic is given by

$$VT = eV + f \tag{3.2}$$

where

$$e = \frac{2(Q-1)}{\left(\sum_{i=1}^{Q}|D_i|\right)\left(\sum_{i=1}^{Q}|D_i|^{-1}\right) - 3Q + 2 + 2(Q-1)(\sum_{i=1}^{Q} n_i - 2)}$$

and

$$f = e[(Q-1)e-1]\left(\sum_{i=1}^{Q} n_i\right)\left(\sum_{i=1}^{Q} n_i - 1\right).$$

Note that VT in (3.2) is distributed as χ^2 with degrees of freedom $e^2(K-1)(\sum_{i=1}^{Q} n_i)(\sum_{i=1}^{Q} n_i - 1)$.

3. Potthof and Whittinghill's second test

First define

$$U = \sum_{i=1}^{n} n_i^2 - \sum_{i=1}^{n} n_i - 2\lambda \sum_{i=1}^{Q} |D_i| n_i.$$

Then the test statistic is given by

$$UT = gU + h$$

where

$$g = \frac{\sum_{i=1}^{Q}|D_i|^2}{\sum_{i=1}^{Q} 0.5|D_i|^2 + \lambda \sum_{i=1}^{Q}|D_i|^3}$$

and

$$h = g(g+1)\lambda^2 \sum_{i=1}^{Q} |D_i|^2.$$

If λ is unknown, as is usually the case, the following estimate is used

$$\lambda^* = \sqrt{\frac{\sum_{i=1}^{Q} n_i^2 - \sum_{i=1}^{Q} n_i}{\sum_{i=1}^{Q} |D_i|^2}}.$$

4. Likelihood ratio statistic

Discussed in [11] is the likelihood ratio statistic. This is given by

$$LRS = 2\left(\sum_{i=1}^{Q} n_i \ln \frac{n_i}{|D_i|} - \sum_{i=1}^{Q} n_i \ln \frac{\sum_{i=1}^{Q} n_i}{\sum_{i=1}^{N} |D_i|} \right) \tag{3.3}$$

where for quadrats with zero counts the convention $0 \ln 0 = 0$ is employed [11]. Then (3.3) has a χ^2 distribution with $Q-1$ degrees of freedom.

5. The Score test

The test statistic is given by

$$SC = \left(\frac{\sum_{i=1}^{Q} n_i}{|D_i|} \right)^2 \left(\sum_{i=1}^{Q} \frac{|D_i|^2}{n_i} \right) - \sum_{i=1}^{Q} n_i. \tag{3.4}$$

Note that we add 0.5 to every n_i if the observed count is zero in any one n_i to avoid division by zero. Under the null hypothesis, (3.4) has an asymptotic χ^2 distribution with $Q-1$ degrees of freedom.

For these tests we make use of parametric bootstrapping. In this method the p-value is approximated using Monte-Carlo resamples generated under the null hypothesis of homogeneity [11]. First the test statistic is calculated based on the observed counts n_i, denoted τ. Next, Q Poisson random values from a Poisson$(\hat{\lambda}|D_i|)$ distribution for $i = 1, ..., Q$, using the estimate for λ, $\hat{\lambda} = n(\mathbf{x})/|D|$. We then calculate the test statistic for this sample generated under H_0, denoted τ^*. This process is done T times and the bootstrap p-value is given in [11] as

$$\text{bootstrap } p - \text{value} = \frac{\sum_{i=1}^{T} I(\tau_i^* \geq \tau) + 1}{T+1}. \tag{3.5}$$

Thus the null hypothesis of homogeneity is rejected if the bootstrap p-value is less than the significant level α. Authors in [11] studied size distortions of the tests stated above, considering asymptotic and parametric bootstrap tests. For point patterns with a large number of points, all but the asymptotic UT test are recommended. The asymptotic UT test is not recommended because of its conservative nature. For a small number of points, none of the parametric bootstrap tests using the five test statistics perform consistently better or worse than the others. However, Pearson's χ^2 test and the VT test statistic are recommended for a small number of points over the other three tests because of their mild size distortions. Further discussion on these findings can be found in [11].

3.2 Modified test for first order homogeneity

The test for homogeneous intensity is modified in [25, 26] to account for possible spatial dependence in the data. The argument is made that assuming that the counts are independent is not a concrete assumption and this assumption is especially violated in the study of spatial point patterns because points close to each other are likely to be dependent [5, 26]. Therefore the test for homogeneity is modified by using a random subset of 50% of the quadrats instead of all the quadrats to calculate the test statistic. Furthermore the number of times we repeat this process is limited to 99 to avoid the scenario where independence is violated since if T were 999 for example, some of the samples would have dependence depending on the number of quadrats used [26]. This method for testing for first-order homogeneity is now summarised.

Steps to follow in testing for first-order homogeneity

1. We obtain the point pattern of interest, i.e. the one we want to determine whether it is stationary or not;
2. Calculate the MLE for λ, assuming that the pattern is homogeneous under H_0;
3. For $i = 1$ to 99:
   (a) Simulate a homogeneous point pattern on the same window as the original data using the MLE for λ
   (b) Sample 50% of the quadrats
   (c) Calculate the test statistics based on the simulated point pattern and the sampled quadrats, i.e. calculate τ_i^* for each of the tests
   (d) Calculate the test statistics based on the original point pattern, and the same 50% sampled quadrats τ
4. Calculate the bootstrap p-value for each test statistic as:
$$\text{bootstrap } p - \text{value} = \frac{\sum_{i=1}^{99} I(\tau_i^* \geq \tau) + 1}{99 + 1}$$
5. Reject the null hypothesis of homogeneity if the p-value is less than α.

3.2.1 Simulations

In this section we do a few simulations to show that this modified test works well. However, we shall not delve too deeply into the simulations. Extensive simulations are conducted in [25, 26] to show that this modified test performs satisfactorily and the reader is referred to those references for a more complete study.

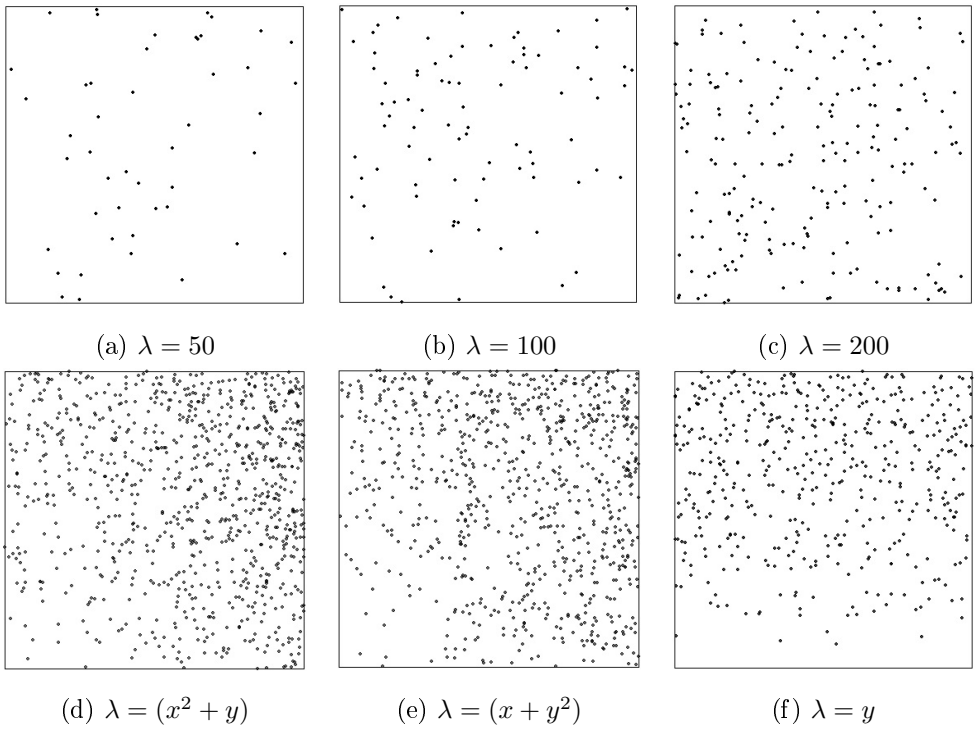
We start with homogeneous Poisson point patterns on a $[0,1] \times [0,1]$ window with different values of λ. We then simulate non-stationary Poisson point patterns again on the unit square for different intensity functions. The realisations for the different set-ups are shown in Figure 3.1. The results of the tests are summarised in Table 3.1. As expected, for the three cases with constant intensity functions the null hypothesis of stationarity is not rejected and thus we conclude that the point patterns are stationary. The opposite is true for the other three patterns. All the p-values are less than 0.05 so the null hypothesis of stationarity is rejected at a 5% level and we conclude that the point patterns are non-stationary, as expected.

Label	$\lambda(x)$	χ^2 test	P & W test	P & W test 2	LR statistic	Score test
(a)	50	0.53	0.39	0.44	0.51	0.41
(b)	100	0.49	0.39	0.4	0.5	0.34
(c)	200	0.24	0.24	0.31	0.24	0.32
(d)	$\lambda = (x^2 + y)$	0.01	0.01	0.01	0.01	0.01
(e)	$\lambda = (x + y^2)$	0.01	0.01	0.03	0.01	0.01
(f)	$\lambda = y$	0.01	0.01	0.01	0.01	0.01

Table 3.1: p-values for test of homogeneity for different point patterns on a unit square window, where the intensity function is $\lambda(x)$ and the Label column corresponds to the labels given in Figure 3.1.

As mentioned in Section 2.6.1, a line pattern in $\mathbb{R}^2$ is considered stationary if the point pattern on the representation space $C = [0, 2\pi] \times [0, \infty]$ is stationary as well. Therefore we can extend the tests to test for stationarity of a line pattern. Figure 3.2 shows line patterns in $\mathbb{R}^2$ with the corresponding point pattern on the representation space which in this case is $C = [0, 2\pi] \times [0, r_{max}]$, where $r_{max} = \sqrt{1^2 + 1^2} = 0.707$ since the simulations are done on a unit square window.

Table 3.2 shows the results of the stationarity tests applied to the point patterns on C shown in Figure 3.2. As one would expected the null hypothesis of stationarity is not rejected for the point pattern on C corresponding to Uniform p and α. The null hypothesis of stationarity is rejected for all the other cases where one or both of p and α were non-stationary and we then conclude that the point patterns, and consequently the line patterns, are non-stationary for these three cases.



(a) $\lambda = 50$ (b) $\lambda = 100$ (c) $\lambda = 200$

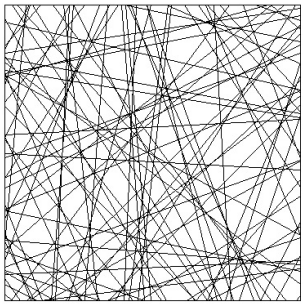
(d) $\lambda = (x^2 + y)$ (e) $\lambda = (x + y^2)$ (f) $\lambda = y$

Figure 3.1: Realisation of Poisson point patterns on a unit square window generated using the `rpoispp` function with the intensity functions as arguments.

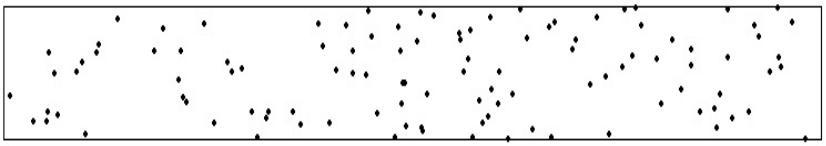
.

	χ^2 test	P & W test	P & W test 2	LR	Score test
Stationary p and α	0.36	0.37	0.45	0.36	0.4
Non-stationary p and stationary α	0.01	0.01	0.07	0.01	0.01
Stationary p and non-stationary α	0.01	0.01	0.02	0.01	0.01
Non-stationary p and non-stationary α	0.01	0.01	0.01	0.01	0.01

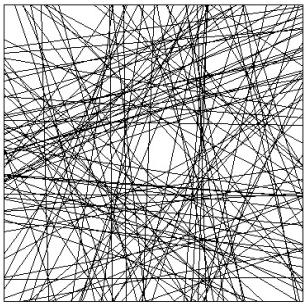
Table 3.2: p-values for test of homogeneity for different line configurations.



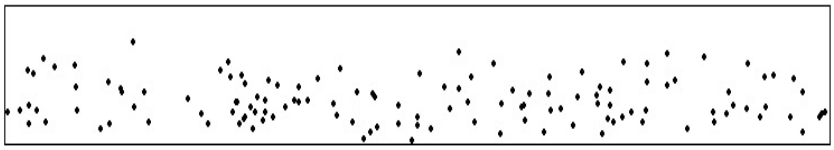
(a) Stationary line pattern



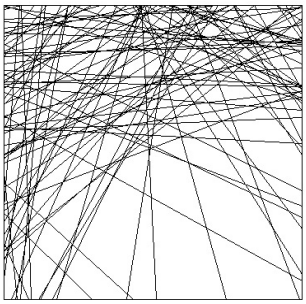
(b) Stationary representation



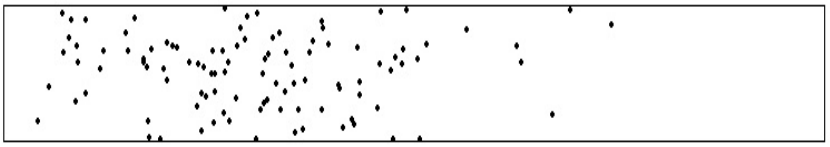
(c) Non-stationary p



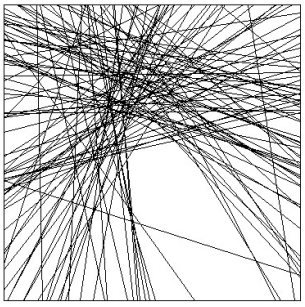
(d) Non-stationary representation



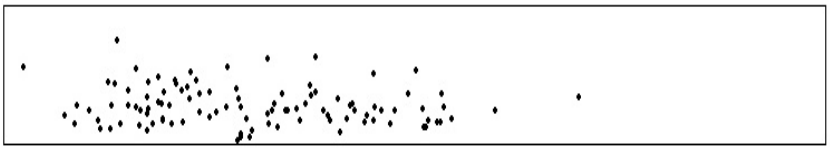
(e) Non-stationary α



(f) Non-stationary representation



(g) Non-Stationary p and α



(h) Non-stationary representation

Figure 3.2: Line patterns in a [0,1]×[0,1] window $\subset\mathbb{R}^2$ with the corresponding representation on $C = [0, 2\pi] \times [0, r_{max}]$. The horizontal axis corresponds to $[0, 2\pi]$ and the vertical axis corresponds to $[0, r_{max}$.]

3.3 Conclusion

In this chapter we briefly considered various tests of first-order homogeneity for point patterns, particularly the modified version proposed by Kraamwinkel in [25]. We investigated how these might be extended for line processes by making use of the representation space $C = [0, 2\pi] \times [0, r_{max}]$. To our knowledge, this is the first application of these tests to line patterns. We now move on from first-order properties to discuss second-order properties in Chapter 4.

Chapter 4

Second order properties

4.1 Introduction

In Chapters 2 and 3, the focus was mainly on the first-order properties of point processes. An important area of point process analysis is that of point-to-point interactions, as briefly discussed in Section 2.4. For instance, we may ask questions about the influence of one point's location on another point within its vicinity. This property was briefly discussed in Section 2.2.2 in the discussion of second-order properties of point processes. In the following sections we explore characterisations of the second-order properties of point processes, starting with the simplest case, the K-function in section 4.2, then the cross-K function in Section 4.3. Finally we build on these ideas to study spatial interactions between points and linear structures in Section 4.4.

4.2 The K-function

4.2.1 Theoretical definition

As mentioned in Chapter 1, the K-function is another way to characterise the second-order properties of a point process. The second-moment measures relate to covariance and correlations in traditional statistics. Just like a random variable X in the traditional sense can have a positive correlation with another random variable Y, such positive correlations exist in the study of point processes. In particular positive associations manifest when we have a point pattern that has clusters [5]. Recall the Matérn cluster process from Section 2.4. There is a positive association between points in the same cluster because they are generated from a 'parent' point. The points are in a way forced to be within a certain distance r of

the parent point as though the parent point has a drawing influence on the offspring points.

The two random variables X and Y can again have a negative association [5]. Matérn's Model I which uses dependent thinning [5] discussed in Section 2.4 as well generates a point pattern that exhibits negative association. As was discussed, this mechanism generates a point pattern that no point is within a distance r of another point. This is negative association since the individual points influence the location of other points. In a way the points are pushing against each other.

The last and perhaps least interesting case is when X and Y have zero covariance. This is the CSR case of point patterns that was discussed in Section 2.3. The points in this case in no way interact with each other. They have no influence on the location of other points within our window of interest.

It is with good reason that we first studied the definitions of stationarity and homogeneity because as Baddeley *et al.* mention in [5], in order to accurately study correlation one must have estimated the mean i.e. the intensity in our case. This, as argued in [5], is important if the researcher is to avoid issues of spurious correlation.

Now, to define the K-function, let $\mathbf{X}$ be a stationary and isotropic point pattern with intensity λ. This is an important assumption of the K-function. The K-function for the point process $\mathbf{X}$ is defined by [5, 18, 14],

$$K(r) = \lambda^{-1} E[\text{Number of points within distance } r \text{ of } x | \mathbf{X} \text{ has a point at } x]. \tag{4.1}$$

This function is defined for $r \geq 0$. The importance of the stationarity assumption is seen above since we divide by a constant, non-spatially-varying intensity λ. Alternatively we can think of the expectation in terms of the number of additional points within distance r of an arbitrary point [18]. For understanding of an arbitrary point we refer to Diggle in [18]. We suppose we have a large number of events $n < \infty$ in a finite region $|D| < \infty$. An 'arbitrary' event then is a point selected randomly from these n points.

As we have already mentioned, the K-function is related to the second-order moment $\lambda_2(x, y)$ defined in equation (2.4). We again assume that we have a stationary point process, and recall that this implies that $\lambda_2(x, y) = \lambda_2(||x - y||) = \lambda_2(r)$. To obtain this relation between the K-function and the second-order property we refer to Diggle in [18]. We assume we have a simple point process. This is defined as a point process for which no two events can occur at the same location [10]. To compute the number of additional points within a distance r of an arbitrary point we integrate $\lambda_2(x|o)$ (the intensity at location x given that there is a point at the origin o) over a circle with centre o and radius r.

Note that since we assumed stationarity, the following holds

$$\lambda_c(x|o) \quad = \quad \frac{\lambda_2(x, o)}{\lambda(o)} = \frac{\lambda_2(x - o)}{\lambda} = \frac{\lambda_2(x)}{\lambda}. \tag{4.2}$$

Therefore using equation (4.2) and following [18] we have

$$\begin{aligned}\lambda K(r) &= \int_0^r \int_0^{2\pi} \lambda_c(x|o) x d\theta dx \\ &= 2\pi \int_0^r \frac{x\lambda_2(x)}{\lambda} dx.\end{aligned}$$

We re-arrange the above to get

$$(2\pi)^{-1}\lambda^2 K(r) = \int_0^r x\lambda_2(x)dx. \tag{4.3}$$

Finally, we apply the Fundamental Theorem of Calculus and re-arrange further giving

$$\lambda_2(r) = (2\pi r)^{-1}\lambda^2 K'(r). \tag{4.4}$$

From these calculations and particularly (4.3) and (4.4) we see the relationship between the K-function and second-order properties.

4.2.2 Estimation of the K-function

Equation (4.1) is the theoretical definition of the K-function. It is very seldom that we will know this in practice therefore we need to estimate it for an observed point pattern. When the K-function was defined in equation (4.1) we made mention of the number of points within 'distance r'. This gives us an indication that in our estimation point-to-point distances will play a major role. For this discussion and subsequent ones, unless stated otherwise, $\mathbf{x}$ is the observed point pattern. The observed point pattern $\mathbf{x}$ has n points $\{x_1, x_2, ..., x_n\}$ where each $x_i \in \mathbb{R}^2$, so that $x_i = (x_{i1}, x_{i2})$ (although we shall just stick to the less cumbersome notation x_i). Let the Euclidean distance between two points x_i and x_j be d_{ij}, that is $d_{ij} = ||x_i - x_j||$. As usual D will be our window of observation. Now, as argued in [5] if we consider pair-wise distances they may contain some information on the configuration of the points. For example, if the average of the pairwise distances is small, this might be indicative of clustering. If the point pattern constitutes of regularly spaced points, it is reasonable that the average of the pairwise distances is larger, or at least larger than what we would expect under CSR.

We now derive the estimate of the K-function following the arguments in [5]. The starting point is the following summation

$$\hat{P}(r) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i} I(d_{ij} \leq r) \tag{4.5}$$

where $I(d_{ij} \leq r) = \begin{cases} 1 & \text{if } d_{ij} \leq r \\ 0 & \text{if } d_{ij} > r \end{cases}$.

For a given distance $r \geq 0$, equation (4.5) first counts the number of (distinct) pairwise distances in the observed point pattern that are less than r, i.e. all pairs of points x_i and x_j $i \neq j$ such that the distance

d_{ij} is less than, or equal to r. This sum is then divided by $n(n-1)$ to get the proportion of d_{ij} values less than r. Note that $n(n-1)$ is the total number of pairs of distinct points.

To aid in the derivation, define $n_i(r)$ as the number of points within distance r of point x_i. Then we have that $n_i(r) = \sum_{j \neq i} I(d_{ij} \leq r)$. For a simple illustration we can study Figure 4.1 where we limit our attention to an arbitrary point B. All the circles drawn are centered at this point and if we let the first circle have radius r_1, then $n_B(r_1) = 5$. This counting is done for all the other points in the point pattern for the distance r_1. We then increase the distance to r_2 and do the same calculation. The average number

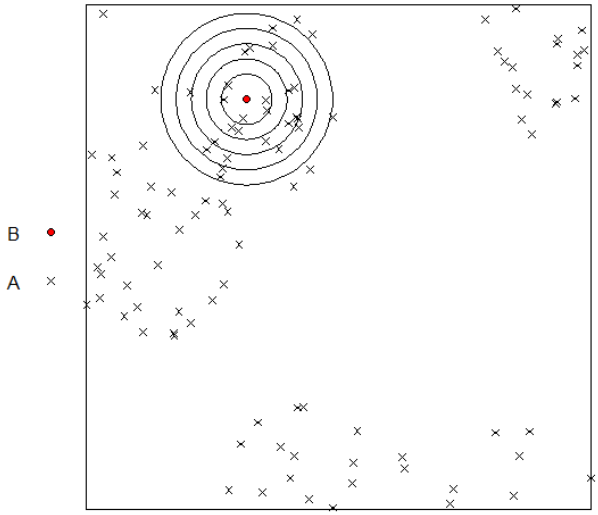


Figure 4.1: Illustration of increasing the radius r.

of points within distance r of a point is then given by $\bar{n}(r) = \frac{1}{n}\sum_{i=1}^{n} n_i(r)$. This quantity is close to what we would require of an estimate of $K(r)$ since the definition of the K-function is concerned with the 'average' number of additional points within distance r of an arbitrary point. We can then re-write equation (4.5) using this new definition as follows

$$\hat{P}(r) = \frac{1}{n(n-1)} \sum_{i=1}^{n} n_i(r) = \frac{n}{n(n-1)} \bar{n}(r) = \frac{1}{n-1} \bar{n}(r). \tag{4.6}$$

It is now important to standardise this value somehow because this average $\bar{n}(r)$ is dependent on the number of points in the dataset. In order to enable comparison between datasets with different numbers of points we standardise this average by dividing by an estimate of the intensity as given in [5] as $\tilde{\lambda} = (n-1)/|D|$. Other estimates use the usual maximum likelihood estimate of λ but Diggle argues in [18] that this is not an important issue when n is large. So we shall also stick to $\tilde{\lambda}$ in our discussions.

Having said the above, we then standardise (4.6) as

$$\bar{n}(r) / \left(\frac{n-1}{|D|} \right) = |D| \hat{P}(r).$$

A tentative estimate for the K-function is then given by

$$\tilde{K}(r) = \frac{|D|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i} I(d_{ij} \leq r).$$

Thus we can see that the estimate above asserts that the K-function is related to pairwise distances and an average of some form [18]. The estimate above however is a biased estimator of the K-function. This is due to what is called edge-effects.

4.2.3 Edge effects

We now briefly discuss edge effects using the ideas laid out in [14]. Suppose we have a similar setting to the one laid out at the start of the section, i.e. we have a point pattern $\mathbf{x}$ on $D \subset \mathbb{R}^2$ with location of events (x_{i1}, x_{i2}). Suppose, as is the case in our present discussion, that we are concerned with inter-point distances. The window of observation D is usually just a subset of a much larger area where the process is observed. The consequence of this fact is clear if we consider the distance to the nearest event for some analysis. The points near the boundary of the window D will have larger expected nearest-neighbour distances than points that are more to the center of D. Therefore estimators that use this distance will be biased. Taking this discussion further, we consider the situation where points near the boundary interact with points just outside the boundary [18]. Since we are confined to the window D, the points outside the window are essentially unobserved. Any information that we could have had concerning these interactions is lost.

Edge effects do not only show up when considering pairwise distances. We briefly mentioned them in Section 2.2.1 when discussing the intensity function, in particular when discussing kernel estimators of the intensity function. In [5] it is argued that for the uncorrected estimate of the intensity in Table 2.1, points u closer to the boundary of D would contribute less to the sum. That is, a negative bias is associated with points u closer to the boundary because of edge effects.

We now consider an example of an edge correction used in estimating the K-function called the isotropic edge correction. Let $\mathbf{X}$ be a point process on $\mathbb{R}^2$. Let x_i and x_j be points belonging to $\mathbf{X}$, and let x_i fall in the observational window D. As before $d_{ij} = ||x_i - x_j||$ is the Euclidean distance between the two points. Now, the point x_j must lie at some location on the disc $B(x_i, d_{ij})$ of radius d_{ij} centered at x_i. If we further assume that $\mathbf{X}$ is isotropic, then x_j can lie anywhere on the disc with uniform probability. Consider then the following quantity

$$p(x_i, d_{ij}) = \frac{l(D \cap \partial[B(x_i, d_{ij})])}{2\pi d_{ij}} \tag{4.7}$$

where $\partial[B(x_i, d_{ij})]$ is the boundary of the disc $B(x_i, d_{ij})$. The numerator is the length (l) of the boundary of the disc falling in the window D. Therefore, the quantity in (4.7) is the probability that x_j falls in D,

which is just the proportion of the circumference of the disc lying in D. (Recall that d_{ij} is the radius of the disc, therefore the length of the circumference is rightly $2\pi d_{ij}$.) We then note that if d_{ij} increases, i.e. we consider a point x_j that is further from x_i, this probability decreases. This makes intuitive sense since as the disc increases, less of it will lie on the window. Thus, as pointed out in [5], we are less likely to observe large distances. Also note that for points that are pairs of points that are more in the centre of the window, the whole disc of radius d_{ij} for example will lie entirely inside the window so that $p(x_i, d_{ij}) = 1$.

One way of correcting for edge effects then is to introduce as weight of some form. The form of this weight is the reciprocal of $p(x_i, d_{ij})$. Each pair of points x_i and x_j is weighted by $p(x_i, d_{ij})^{-1}$ [5]. Once again note that for points towards the center of the window the edge correction factor will be 1 meaning that in effect no edge correction is applied. This then gives the K-function with the isotropic correction [5]

$$\hat{K}_{iso}(r) = \frac{|D|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i} I(d_{ij} \leq r) \frac{1}{p(x_i, d_{ij})}. \tag{4.8}$$

For large values of d, it is possible that $p(x, d) = 0$, i.e. the disc lies entirely outside the window D. Define R_{min} to be the smallest distance such that there is a location $a \in D$ where $p(a, d) = 0$. More clearly this location, which should not be necessarily a data point, is such that the circle of radius d centered at this point lies entirely outside the window of observation D i.e. $l(D \cap \partial[B(a, d)]) = 0$. Then for $r < R_{min}$ the estimator in equation (4.8) is an unbiased estimate of the K-function [5].

If we would like to use the estimator for larger distances, a bit more modification is required. Define $D^*(r) \subset D$ to be the region that contains all locations a such that $p(a, r) > 0$. That is, $D^*(r)$ contains all locations a such that the circle centered at location a with radius r lies, at least partially, inside the window. Let R^*_{min} be the smallest distance such that $|D^*(r)| = 0$. That is, R^*_{min} is the smallest distance such that the area of the region containing locations a such that $p(a, d) > 0$ is 0. Put differently, R^*_{min} is the smallest distance such that there are no locations a where $p(a, d) > 0$. The final estimator is then given in [5] by

$$\hat{K}(r) = \frac{|D|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i} I(d_{ij} \leq r) \frac{|D|}{p(x_i, d_{ij})|D^*(d_{ij})|} \tag{4.9}$$

where $|D^*(d_{ij})|/|D|$ is the proportion of the window area occupied by $D^*(d_{ij})$. Equation (4.9) is an unbiased estimator of the K-function for $r < R^*_{min}$ [18]. The newly defined terms will be replaced by $e_{ij}(r)$ to make the notation less cumbersome

$$\hat{K}(r) = \frac{|D|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i} I(d_{ij} \leq r) e_{ij}(r). \tag{4.10}$$

A final point on edge corrections is that according to [5] the choice of edge correction is not that all important. All that is important is that at least one is applied. Furthermore, if we get different results when

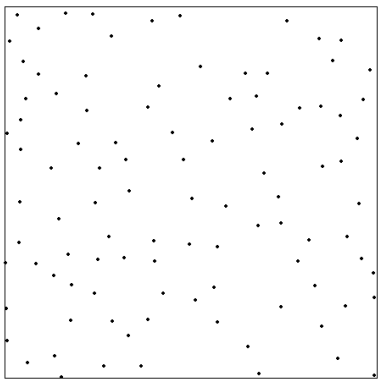
we use different edge corrections, it is possible that the point pattern is not homogeneous, which was one of the assumptions required when defining the K-function and deriving its estimate. Therefore, as previously mentioned, it is important that one determines whether or not the point pattern has homogeneous intensity before using the K-function.

4.2.4 Simulations

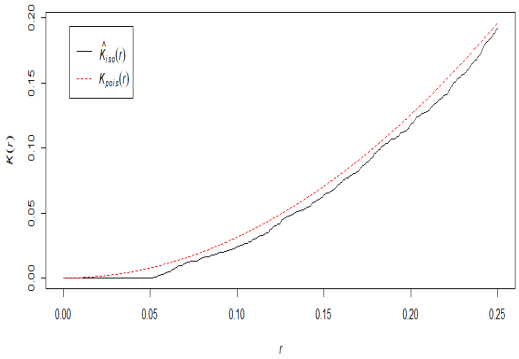
As alluded to earlier, the K-function can be used, at least in part, to distinguish between clustered, regular and random point patterns. The 'null hypothesis' is taken to be that the point pattern is completely random. The homogeneous Poisson process is used as a benchmark against which to compare whatever estimate of a summary function we use. The K-function for the homogeneous Poisson process can be easily determined. Recall $\bar{n}(r)$ which was defined as the average number of points within distance r of a typical point. Using the CSR properties given in Section 2.3, particularly property 2, the average number of points falling in a disc of radius r is $\lambda \times \pi r^2$ since πr^2 is the area of the disc. Thus after standardising, the K-function for a completely spatial random pattern is $K(r) = \pi r^2$.

Through simulation we now see how the empirical estimate for K-function for different point configurations compares with the one for CSR. We begin with a regular point pattern. Recall that for regular point patterns, the interpoint distances are larger than what we expect under CSR. For regular point patterns the simple sequential inhibition mechanism was used which was described in Section 2.4, and realisations were generated using the `rSSI` function in `spatstat`. The `Kest` function, also in the `spatstat` package, was used to calculate an estimate of the K-function for the simulated datasets. Figure 4.2 shows the simulated regular point pattern and the corresponding K-function estimate. Since the interpoint distances are larger in the regular patterns than what we expect under CSR, this implies that there are a few number of points within a certain distance r of an arbitrary point. This is reflected in the graphs as the K-function estimates for both graphs, denoted by $\hat{K}_{iso}(r)$, lie below the expected curve under complete spatial randomness, denoted by $K_{pois}(r)$.

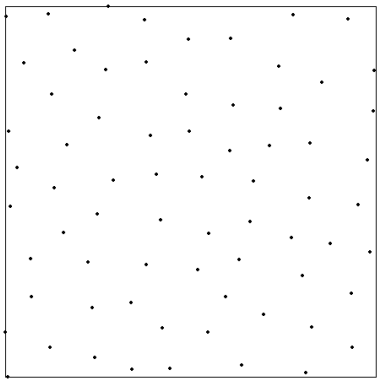
We now consider a clustered point pattern. Clustered point patterns were generated using Matérn cluster process described in Section 2.4. For clustered point patterns, the inter-point distances are smaller than expected. This is clear because 'offspring' points are uniformly distributed within a disc of radius r and so the distances between points in the same cluster will be smaller than under CSR. This then translates to the number of points within a certain distance of an arbitrary point being larger than expected. As can be seen in Figure 4.3, the empirical estimate lies above the graph for what is expected under CSR.



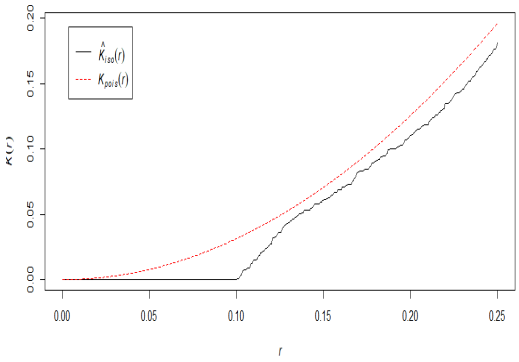
(a) SSI realisation with $r = 0.05$



(b) K-function estimate



(c) SSI realisation with $r = 0.10$



(d) K-function estimate

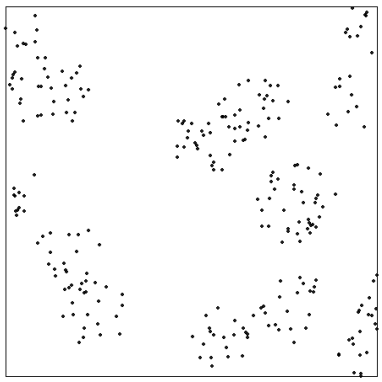
Figure 4.2: Regular point patterns with corresponding K-function estimates, generated using `rSSI` and `Kest` respectively on a unit square window.

4.2.5 Confidence intervals

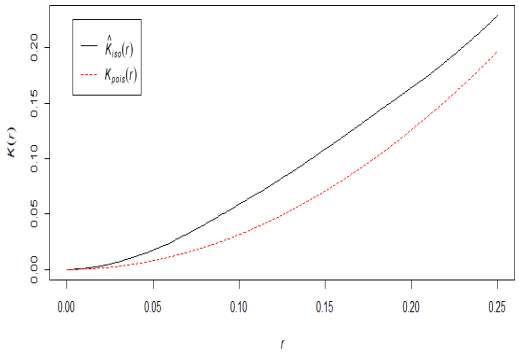
Once we have an estimate of any sort it is good statistical practise to determine the precision of the estimate through calculating the variance of the estimate. There are two methods that will be discussed in this section namely, the block bootstrap method and Loh's bootstrap [5, 18].

4.2.5.1 Block bootstrap

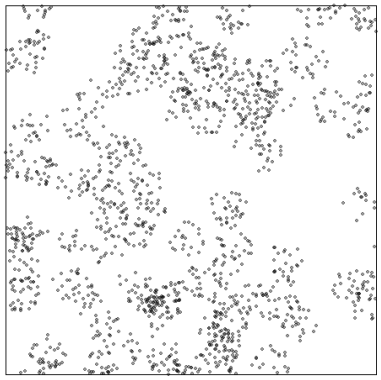
The first method of variance estimation is the block bootstrap method from in [18, 5]. One assumption for this method is that the window D is rectangular. We divide this window into Q equal quadrats (as we did for quadrat counts), say $D_1, D_2, ..., D_Q$. We then focus on a quadrat D_l and a point x_i found in D_l and another point x_j that may lie anywhere on the window of observation. For this quadrat D_l consider the following quantity



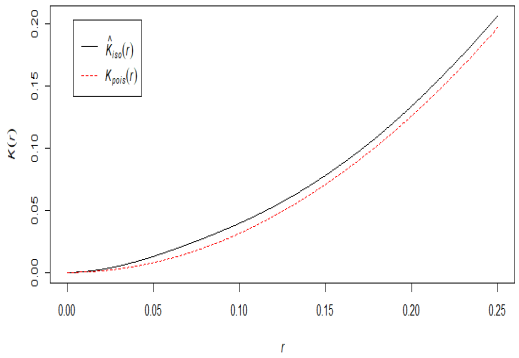
(a) `rMatClust`: $\kappa = 15, r = 0.10, \mu = 20$



(b) K-function estimate



(c) `rMatClust`: $\kappa = 100, r = 0.05, \mu = 20$



(d) K-function estimate

Figure 4.3: Clustered point patterns with corresponding K-function estimates, generated using `rMatClust` and `Kest` respectively on a unit square window.

$$\hat{K}(r, D_l) = \frac{Q|D|}{n(n-1)} \sum_{x_i \in D_l} \sum_{j \neq i} I(d_{ij} \leq r) e_{ij}(r)$$

which is the estimate of the K-function for the quadrat D_l, such that $x_i \in D_l$ and x_j is found anywhere (possibly even in D_l). For a fixed value of r we then have Q of these K-functions for the quadrats. Note that to get the estimate from equation (4.10) we average over the quadrats. That is

$$\hat{K}(r) = \frac{1}{Q} \sum_{l=1}^{Q} \hat{K}(r, D_l).$$

An estimate for the variance is calculated as follows

$$\widehat{\text{var}}[\hat{K}(r, Q)] = \frac{1}{m-1} \sum_{l=1}^{m} \left[\hat{K}(r, D_l) - \hat{K}(r)\right]^2.$$

We now make a couple of assumptions: the Q $\hat{K}(r)$ estimates are independent and identically distributed. Assuming that the estimates are independent implies a further implicit assumption that the quadrat counts are independent as well. As Diggle points out in [18], the independence assumption holds for the

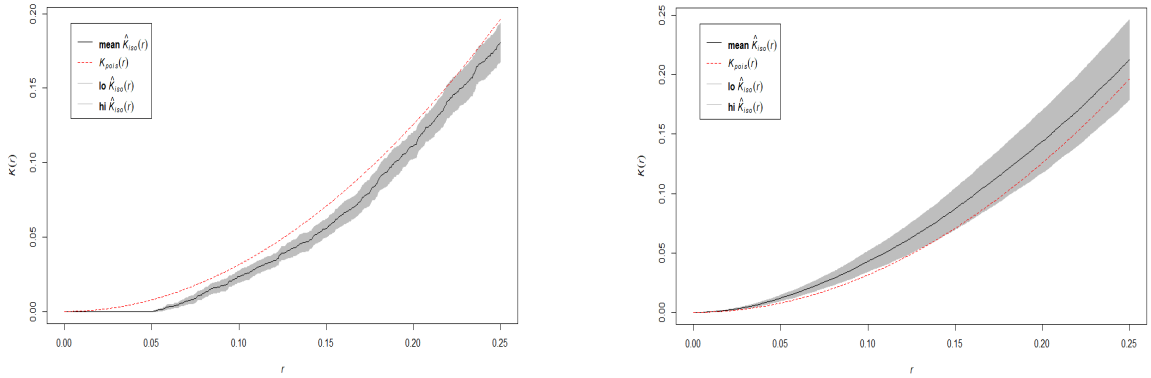
homogeneous Poisson process but not generally. Care must be taken therefore when this estimate for the variance is used. An estimate for the variance of $\hat{K}(r)$ is obtained as follows

$$\begin{aligned}
\widehat{\text{var}}(\hat{K}(r)) &= \widehat{\text{var}}\left[\frac{1}{Q}\sum_{l=1}^{Q}\hat{K}(r,B_l)\right] \\
&= \frac{1}{Q^2}\sum_{l=1}^{m}\widehat{\text{var}}\left[\hat{K}(r,B_l)\right] \text{ since independence assumed} \\
&= \frac{Q}{Q^2}\widehat{\text{var}}\left[\hat{K}(r,D)\right] \text{ since identical distribution assumed} \\
&= \frac{1}{Q}\widehat{\text{var}}\left[\hat{K}(r,D)\right].
\end{aligned}$$

The standard error for the estimate is given by $\sqrt{\frac{\widehat{\text{var}}[\hat{K}(r,D)]}{Q}}$. If we assume normality, an approximate $100(1-\alpha)\%$ confidence interval for $K(r)$ is given by

$$\hat{K}(r) \pm z_{\alpha/2}\sqrt{\frac{\widehat{\text{var}}[\hat{K}(r,D)]}{Q}}. \tag{4.11}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)^{th}$ percentile of the standard normal distribution. As Baddeley *et al.* point out in [5], the confidence intervals are calculated for each fixed value of r and thus hold just for that single value of r. They are therefore called pointwise confidence intervals of the estimate, because they hold for one value of r at a time. Hence the confidence that the true K-function lies within the confidence intervals, for all values of r for which it is calculated, is less than 95% if $\alpha = 0.05$ [5]. Figure 4.4 shows K-functions for two point configurations, with estimates of the confidence intervals calculated using the `varblock` function. The number of blocks Q was 25.



(a)K-function estimate for a regular point pattern (b)K-function estimate for a clustered point pattern

Figure 4.4: K-function estimates with 95% confidence intervals. These were calculated using the `varblock` function in `spatstat`. **lo** $\hat{K}_{iso}$ and **hi** $\hat{K}_{iso}$ are the lower and upper confidence limits respectively. **mean** $\hat{K}_{iso}$ is the estimate of the K-function.

4.2.5.2 Loh's bootstrap

Loh's bootstrap is considered in [5] and [28]. In this method we consider what Baddeley *et al.* call local K-functions where equation (4.10) is broken down such that we have n K-functions, one for each point x_i, for each value of r. The local K-function for a point x_i is given by

$$\hat{K}(r, x_i) = \frac{|D|}{n-1} \sum_{j \neq i} I(d_{ij} \leq r) e_{ij}(r).$$

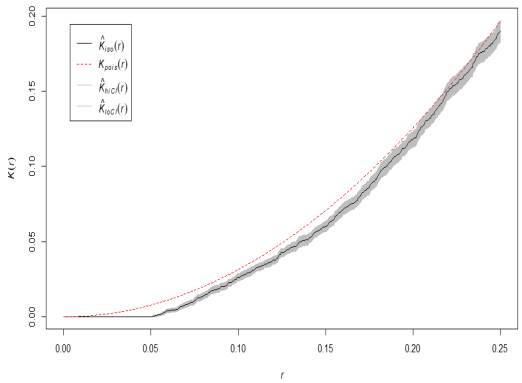
To obtain the original estimate given in equation (4.10) we find the average of the n local K-functions as

$$\hat{K}(r) = \frac{1}{n} \sum_{i=1}^{n} \hat{K}(r, x_i).$$

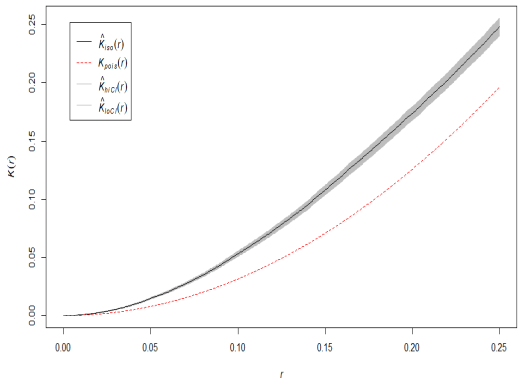
We then do the usual bootstrapping procedure done in classical statistics. We take an independent random sample of the points size n with replacement. Denote this sample of size n as $x'_1, x'_2, ..., x'_n$. We then calculate a resampled version of the K-function as follows

$$K^*(r) = \frac{1}{n} \sum_{l=1}^{n} \hat{K}(r, x'_l).$$

This process is repeated a large number times T, which is usually 999 in bootstrap approaches. This gives a bootstrap sample distribution of the estimate $\hat{K}(r)$. The confidence intervals can then be simply calculated by determining the relevant percentiles based on the level of significance chosen. That is, for $\alpha = 0.05$, we need to find the 2.5^{th} and 97.5^{th} percentiles to construct 95% confidence intervals. Figure 4.5 shows the confidence intervals calculated using this method for the same point patterns as in Figure 4.4. The main difference, particularly with the clustered point pattern, is that the confidence intervals are narrower. For the rest of our discussions we will focus on Loh's bootstrap, particularly because we do not have to concern ourselves with division of the window into a certain number of blocks.



(a)K-function estimate for a regular point pattern



(b)K-function estimate for a clustered point pattern

Figure 4.5: K-function estimates with 95% confidence intervals. These were calculated using the `lohboot` function in `spatstat`. $\hat{K}_{loCl}$ and $\hat{K}_{hiCl}$ are the lower and upper confidence limits respectively. $\hat{K}_{iso}$ is the estimate of the K-function. The number of resamples was 999.

4.3 The cross-K function

4.3.1 Theoretical definition

In Section 4.2 the case where the point pattern had no marks, i.e. an unmarked point pattern, was discussed. That is, we were only interested in one type of points and their interdependence. In this section we discuss point patterns where in addition to locations, the points have a corresponding mark as discussed in Section 2.5.

The K-function has an extension called the cross-K or cross-type K function in literature. As mentioned in [29], in analysis of multitype points, we limit the discussion to two types of points at a time. To set the scene here we assume that the point process has two marks so that we have points of type i with intensity λ_i and points of type j with intensity λ_j. The theoretical definition of the cross-K function is then given by [5, 20, 29]

$$K_{ij}(r) = \frac{1}{\lambda_j} E[\text{number of type } j \text{ points within distance } r \text{ of a random type } i \text{ point}]. \tag{4.12}$$

The cross-K function defined in equation (4.12), analogously to the K-function, measures the spatial dependency between points of type i and type j [5]. We note then that if $i = j$, then we have $K_{ii}(r)$ which is just the K-function defined in Section 4.2. The distinction between the two functions is clear when $i \neq j$ because in that case, as we shall see, we only concern ourselves with distances between the points if they are of a different type. That is to say we are not concerned with the distances between points of the same type. We want to know how points of one type relate to the other type.

4.3.2 Estimation of the cross-K function

The derivation of the estimate for the cross-K function is similar to the way the ordinary K-function was derived, the major difference being notation. We assume we have a point pattern with type i points and type j points, as has been the assumption in this section. Let x_{i_k} be the location of the k^{th} type i event and x_{j_l} be the location of the l^{th} type j event. Let n be the total number of type i points such that $\{x_{i_1}, x_{i_2}, ..., x_{i_n}\}$ is the set of all the locations of type i points. Similarly let $\{x_{j_1}, x_{j_2}, ..., x_{j_m}\}$ be the set of locations of all m of type j points. Lastly denote the distance between the k^{th} type i event and the l^{th} type j event as d_{i_k,j_l}, i.e. $d_{i_k,j_l} = ||x_{i_k} - x_{j_l}||$.

Once again define the following quantity

$$\hat{P}_{ij}(r) = \frac{1}{n \times m} \sum_{k=1}^{n} \sum_{l=1}^{m} I(d_{i_k,j_l} \leq r) \tag{4.13}$$

which, for a given distance r, is the proportion of d_{i_k,j_l}'s less than r. For a given distance r we count the number of type j points that are within distance r of type i points. To break down equation (4.13) we define $n_{i_k}(r)$ to be the number of type j points within distance r of the k^{th} type i point found at location k, i.e. $n_{i_k}(r) = \sum_{l=1}^{m} I(d_{i_k,j_l} \leq r)$. The average number of points of type j within distance r of a point of type i is then $\bar{n}_i(r) = \frac{1}{n} \sum_{k=1}^{n} n_{i_k}(r)$. Thus the proportion in equation (4.13) can be re-written as

$$\hat{P}_{ij}(r) = \frac{1}{nm} \sum_{k=1}^{n} n_{i_k} = \frac{1}{m} \bar{n}_i(r).$$

From the above it is clear that $m\hat{P}_{ij}(r)$ is the average number of type j points within distance r of a type i point. All that is left now is to standardise so that, as argued in Section 4.2.2, we can compare datasets with varying number of points. The standardisation in this context is done by dividing the average by an estimate of the intensity of type j points i.e. $\hat{\lambda}_j = m/|D|$:

$$\frac{\left[m \times \hat{P}_{ij}(r)\right]}{m/|D|} = |D|\hat{P}_{ij}(r) = \frac{|D|}{n \times m} \sum_{k=1}^{n} \sum_{l=1}^{m} I(d_{i_k,j_l} \leq r).$$

Therefore, the stadardised average number of points of type j within distance r of a typical point of type i, corrected for edge effects as discussed in Section 4.2.2 is given by

$$\hat{K}_{ij}(r) = \frac{|D|}{nm} \sum_{k=1}^{n} \sum_{l=1}^{m} I(d_{i_k,j_l} \leq r) e_{i_k,j_l}(r) \tag{4.14}$$

The confidence intervals using Loh's bootstrap are calculated in a similar way to the K-function. The local cross-K function for a point of type i at location k as follows is first calculated as follows

$$\hat{K}_{ij}(r, x_{i_k}) = \frac{|D|}{m} \sum_{l=1}^{m} I(d_{i_k,j_l} \leq r) e_{i_k,j_l}(r). \tag{4.15}$$

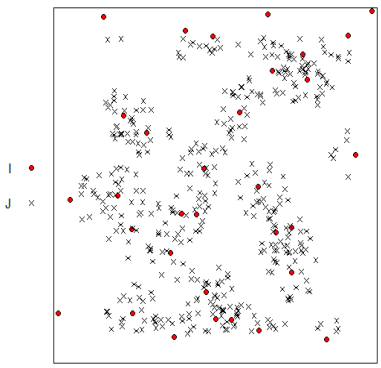
It is clear that $\hat{K}_{ij}(r) = \frac{1}{n_i} \sum_{k=1}^{n_i} \hat{K}_{ij}(r, x_{i_k})$. We resample from the n local cross-K functions by taking a simple random sample of size n with replacement. Suppose our sample gives us $\hat{K}_{ij}(r, x_{i_k})$'s associated

with the points $x'_{i_1}, x'_{i_2}, ..., x'_{i_n}$. A resampled version of $\hat{K}_{ij}(r)$ is calculated as

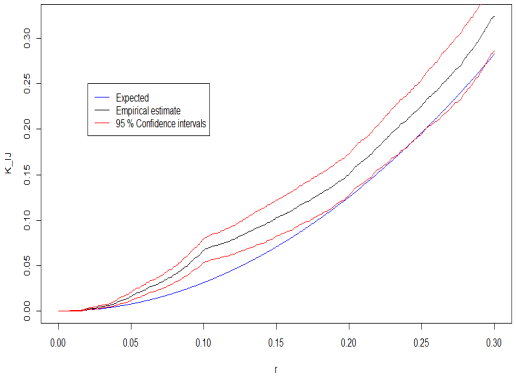
$$K^*_{ij}(r) = \frac{1}{n}\sum_{k=1}^{n} \hat{K}_{ij}(r, x'_{i_k}). \tag{4.16}$$

We do this a large number of times T, usually 999, which then gives us a bootstrap distribution of $\hat{K}_{ij}(r)$ from which we can find $100(1-\alpha)\%$ confidence intervals.

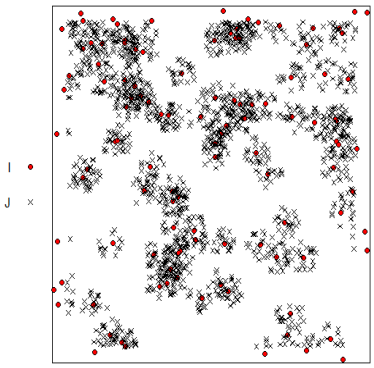
Figure 4.6 shows two realisations using the `rMatClust` function and the corresponding cross-K function estimates. To get a marked point pattern as seen in the figures, the argument `saveparents` of the `rMatClust` function was set to `True` and so the parent locations were stored as an attribute of the resultant point pattern. This was then extracted and the point pattern of the 'parent' points, labelled I was superimposed on the offspring point pattern, labelled J. For Figure 4.6(a) there were 32 type I points and 437 type J points, and for (c) there were 113 type I points and 1854 type J points. At the writing of this document there was no function to calculate the confidence intervals using Loh's bootstrap for the cross-K function, so this was coded using `R` and the steps outlined above.



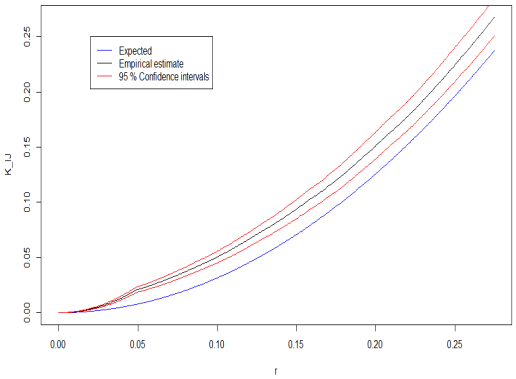
(a) `rMatClust`: $\kappa = 20, r = 0.10, \mu = 20$



(b) Cross-K function estimate



(c) `rMatClust`: $\kappa = 100, r = 0.05, \mu = 20$



(d) Cross-K function estimate

Figure 4.6: Clustered point patterns with corresponding K-function estimates, generated using `rMatClust` and `Kcross` respectively. Displayed in red are the 95% confidence intervals calculated using Loh's bootstrap with 999 resamples.

4.4 Points and linear structures

4.4.1 Introduction

In this section we bring together the ideas discussed in Sections 4.2 and 4.3 to attempt to characterise the spatial relationship between points and lines. The extension of the functions defined in the previous sections, the empirical estimate derived and confidence intervals discussed as well. This extension was proposed by Comas *et al.* in [13] and the definitions and arguments used here follow from there.

4.4.2 Definitions and estimation

We start by defining a line segment and a linear network in similar fashion to Comas *et al.* [13]. A line segment with endpoints a and b is defined as $[a, b] = \{ta + (1 - t)b : 0 \leq t \leq 1\}$. A linear network is defined as the union of the lines segments i.e. $L = \bigcup_{i=1}^{n_l} l_i$. A simple linear network from the `spatstat` package is shown in Figure 4.7. Let $\boldsymbol{X}$ be a spatial point process that generates a set of countable events $x_1, x_2, ..., x_n$ on a plane, say, D. For our purposes we shall assume that the point pattern generated is stationary and hence has a constant intensity λ. Let $|L|$ and $|D|$ be the length of the linear network and the area of the observational window, respectively.

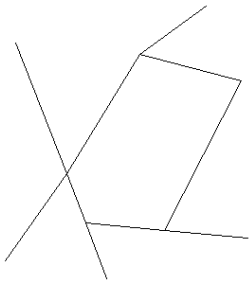


Figure 4.7: A simple linear network from the `spatstat` package.

The theoretical function for detecting spatial dependency between linear network and point pattern is defined in [13] as follows:

$$K_{LX}(r) = \frac{1}{\lambda_L} E\left[\int_L [I(0 < ||\boldsymbol{x} - y|| \leq r) | \boldsymbol{x} \in \boldsymbol{X}] dy \right] \tag{4.17}$$

where $I(0 < ||\boldsymbol{x} - y|| \leq r) = \begin{cases} 1 & \text{if } ||\boldsymbol{x} - y|| \leq r \\ 0 & \text{if } ||\boldsymbol{x} - y|| > r \end{cases}$ and $||\boldsymbol{x} - y||$ calculates the Euclidean distance between a point of the point pattern $\boldsymbol{x}$ and a point y lying on the linear network L. Comas *et al.* then mention that $\lambda_L K_{LX}(r)$ is the expected length of L falling in a disk $b(x, r)$ [13]. Also note that we also assume that the linear network is stationary, so $\lambda_L = |L|/|D|$.

It makes intuitive sense that if there is some spatial dependency between the point pattern and the linear network the length of L falling in the disks centered at the points of $\boldsymbol{X}$ would be larger than expected. Consider a simple example shown in Figure 4.8. Here the linear network is a single straight line and from a visual study we have an initial feel that the points are drawn towards the line.

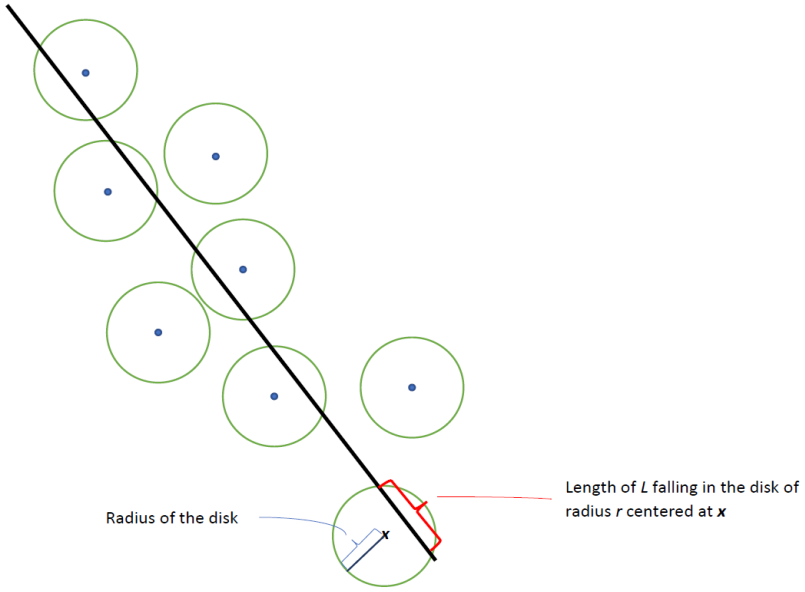


Figure 4.8: A simple linear network and a point pattern clustered around it.

For completeness we consider Figure (4.9) where the points seem to be repelled by the linear network.

As in traditional statistics, we have to derive an estimate for (4.17) above. The first step perhaps would be to determine an estimate of the integral term. One way of doing this would be through a Riemann sum. We divide the linear network L into n_L roughly equal partitions as done in [13]. Note then that $|L|/n_L$ is the length of each partition. An estimate of the integral is then

$$\hat{I}(r) = \sum_{i=1}^{n}\sum_{j=1}^{n_L} I(0 < ||x_i - y_j|| \leq r)\frac{|L|}{n_L} \tag{4.18}$$

where y_j is the middle of the j^{th} line segment. Since $\lambda_L K_{LX}(r)$ is the expected length of L falling in a disk $B(x, r)$, the average of (4.18) above is then

$$\bar{\hat{I}}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_L} I(0 < ||x_i - y_j|| \leq r)\frac{|L|}{n_L} \tag{4.19}$$

Finally we standardise (4.19), as we did in Sections 4.2 and 4.3, by dividing by the intensity of the linear

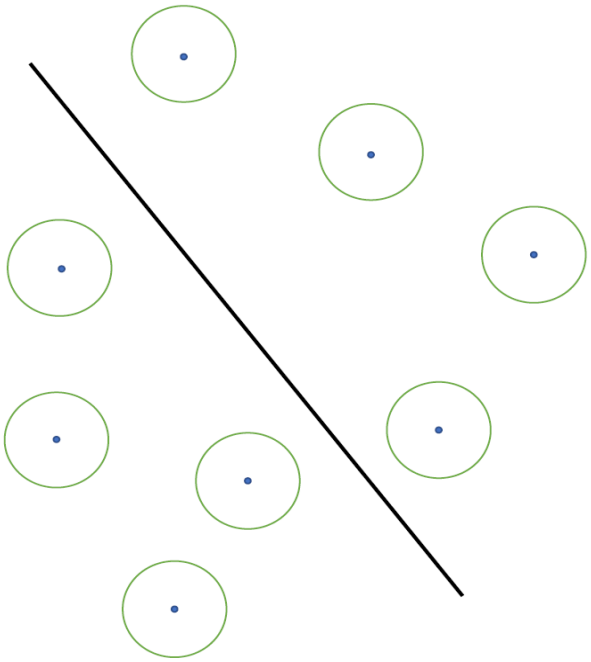


Figure 4.9: A simple linear network and a point pattern being repelled by the linear network.

network $\lambda_L = |L|/|D|$

$$\frac{\bar{\hat{I}}(r)}{\lambda_L} = \frac{|D|}{|L|}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_L} I(0 < ||x_i - y_j|| \leq r)\frac{|L|}{n_L}$$

To arrive to the final form of the empirical estimate we add a term to take into account edge effects, and so the empirical estimate of (4.17) is

$$\hat{K}_{LX}(r) = \frac{|D|}{n \times n_L}\sum_{i=1}^{n}\sum_{j=1}^{n_L} {w_{ij}}^{-1} I(0 < ||x_i - y_j|| \leq r) \tag{4.20}$$

where ${w_{ij}}^{-1}$ is an edge correction whose importance was discussed in Section 4.2.2.

We now investigate if there are any differences if we define (4.17) with the linear network as the point of reference. Define this alternative as follows

$$K_{XL}(r) = \frac{1}{\lambda}E[N_{XL}(r)] \tag{4.21}$$

where $N_{XL}(r)$ is the number of points of the point pattern $\boldsymbol{X}$ within distance r of the linear network L. Once again, if there is a tendency of points to cluster around a linear network then there will be more points closer to the linear network than expected. A visual representation of how this would be different from the first case is shown in Figure 4.10.

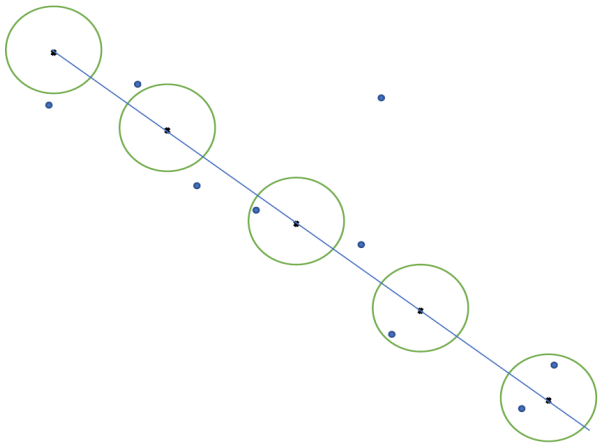


Figure 4.10: A simple linear network and a point pattern clustered around it.

Note that we picked a few points on the linear network to illustrate this second case. If we would like a good estimate of (4.21) then we would have to take more points from the linear network. So, the main question is which points on the linear network will be used as a point of reference. We shall keep the same method as in the first part, that is, the linear network will be divided into n_L roughly equal partitions. The middle points of the resulting line segments will be our points of reference.

Let d_{ji} be the Euclidean distance between the middle of the j^{th} line segment of L and a point x_i of the point pattern. Calculate all the pairwise distances between the mid-points of the line segments of L and points x_i and the point pattern (there are $n \times n_L$ of these in total). For each r determine the number of d_{ji} values less than r. The total number of these values less than r is $\sum_{j=1}^{n_L}\sum_{i=1}^{n} I(0 < ||y_j - x_i|| \leq r)$. Put differently, the above counts the number of times a point x_i is within distance r of a point y_j of the linear network. Now, the proportion of d_{ji} values less than r is given by

$$P_{ji}(r) = \frac{1}{n \times n_L}\sum_{j=1}^{n_L}\sum_{i=1}^{n} I(d_{ji} \leq r).$$

Let $c_j(r)$ be the total number of points of the point pattern within distance r of the j^{th} middle point i.e.,

$$c_j(r) = \sum_{i=1}^{n} I(d_{ji} \leq r).$$

The average number of points of $\boldsymbol{X}$ within distance r of a 'point' of the linear network is $\bar{c}(r) = \frac{1}{n_L}\sum_{j=1}^{n_L} c_j(r)$. Then

$$P_{ji}(r) = \frac{1}{n_L \times n}\sum_{j=1}^{n_L} c_j(r) = \frac{n_L}{n \times n_L}\bar{c}(r) = \frac{1}{n}\bar{c}(r).$$

So then $n \times P_{ji}(r) = \bar{c}(r)$ is the average number of points of the point pattern $\boldsymbol{X}$ within distance r of

a 'point' of the linear network. Standardise this by dividing by $\hat{\lambda} = n/|D|$ which is an estimate of the intensity of the point pattern

$$\frac{n \times P_{ij}(r)}{\frac{n}{|D|}} = \frac{\bar{c}(r)}{\frac{n}{|D|}} = \frac{|D|}{n}\bar{c}(r) = \frac{|D|}{n}\frac{1}{n_L}\sum_{j=1}^{n_L}\sum_{i=1}^{n} I(d_{ji} \leq r).$$

The estimate of (4.21) along with the edge corrections is then given by

$$\hat{K}_{XL}(r) = \frac{|D|}{nn_L}\sum_{j=1}^{n_L}\sum_{i=1}^{n} {w_{ji}}^{-1} I(0 < ||y_j - x_i|| \leq r)$$

which is similar estimate we obtained in (4.20) but not exactly the same because of the edge correction. When edge corrections are used $\hat{K}_{XL}(r)$ and $\hat{K}_{LX}(r)$ are not equal but are positively correlated [20].

4.4.3 Confidence intervals: Loh's bootstrap

Confidence intervals were initially discussed in Section 4.2.5. For the estimate given by equation (4.20) we will consider confidence intervals using Loh's bootstrap extended from that given in Section 4.2.5. The block bootstrap method is less than ideal because of the nature of the problem at hand. That is, we are working with linear networks so it would not make much sense to sample some blocks with roads because of the way the linear network is connected. We then only consider Loh's bootstrap because this only involves sampling the points only and not the roads which makes sense because the road can be considered to be a permanent structure, while the points are generated by some stochastic mechanism.

Therefore, as an extension to Loh's bootstrap introduced earlier, the way the confidence intervals are calculated is as follows:

1. Define and calculate the local linear network K-function:

$$\hat{K}_{LX_i}(r) = \frac{|A|}{n_L}\sum_{j=1}^{n_L} I(||x_i - y_j|| \leq r){w_{ij}}^{-1}.$$

   which calculates the empirical function for a single point $x_i \in \boldsymbol{X}$.

2. Resample from the n local linear network cross K-functions by taking a simple random sample of size n with replacement and calculate a new resampled version of the function: $K^*_{LX}(r) = \frac{1}{n}\sum_{i=1}^{n}\hat{K}_{LX_i}(r)$.

3. Do this 999 times to obtain a bootstrap distribution of $\hat{K}_{LX}(r)$ and determine confidence intervals by finding relevant percentiles.

The way the confidence intervals are calculated perhaps gives more justification to why the function was defined as equation (4.17) and not (4.21). If (4.21) was used, the calculation of confidence intervals would

have involved sampling points on the linear network in step 2 above. This does not make much sense because the linear network is usually considered a permanent structure. Using the bootstrap is valid if you have a sample that you believe is representative of the whole population. The multiple resamples are an attempt to mimic sampling variability. The sampling of the linear network is therefore incorrect because we have it in its entirety because of its permanence.

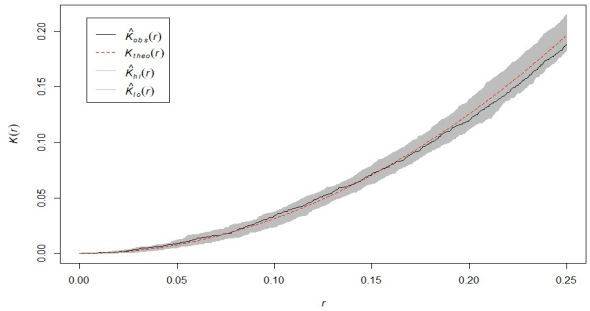
4.5 Hypothesis Tests and Simulation Envelopes

In addition to having confidence intervals for the empirical estimate of the summary function, we could also consider simulation envelopes which are also a graphical method developed using a Monte Carlo approach. Monte Carlo approaches to spatial point patterns have been widely used in writings such as in [4, 5, 19, 24]. These methods enable us to perform valid hypothesis tests and draw conclusions based on our data.

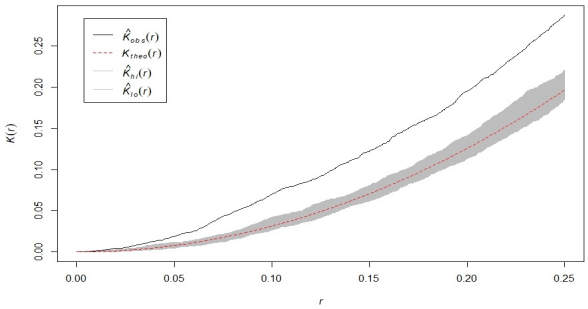
4.5.1 Simulation envelopes

Suppose the question at hand is if a point pattern we have is completely random. The procedure, as described in [5], is to simulate N realisations under the null hypothesis, that of CSR, with the same number of points as the original point pattern and on the same window. The idea is that if the null hypothesis of independence were true, the original dataset would be statistically equivalent to the N datasets simulated under CSR. Therefore the observer would not be able to pick the original dataset if it was laid out together with the N simulated ones. For each of these realisations, calculate the appropriate summary function, such as the K-function, then determine the maximum and minimum values of these summary functions. These values are then the upper and lower simulation envelopes under CSR. Loosely speaking, if the estimated summary function lies between these envelopes one could say that it is not significantly different from CSR (we shall discuss the care that should be taken when dealing with simulation envelopes shortly).

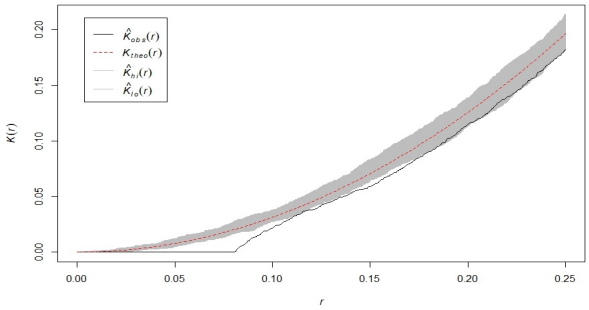
Examples of simulation envelopes are shown in Figure 4.11. Each point pattern had around 100 points and N was taken to be 99. It is clear that for Figure 4.11(a) the estimate of the K-function, $\hat{K}_{obs}(r)$, lies within the simulation envelopes, as expected because the original point pattern was a homogeneous Poisson point pattern. Figure 4.11(b) shows that $\hat{K}_{obs}(r)$ lies outside the simulation envelopes, simulated assuming the null hypothesis of CSR. Again this is not surprising because the point pattern generated was a clustered point pattern. Lastly Figure 4.11(c) shows that $\hat{K}_{obs}(r)$ lies below that simulation envelopes, at least for some values of r, again as expected because it is calculated for a regular point pattern.



(a) Homogeneous Poisson point process(CSR)



(b) Clustered point pattern



(c) Regularly spaced point pattern

Figure 4.11: Pointwise simulation envelopes with the K-function for different point configurations generated using the `envelope` function in `spatstat`. $\hat{K}_{obs}(r)$ is the K-function estimate for the originally simulated dataset, K_{theo} is the expected curve under independence, $\hat{K}_{hi}(r)$ and $\hat{K}_{lo}(r)$ represent the under and lower simulation envelopes respectively.

Authors such as in [4] have cautioned the use of simulation envelopes, however. This caution is mainly about their interpretation rather than their statistical validity. It is argued that it is incorrect, for example, to attach a significance level of $2k/(N+1)$ if the estimate veers outside the simulation envelopes at any point and conclude that the data are in fact not CSR at this level of significance. (Note that the k in this case is the `nrank` parameter of the `envelope` function. If $k = 1$, the maximum and minimum envelopes will be plotted, if $k = 5$ the fifth largest and fifth smallest envelopes will be plotted and so on.) Further information on simulation envelopes and a corresponding and valid pointwise test and other tests based on a Monte Carlo approach can be found in [4, 5]. We shall only consider the Maximum Absolute Deviation test in the next sub-section.

4.5.2 Maximum Absolute Deviation Test

Building on from the Monte Carlo set up in the previous section, we consider a Monte Carlo test as given in [5]. To perform the Monte-Carlo test we need to be able to simulate N (which is usually taken as 19, 99, 999) point patterns under the null hypothesis. Furthermore, we need to find a way to reduce each of the simulated datasets and the original dataset to a single value H. The value from the observed dataset

h_o is compared to the N other values $\{h_1, h_2, .., h_N\}$. The null hypothesis is rejected at a significance level of $1/(N+1)$ if $h_o > h_{max} = \max\{h_1, h_2, .., h_N\}$.

The value of the test statistic that we shall reduce the point patterns to is the Maximum Absolute Deviation (MAD). This is defined as the maximum difference between the summary function calculated for the dataset and the theoretical function under the null hypothesis [4, 5]. In our case the summary function is the estimate of equation (4.17) given by equation (4.20) and the theoretical function under the null hypothesis of independence between the points and the linear network is πr^2. Thus the MAD value is:

$$H = \max_{0 \leq r \leq r_{max}} |\hat{K}_{LX}(r) - \pi r^2|. \tag{4.22}$$

More generally we reject the null hypothesis if h_o is greater than the k_{th} largest value, giving a significance level of $\alpha = k/(N+1)$. This test is one sided since we are considering the absolute value of the largest deviation from the null hypothesis curve. This means that even if there is a repellent relationship, that is $\hat{K}_{LX}(r)$ lies below πr^2, the quantity in equation (4.22) above would still be positive. Furthermore, perhaps more clearly, if the value of H is small, that implies that the empirical curve is close to the curve representing the null hypothesis. Alternatively, we could calculate a p-value for the test which is calculated as:

$$\text{p-value} = \frac{\sum_{i=1}^{N} I(h_i \geq h_o) + 1}{N+1}.$$

4.5.3 Second deviation test

Diggle and Chetwynd in [19] proposed another method of assessing departures from the null hypothesis and the authors in [13] use this method as well in their study of the defined function, given by equation (4.17). The following function is defined:

$$H(r) = K_{LX}^{theo}(r) - K_{LX}(r) \tag{4.23}$$

where $K_{LX}^{theo}(r) = \pi r^2$ is the theoretical value of the function under the null hypothesis of independence and $K_{LX}(r)$ is as defined in equation (4.17). When implemented in [19], it was implemented for point-to-point relationships and the context was that of spatial aggregation of diseases. Two types of points were considered, the cases (type 1) and the controls (type 2). To investigate spatial aggregation, the authors defined $K_{11}(r) - K_{22}(r)$ so that positive values of this quantity implied some spatial clustering of type 1 events, i.e. the cases, over the degree of spatial clustering of type 2 events i.e. the controls. Therefore in our present context of points and linear networks, we may interpret positive values of $K_{LX}(r) - K_{LX}^{theo}(r)$ as clustering of points around the linear network. More clearly we can say that positive values imply that there is more 'length of road', on average, than there would be if the points were independent of the road.

We note that in most applications we will not have the exact theoretical form of $K_{LX}(r)$ hence we estimate

it with $\hat{K}_{LX}(r)$ so that equation (4.23) becomes

$$\hat{H}(r) = K_{LX}^{theo}(r) - \hat{K}_{LX}(r). \tag{4.24}$$

Since the function is calculated for a range of r values, we need to find a way to reduce it to a single test statistic value, as for the MAD test. Let K be the number of r values for which the function is evaluated. The test statistic suggested in [13] and originally in [19] is

$$H = \sum_{k=1}^{K} \frac{\hat{H}(r_k)}{\sqrt{\text{var}[\hat{H}(r_k)]}}. \tag{4.25}$$

Under the null hypothesis, H is approximately normally distributed with mean $E[H] = 0$ and variance

$$\text{var}[H] = K + 2\sum_{j=2}^{K}\sum_{k=1}^{j-1} \text{corr}[\hat{H}(r_j), \hat{H}(r_k)]$$

as given in [19].

This can be extended to cases where we want to compare summary functions calculated from point and line configurations for different values of d. The notation changes slightly to indicate this. Equation (4.24) becomes

$$H(r) = K_{LX}^{d_1}(r) - K_{LX}^{d_2}(r)$$

where d_1 and d_2 are values that determine the strength of attraction of the points to the linear network. Under H_0, $d_1 = d_2$. The approach is then to determine the approximated sampling distribution of (4.25) under H_0 for a given value of $d = d_1 = d_2$, where $\hat{H}(r_k) = \hat{K}_{LX}^{d_1}(r) - \hat{K}_{LX}^{d_2}(r)$. The test statistic based on the data is calculated as

$$H_{obs} = \sum_{k=1}^{K} \frac{\hat{H}(r_k)}{\sqrt{\text{var}[\hat{H}(r_k)]}} \tag{4.26}$$

where $\hat{H}(r_k) = \hat{K}_{LX}^{d_1}(r) - \hat{K}_{LX}^{emp}(r)$, and as usual, compared to the relevant critical values obtained from the null sampling distribution. The idea is that if the points in the dataset are within distance d_1 of the linear network and the proposed mechanisim by which the points are displaced is valid, then the empirical estimate $\hat{K}_{LX}^{emp}(r)$ should not be significantly different from $\hat{K}_{LX}^{d_1}(r)$, i.e., the estimate calculated from points-to-line configurations when $d = d_1$. The performance of this test is considered in Section 5.4.

4.6 Conclusion

In this chapter the K-function and cross-K function were discussed as a means of characterising the second order properties of unmarked and marked point processes respectively. The extension of these to characterise the spatial relationship between a linear network and a point pattern, as proposed in [13], was discussed and the empirical estimate derived. Edge effects and the problems they cause when they are

not taken into account properly were discussed along with Ripley's edge correction. Confidence intervals for the empirical estimates were also presented, with preference being given to Loh's bootstrap over the block bootstrap because we need not concern ourselves with dividing the study window in the former. In addition to confidence intervals, another graphical method in the form of simulation envelopes was discussed in the Monte Carlo setting. Finally, formal hypothesis tests namely the MAD test and the test proposed in [19] were presented. We next consider a number of simulations in Chapter 5 including testing the utility of the function defined in equation (4.17) and investigation into the performance of these tests.

Chapter 5

Simulations

In this chapter we conduct a number of simulations and hypothesis tests to investigate how well the function defined in equation (4.17) can pick up various point to line configurations. We consider a number of different settings laid out in Section 5.1 and present the graphical results. A discussion of these results is presented in Section 5.2. Results of formal hypothesis tests for these simulations using the MAD method and the second deviation test (discussed in Section 4.5.3) are discussed in Sections 5.3 and 5.4 respectively.

5.1 Simulations and testing of the empirical estimate

Recall the following function which is an estimate of the theoretical function, defined in equation (4.17), for detecting spatial dependency:

$$\hat{K}_{LX}(r) = \frac{|D|}{n \times n_L} \sum_{i=1}^{n} \sum_{j=1}^{n_L} {w_{ij}}^{-1} I(0 < ||x_i - y_j|| \leq r) \tag{5.1}$$

where n is the number of points in the point pattern, n_L is the number of divisions of the linear network L and $|D|$ is the area of the observation window. In this section we do a number of simulations to test the function defined in [13] using equation (5.1) as the estimate. In Section 2.6 the Poisson line process was introduced along with some alterations. We shall use a line process to simulate a linear network. While this is the most ideal situation and road networks are not Poisson line patterns, authors have used this method to represent road networks [9].

There are a number of scenarios we consider to test the function's ability to pick up spatial dependency. We start with the base case of independence between the points and the linear network. To set this up, a linear network is simulated using a Poisson line process. Independently of that, a Poisson point pattern is simulated and superimposed on the linear network. It is clear that these are independent and because we

have a Poisson point pattern, the function should be πr^2. Non-stationary point patterns superimposed on the linear network are also considered.

We then move on to the other relationships that were discussed in earlier chapters, those of attraction and repulsion. To simulate the case where points are attracted to or repelled by the linear network we follow the way used in [13]. For attraction we simulate points lying exactly on the linear network. It is clear that these points are dependent on the linear network. As a way to 'reduce' this dependence, the points on the network are shifted by some mechanism. So for a point with coordinates (x_i, y_i), $x_i \to x_i' = x_i + d.u$ and $y_i \to y_i' = y_i + d.u$ under the constraint $(x_i', y_i') \in D$, where $U \sim \text{uni}(-1, 1)$ and $d \geq 0$ determines the strength of attraction of the points to the linear network, i.e. as d increases the strength of attraction should decrease. The function will then be tested on these shifted points for varying values of d. Particularly, we shall use 0.01, 0.05 and 0.10 for values of d because all the simulations are done on a unit square window.

The case for a repellent relationship will be done in a slightly different manner. The idea here is to impose a minimum distance between the points and the linear network such that the points seem to be repelled by the linear network. This is similar to Matérn's Model I or the SSI mechanism. A challenge in this case, as we saw with the SSI mechanism in Section 2.4 is that in some cases it may be difficult to fit the desired number of points on the window, while still maintaining the minimum required distance between the points and the linear network. This is especially true for Matérn's Model I because if the user inputs a Poisson parameter of 100 for example with $r = 0.10$, and the window is a unit square, the number of points returned are way less than one would desire if trying to generate a large regular point pattern. This is clearly because of the size constraint of the window and the distance r chosen. The same problem applies to the SSI mechanism because even if the user specifies a certain number of points desired, if the window is relatively small and r is large, the algorithm will stop after a maximum number of attempts is reached and the number of points returned will be less than what the user intended.

To get around the issues discussed above the `rSSI` function in `spatstat` will be used with some alterations. One of the optional arguments of the function is an initial point pattern `x.init`. As explained in the documentation, if this argument is supplied, the returned point pattern will be points that are at least distance r away from `x.init` and the initial point pattern itself. So once we have a simulated linear network, it is broken up into points and converted to a point pattern. This is the 'point pattern' that will be used as `x.init`. Once this is returned, the points that are at least distance r from the points in `x.init` are separated from `x.init` and these points represent the points repelled by the linear network. One caveat is that the returned points are also distance r away from each other, which is not required for this simulation study. It may also return fewer number of points that are required because of that additional constraint. So to achieve the number of points we desire, we can simulate multiple point patterns this way until we have reached the value n we want. Then these point patterns are superimposed and this

point pattern will have points that are at least distance r from the linear network. A slight modification can be made to the function as well, where in the case that a point is within distance r of the linear network it is accepted with a certain probability. This allows the generation of points to be slightly faster while accepting the fact that some points will indeed be closer than r. One could argue that this is more representative of a real world situation because it is only in an idealised world that all points will be strictly at least distance r away from the linear network. One should be wary though about the choice of this probability because if it is too high then most of the points will lie within distance r of the linear network which is not desirable if we would like to simulate the case of a repellent relationship as discussed.

There will be a number of combinations of point and line configurations. In addition to the above, it will be of interest to test how well the function works when the line pattern used as the linear network is non-stationary. We shall consider all four cases of line configurations discussed in Section 2.6:

- Case 1: Uniform p and α
- Case 2: Non-uniform p and uniform α
- Case 3: Uniform p and non-uniform α
- Case 4: Non-uniform p and α

Although the case where p and α are negatively skewed was discussed in Section 2.6, we shall only focus on the case of positive skewness. Emphasis is placed on a non-uniform distribution, not necessarily how it is non-uniform. In summary, for each of the four cases we will consider the independence case as discussed above, and attraction and repulsion relationships. In the simulation results that follow, the resulting K_{LX} estimate is in black, the 95% confidence intervals are in red and the independence curve is in blue.

5.1.1 Independence

Figure 5.1 shows Poisson point patterns with different levels of intensity (λ_x's) superimposed on linear networks formed from line patterns of different λ_L's. Figure 5.2 shows non-stationary point patterns on stationary linear networks, and finally a stationary point pattern on non-stationary linear networks in Figure 5.3.

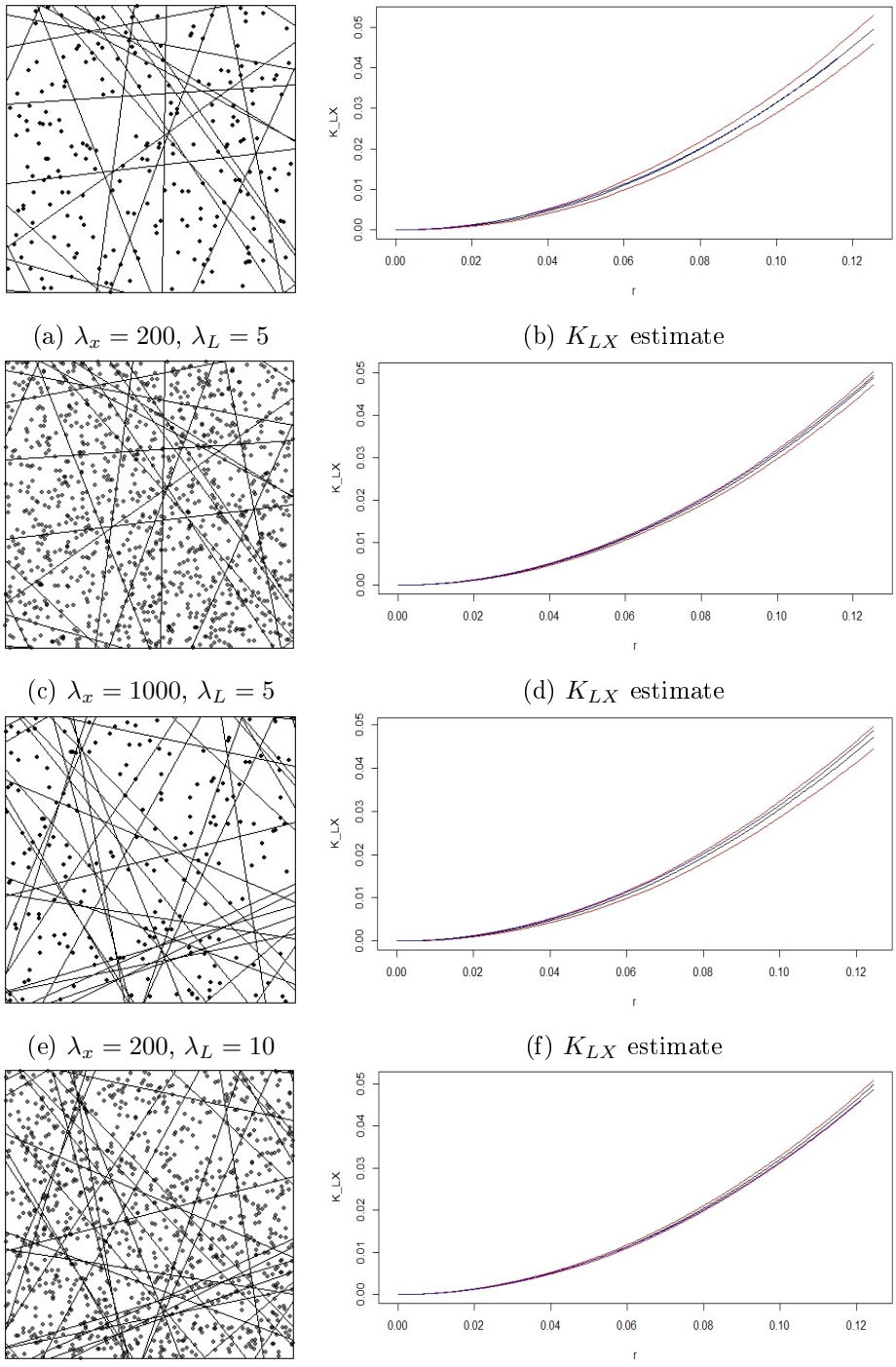
5.1.2 Attraction

In this section we test the function on a set of points that are 'attracted' to the linear network. This is done in the way explained in the introduction to Section 5.1. We first illustrate how points are generated on the

linear network and then shifted as described before. A total of 100 uniform points were generated on the linear network using the `runiflpp` function and the result is shown in Figure 5.4(a). The resulting point patterns shifted by different values of d are shown in (b)-(d). The linear network used is generated using the `rpoisline` function (Case 1) with parameter 5. Figure 5.5 shows the corresponding $\hat{K}_{LX}$ estimates (i.e. corresponding to the shifted point patterns in Figure 5.4(b)-(d)). Figure 5.6 shows the estimates of the function for different values of d and n for Case 1. We then consider Cases 2-4 in Figure 5.7 for $n = 1000$.

5.1.3 Repulsion

Here we consider points repelled by the linear network. A minimum distance of r is imposed between the points and the linear network. The values of r considered are 0.01, 0.05 and 0.10. The probability threshold used is 0.01, so that if a point is closer than the specified distance it is only accepted with probability 0.01. As argued before, for the larger distances some of the points will lie closer than 0.10, for example, because of the constraints. This however enables us to get the desired number of points. We first consider 100 points and the results of imposing a minimum distance between the points and linear network are shown in Figure 5.8. The linear network is generated the same way as above, using the `rpoisline` function with parameter 5. Figure 5.9 shows the corresponding K_{LX} estimates for the configurations in Figures 5.8. Figure 5.10 shows the estimates of the function for different values of r and n for Case 1. We then consider Cases 2-4 in Figure 5.11 for $n = 1000$.



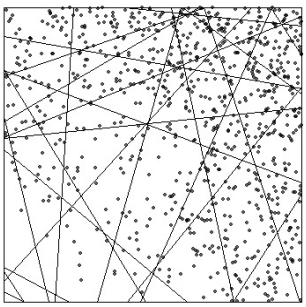
(a) $\lambda_x = 200$, $\lambda_L = 5$ (b) K_{LX} estimate

(c) $\lambda_x = 1000$, $\lambda_L = 5$ (d) K_{LX} estimate

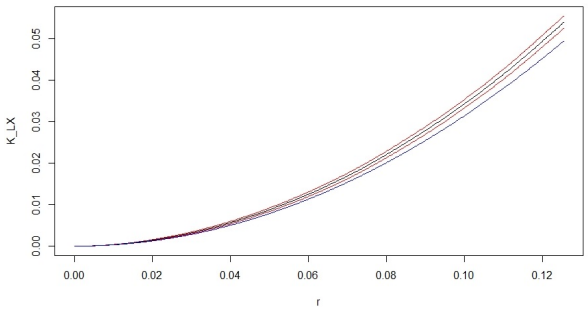
(e) $\lambda_x = 200$, $\lambda_L = 10$ (f) K_{LX} estimate

(g) $\lambda_x = 1000$, $\lambda_L = 10$ (h) K_{LX} estimate

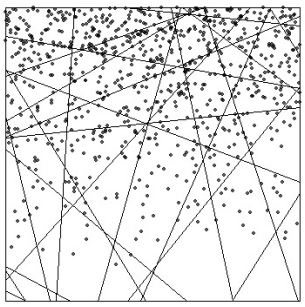
Figure 5.1: Poisson point patterns superimposed on linear network with corresponding K_{LX} estimates. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



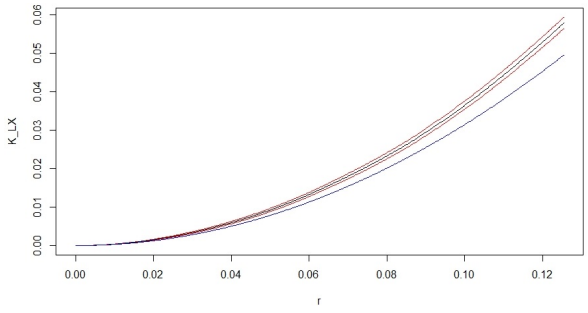
(a) $\lambda = x^2 + y^2$, n=821, $\lambda_L = 5$



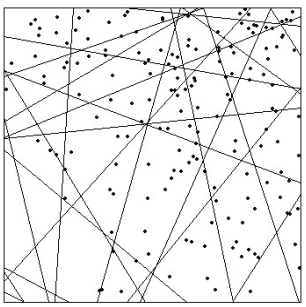
(b) K_{LX} estimate



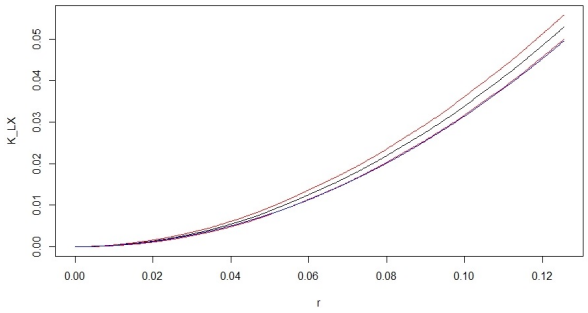
(c) $\lambda = y^2$, n=784 $\lambda_L = 5$



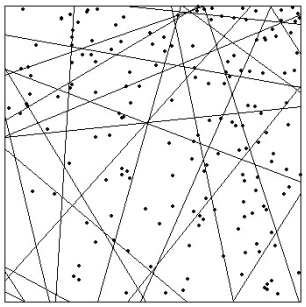
(d) K_{LX} estimate



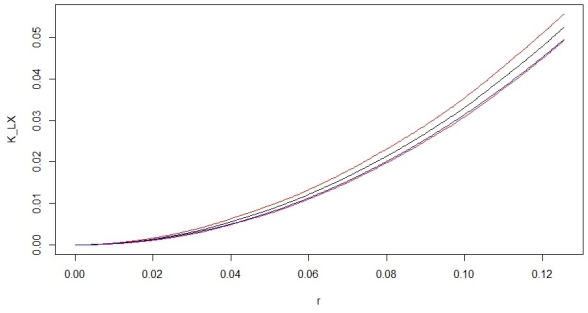
(e) $\lambda = x + y^2$, n=168, $\lambda_L = 5$



(f) K_{LX} estimate

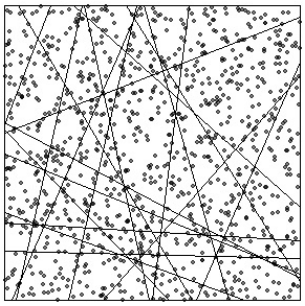


(g) $\lambda = x^2 + y$, n=170, $\lambda_L = 5$

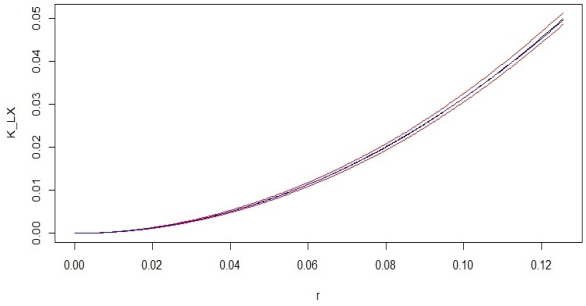


(h) K_{LX} estimate

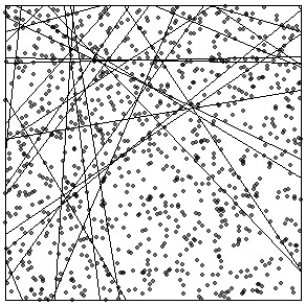
Figure 5.2: Non-stationary Poisson point patterns superimposed on linear network with corresponding K_{LX} estimates. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



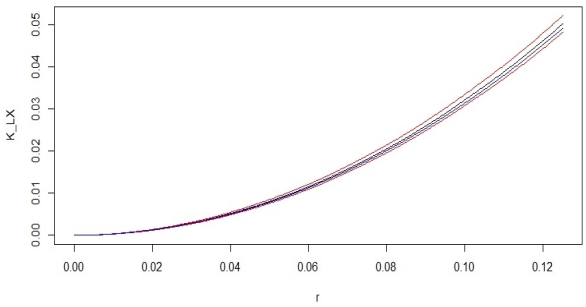
(a) $\lambda = 1000$, Case 2: $\lambda_L = 5$



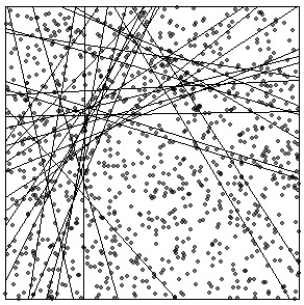
(b) K_{LX} estimate



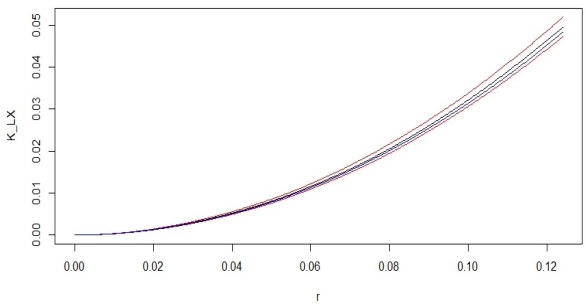
(c) $\lambda = 1000$, Case 3: $\lambda_L = 5$



(d) K_{LX} estimate

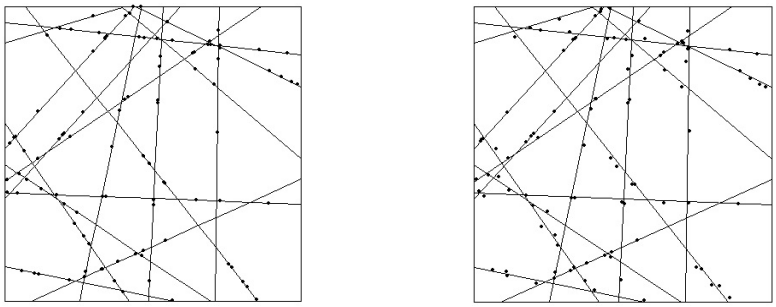


(e) $\lambda = 1000$, Case 4: $\lambda_L = 5$

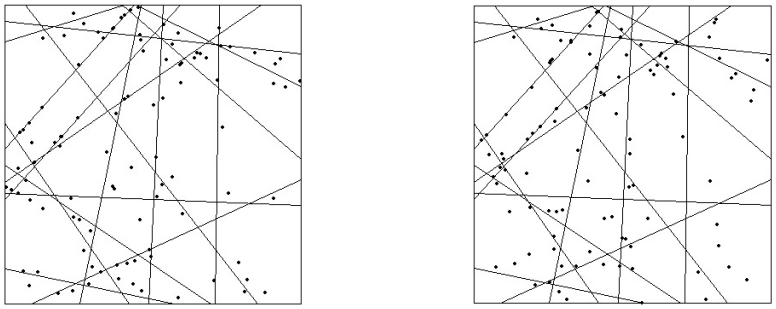


(f) K_{LX} estimate

Figure 5.3: Poisson point patterns superimposed on a non-stationary linear network with corresponding K_{LX} estimates. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.

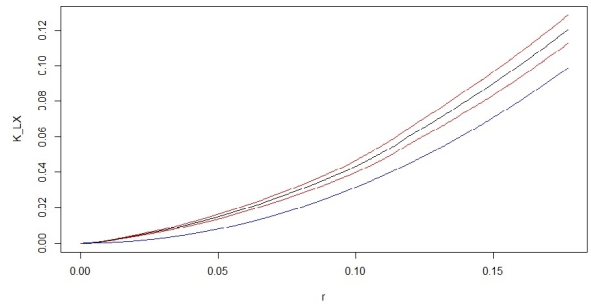


(a) Points lying exactly on the linear network (b) Points shifted by $d = 0.01$

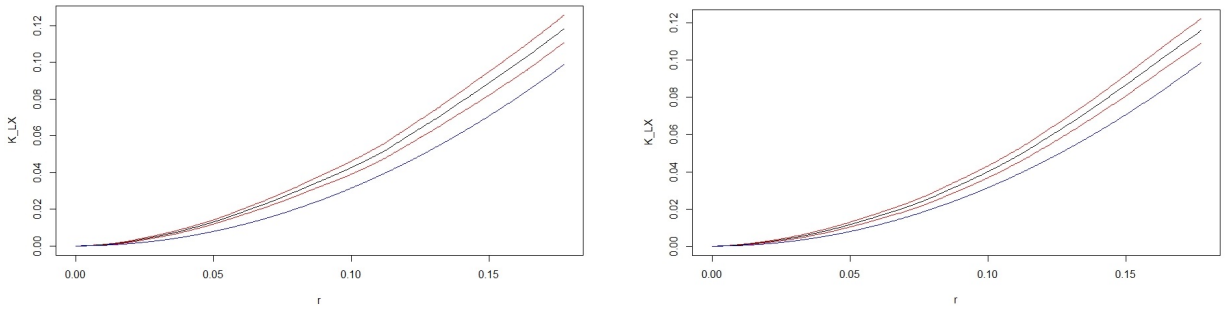


(c) Points shifted by $d = 0.05$ (d) Points shifted by $d = 0.10$

Figure 5.4: Illustration of shifting of points initially on a linear network by different values of d to reduce the degree of dependence on the linear network. The linear network is simulated using `rpoisline` function with parameter 5.

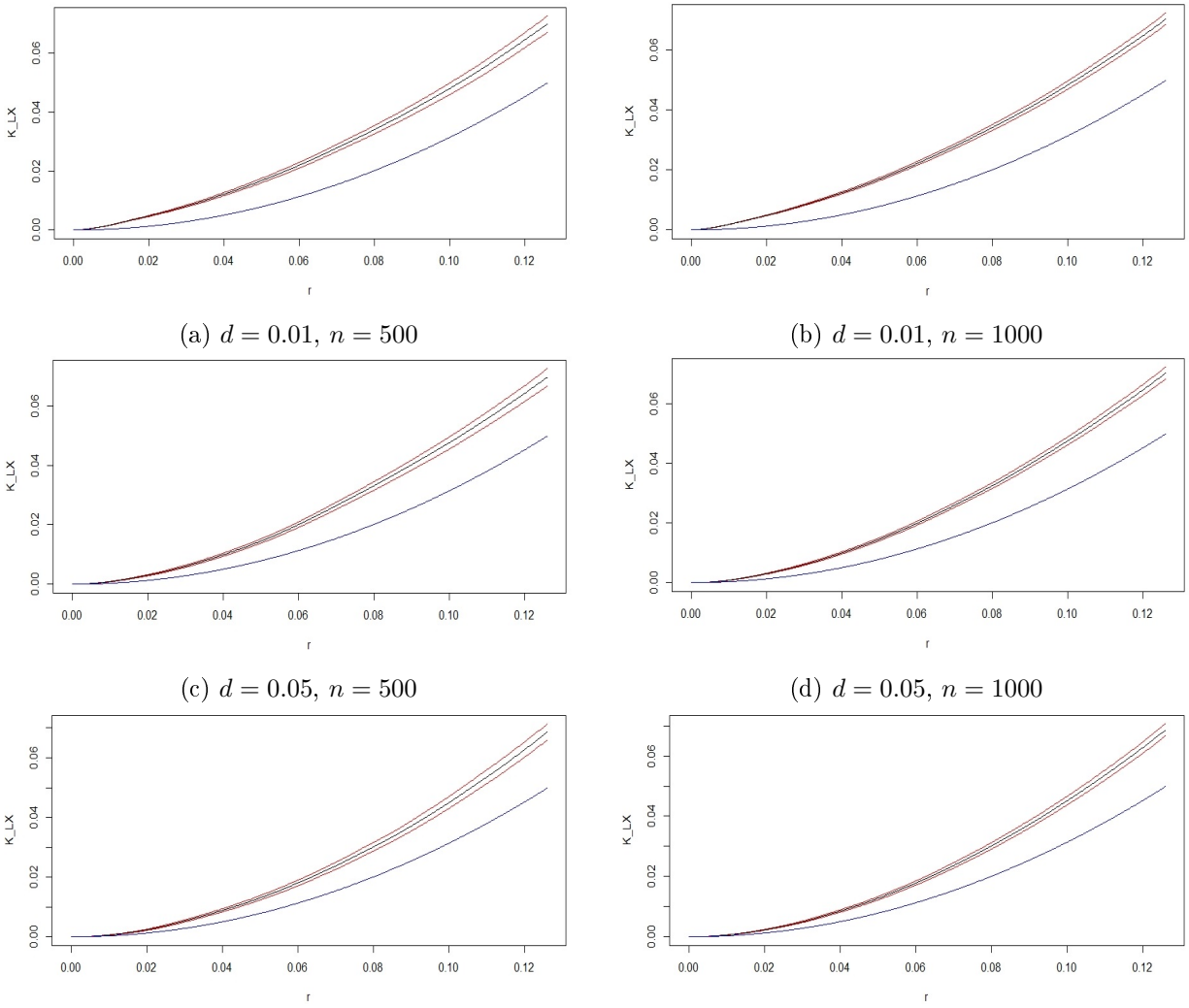


(a) $d = 0.01$



(b) $d = 0.05$ (c) $d = 0.10$

Figure 5.5: K_{LX} estimates for Case 1 for varying values of d, for $n = 100$ points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



(a) $d = 0.01$, $n = 500$

(b) $d = 0.01$, $n = 1000$

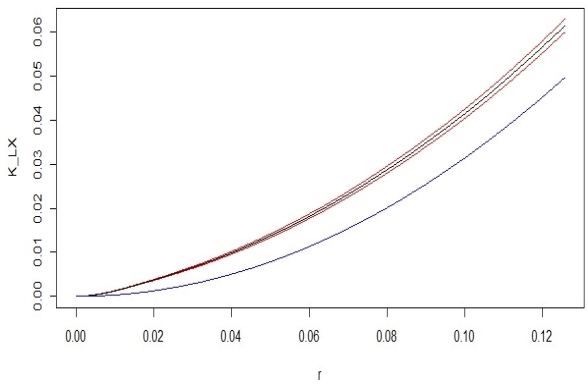
(c) $d = 0.05$, $n = 500$

(d) $d = 0.05$, $n = 1000$

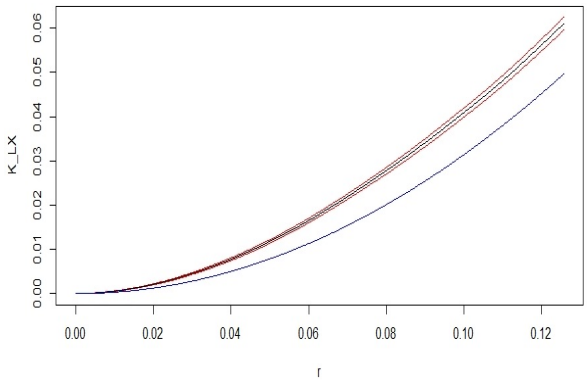
(e) $d = 0.10$, $n = 500$

(f) $d = 0.10$, $n = 1000$

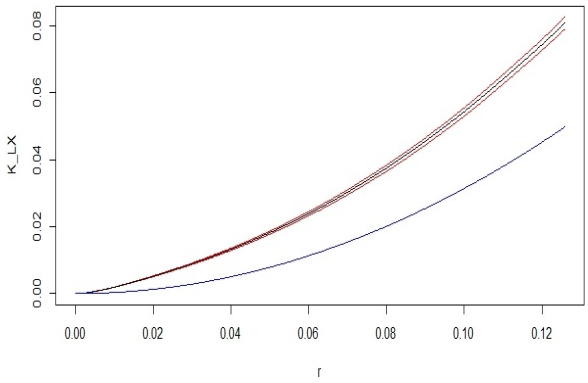
Figure 5.6: K_{LX} estimates for Case 1 for different values of d and n. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



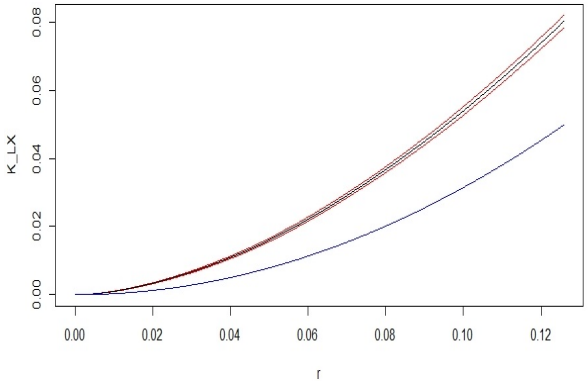
(a) Case 2, $d = 0.01$



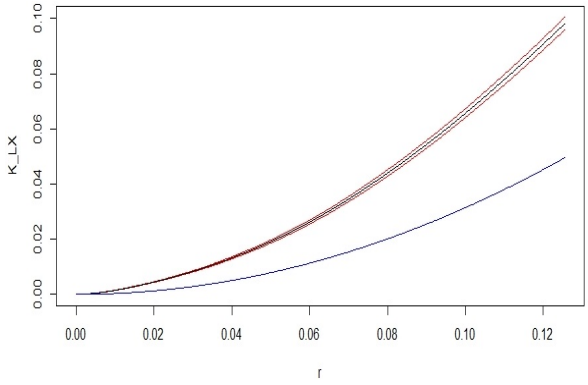
(b) Case 2, $d = 0.05$



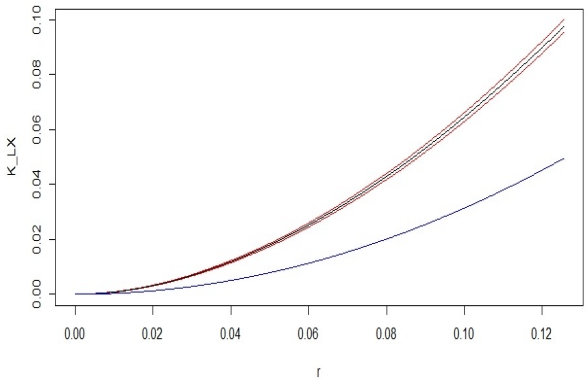
(c) Case 3, $d = 0.01$



(d) Case 3, $d = 0.01$

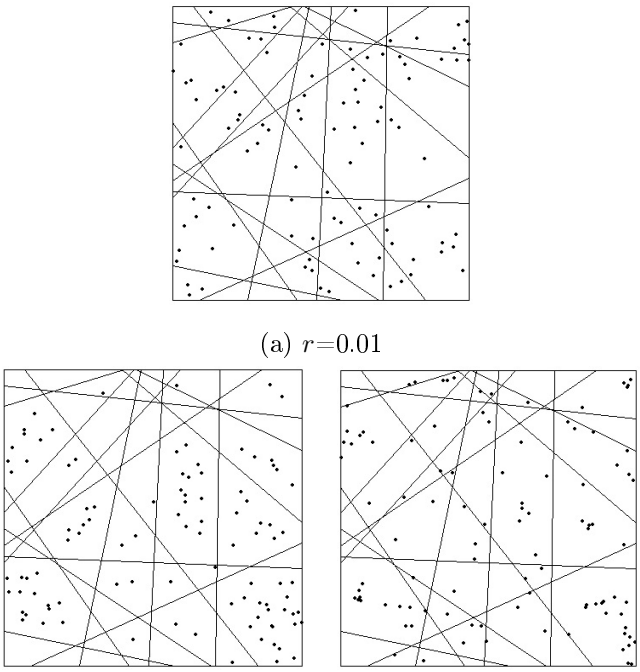


(e) Case 4, $d = 0.01$



(f) Case 4, $d = 0.05$

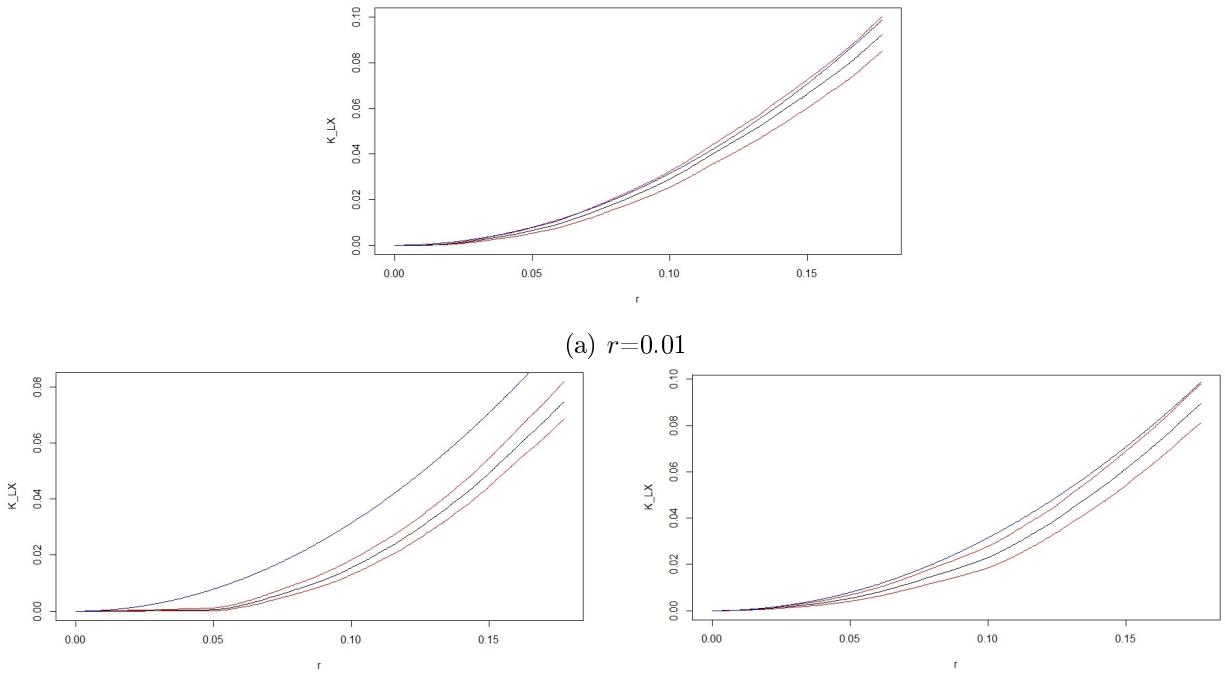
Figure 5.7: K_{LX} estimates for 3 cases and different values of d, $n = 1000$ points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



(a) r=0.01

(b) r=0.05 (c) r=0.10

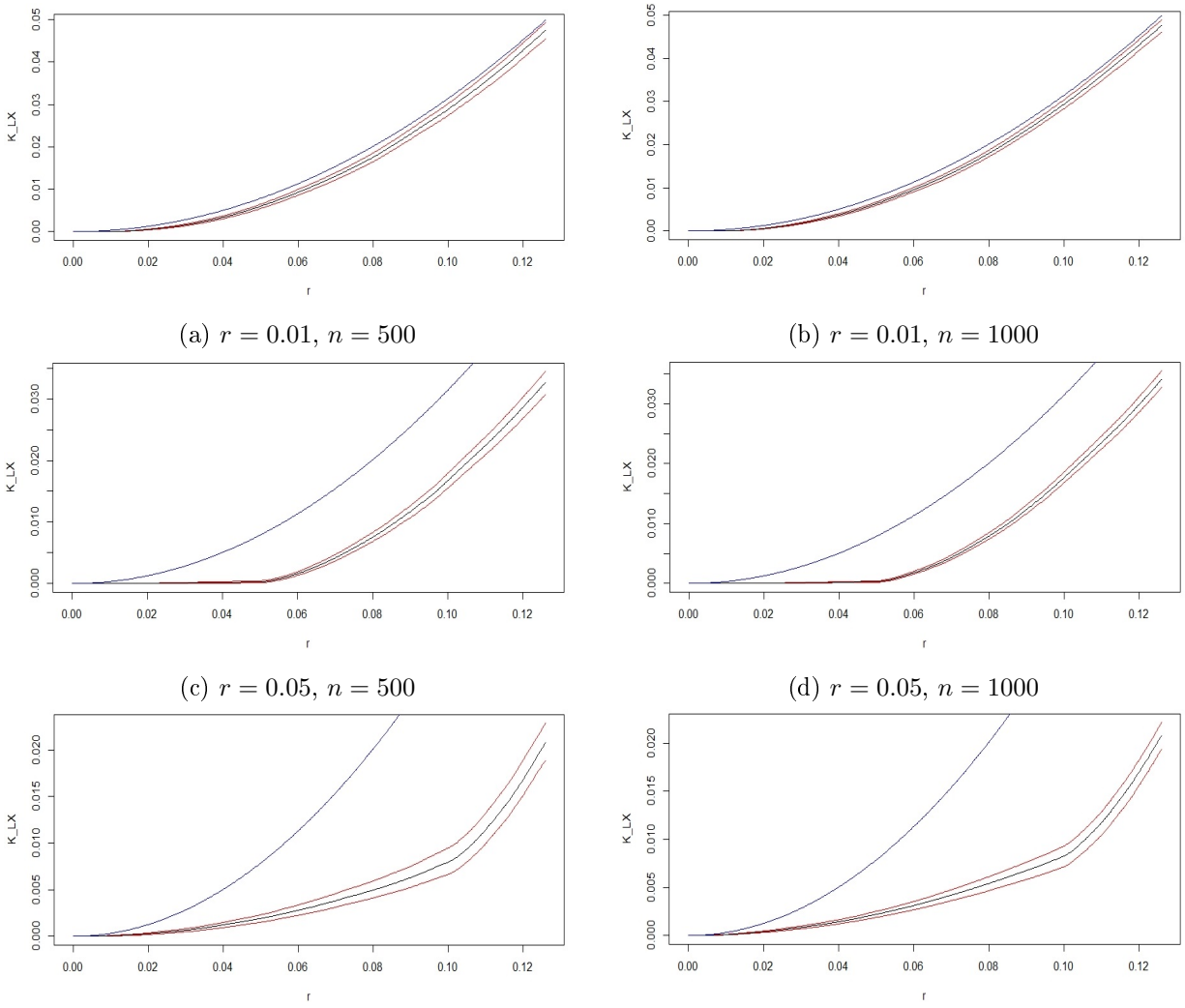
Figure 5.8: Illustration of repulsion of points by linear network for different values of the repulsion distance r. The linear network is simulated using the `rpoisline` function with parameter 5.



(a) r=0.01

(b) r=0.05 (c) r=0.10

Figure 5.9: K_{LX} estimates for different values of r, for n=100 points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



(a) $r = 0.01$, $n = 500$

(b) $r = 0.01$, $n = 1000$

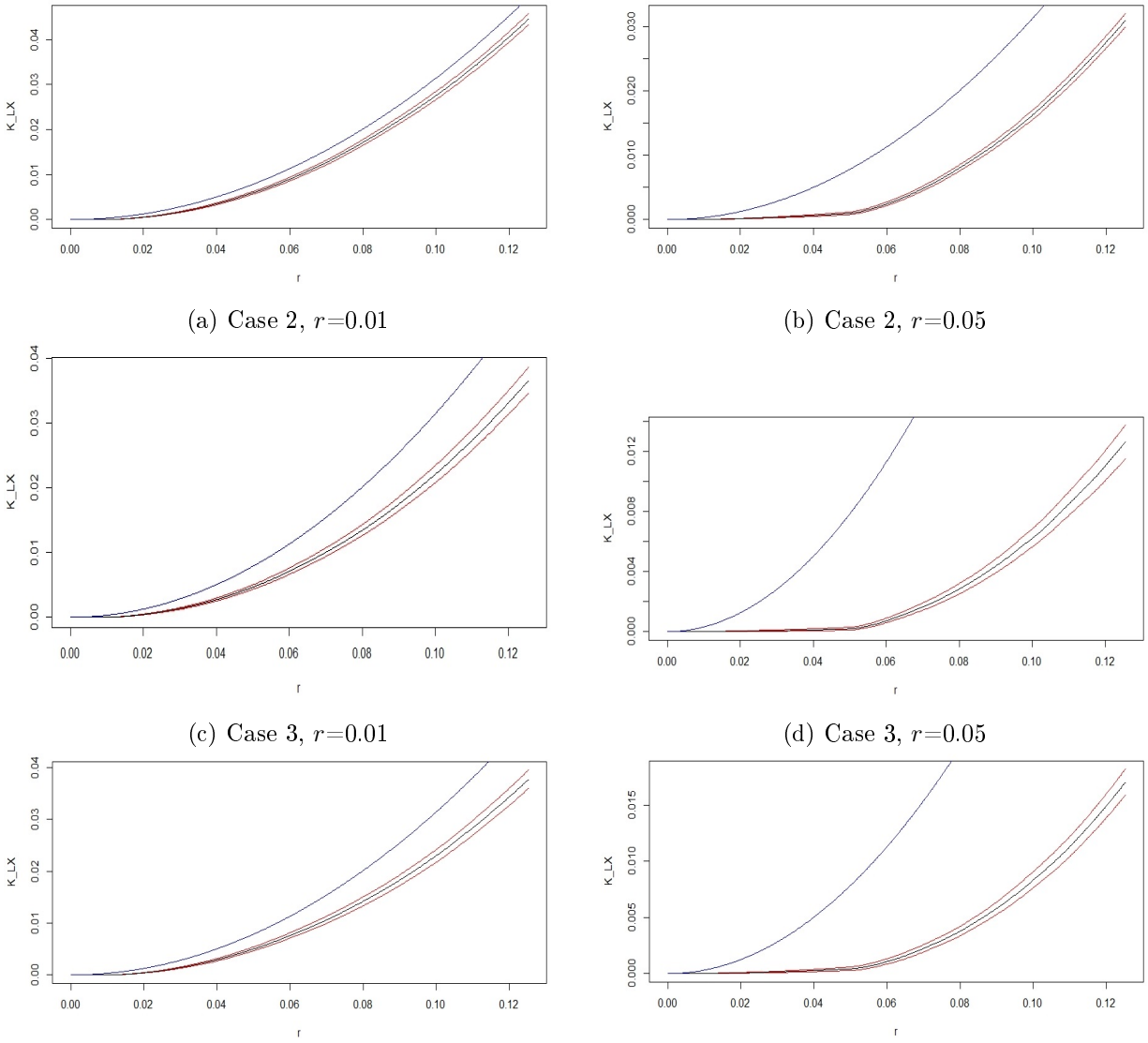
(c) $r = 0.05$, $n = 500$

(d) $r = 0.05$, $n = 1000$

(e) $r = 0.10$, $n = 500$

(f) $r = 0.10$, $n = 1000$

Figure 5.10: K_{LX} estimates for Case 1 for different values of r and n. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue.



(a) Case 2, r=0.01

(b) Case 2, r=0.05

(c) Case 3, r=0.01

(d) Case 3, r=0.05

(e) Case 4, r=0.01

(f) Case 4, r=0.05

Figure 5.11: K_{LX} estimates for cases 2-4 and different values of r, for n=1000 points. The empirical estimate is in black, the confidence intervals in red and the expected curve under independence in blue. The parameter used for the line process was 5 in all cases.

5.2 Discussion

We started with the case of independence where a Poisson point pattern was superimposed on a linear network as shown in Figure 5.1. Since the points are independent of the linear network we expect that $\hat{K}_{LX} \approx \pi r^2$ as was discussed in Section 4.4. Studying the K_{LX} estimates shown in Figure 5.1 we can see that this is the case for the simulated point patterns and linear networks. The K_{LX} estimates lie close to the curve of independence (in blue) of πr^2. The curves for independence also lie within the confidence intervals that are shown in red. The K_{LX} estimates are therefore not significantly different from the expected result under independence. This is of course in line with what we expected because of the way the data was simulated.

Figure 5.2 shows non-stationary Poisson point patterns superimposed on stationary linear network. The two structures were simulated independently so there is no reason to suspect that there is spatial dependency between the two. Figures 5.2 (f) and (h) seem to suggest that the non-stationarity of the point pattern does not impact the estimate, that is, the K_{LX} estimates are all close to πr^2 and the independence curve lies within the confidence intervals of the estimates. This suggests that the estimates are not significantly different from the case of independence. However, Figures 5.2(b) and (d) seem to suggest that there is a spatial relationship of attraction between the points and the line segments since the estimated curve lies above that of independence, and the curve for independence lies outside the confidence intervals of the estimate. The reason behind these results could be the number of points that were used, where for the smaller number of points we have no significant difference and for the large number of points (and thus shorter confidence intervals) the results seem to suggest a significant difference between the estimates and the independence case. But of course the main reason is the non-stationarity of the point pattern. In Section 4.4, one of the assumptions was that of a stationary point pattern and this is violated here.

For non-stationary linear networks we consider Figure 5.3. Stationary Poisson point patterns are superimposed on non-stationary linear networks. Once again, the two structures are simulated independently so there should be no spatial dependency suggested by the K_{LX} estimates. The results shown in figures (b), (d) and (f) show that the K_{LX} estimates of the corresponding point-to-line configuration all seem not to be significantly different from the independence curve. Furthermore, the curve for independence lies within the confidence intervals for all these three cases.

Figure 5.5 shows the K_{LX} estimates for Case 1 when the point patterns are attracted to the linear network as illustrated in Figure 5.4. The estimates of the function lie consistently above the curve for independence. This implies that there is 'more length of road' within distance r of the points than would be expected under independence. Thus there appears to be clustering of points around the linear network which is what we expected because of the way the points were simulated. The Poisson curve of independence also lies outside the confidence intervals of our K_{LX} estimates for all three values of d. Figure 5.6 shows the

K_{LX} estimates for $n = 500$ and 1000 to further illustrate the usefulness of the function on picking up on spatial dependency.

Figure 5.7 shows the K_{LX} estimates for Cases 2-4 for different values of d and $n = 1000$ points. The function appears to be able to pick up the spatial dependency between the points and the linear network. One difference between these three cases and case 1 is that the estimates seem to lie much higher above the independence curve for the same value of d. This could be because of the non-stationarity of the lines. That is, the lines themselves favour the top-right corner of the window so a majority of them lie there. Therefore, if points lying exactly on the network are simulated then shifted, they will be around the area that is highly concentrated with lines. This will then result in there being more 'length of road' around the points than if the line pattern was stationary.

Figure 5.8 shows point-line configurations where the points are repelled by the linear network for Case 1. This can be seen particularly for $r = 0.05$ where the points seem to form clusters away from the line segments as if there is some barrier between the two structures. The K_{LX} estimates for this case are shown in Figure 5.9. The K_{LX} estimates lie below the curve of independence for most, if not all the values for which the function is calculated. This implies there is less length of road within distance r of the points which is in line with our expectations. For Figures 5.9(a) and (c) however for the simulated data the difference between the estimates and the curve for independence does not seem to be significant since the curve lies inside the confidence intervals at least for some values of r (the values on the x-axis, not the minimum distance value). For $r = 0.01$ the reason may be that the barrier distance is not large enough to be reflected in the graph. For $r = 0.10$ the reason may have to do with issues raised before, such as there not being an optimal spot for a point to be at least 0.10 units away from the point linear network while also still being in the window. Recall that these points are accepted with probability 0.01. The case where the repellent relationship is most apparent is $r = 0.05$ where the graph is close to zero at distances less than 0.05 and then almost immediately picks up after 0.05. It is also consistently below the independence curve, and the independence curve lies outside the calculated confidence intervals. Figure 5.10 shows the K_{LX} estimates for $r = 0.01$, 0.05 and 0.10 and for $n = 500$ and 1000 to further illustrate the usefulness of the function on picking up on this repulsion of the points by the line segments.

Figure 5.11 shows the K_{LX} estimates for Cases 2-4 of line configurations for $r = 0.01$ and 0.05. The function behaves in a similar way to what we had in Case 1. For r=0.05 the function almost immediately picks up at 0.05 and is consistently below the independence curve. The independence curve also consistently lies outside the confidence intervals of the estimate. For r=0.01 the estimated curve also lies below the independence case but to a less extent than for r=0.05 because the barrier between the points and the line segments is smaller.

5.3 Power of the MAD test

5.3.1 Setup

In this section we consider the power MAD test for various null and alternative hypotheses. The first null hypothesis is that of CSR, that is, the points are independent of the linear network. The alternative hypothesis is that the points lie within distance d of the linear network. The point pattern under the alternative hypotheses will be generated in the same way that was described in Section 5.1 for $d = 0.01$, 0.05 and 0.10. The simulations will be for $n = 100$, 500 and 1500.

The power of the MAD test will be calculated using the procedure described in [24]. The first step is to simulate M sets of N point patterns under H_0. For each of the N point patterns calculate the test statistic $h = \max_{0 \leq r \leq r_{max}} |\hat{K}_{LX}(r) - \pi r^2|$. Denote the N test statistics from the i^{th} set of point patterns as $H^{(i)}_{1:N} = \{H_1, H_2, ..., H_N\}$, where H_1 is the test statistic for the first point pattern of the i^{th} of the M sets. This means that there will be M of the $H^{(i)}_{1:N}$'s: $H^{(1)}_{1:N}, H^{(2)}_{1:N}, ..., H^{(M)}_{1:N}$. For each of these calculate the critical value at α level of significance. That is, for $H^{(i)}_{1:N} = \{H_1, H_2, ..., H_N\}$ determine the $(1-\alpha)^{th}$ percentile, since this is an upper-tailed test. Therefore at the end there will be M critical values.

Under the alternative hypothesis, simulate T point patterns. Calculate the test statistics $H^{(1)}$,$H^{(2)}$,..., $H^{(T)}$. The power is then estimated as:

$$\hat{\pi} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=1}^{T} I(H^{(t)} > Z(H^{(m)}_{1:N}, \alpha))$$

where $Z(H^{(m)}_{1:N}, \alpha)$ is the critical value calculated at an α level of significance from the m^{th} set of point patterns. For the simulations conducted, M was taken to be 100, $N = 199$ and $T = 1000$. In addition to using K_{LX}, the stabilised version of the function is used, i.e. $\sqrt{K_{LX}(r)/\pi}$ because it is argued that this improves the power of the test by authors such as in [5]. The standard levels of significance i.e. 0.01, 0.05 and 0.10 are considered.

5.3.2 One-sided alternative

We first consider the null hypothesis of CSR. Table 5.1 shows the powers of the MAD test under various alternative hypotheses and values of n. As can be seen in Table 5.1 both versions of the MAD tests perform satisfactorily when the null hypothesis is that of CSR. The power values are all high, even with a few number of points. As the number of points are increased, the power effectively is 1. All the test statistics calculated from the alternative hypotheses are rejected at all levels of significance. Table 5.2 shows the powers of the MAD test when the null hypothesis is that the points are within distance 0.01 of the linear network and the alternative is that the points are within 0.05 or 0.10. The powers of the

n	Value of d under alternative	α	Power using K_{LX}	Power using stabilised K_{LX}
100	0.01	0.01	0.943	1.000
		0.05	0.995	1.000
		0.10	0.998	1.000
	0.05	0.01	0.936	0.990
		0.05	0.993	1.000
		0.10	0.998	1.000
	0.10	0.01	0.850	0.881
		0.05	0.970	0.979
		0.10	0.987	0.993
500	0.01	0.01	1.000	1.000
		0.05	1.000	1.000
		0.10	1.000	1.000
	0.05	0.01	1.000	1.000
		0.05	1.000	1.000
		0.10	1.000	1.000
	0.10	0.01	1.000	1.000
		0.05	1.000	1.000
		0.10	1.000	1.000
1500	0.01	0.01	1.000	1.000
		0.05	1.000	1.000
		0.10	1.000	1.000
	0.05	0.01	1.000	1.000
		0.05	1.000	1.000
		0.10	1.000	1.000
	0.10	0.01	1.000	1.000
		0.05	1.000	1.000
		0.10	1.000	1.000

Table 5.1: Power values for the MAD test when the null hypothesis is CSR and the alternative is the points shifted by a value of d from the linear network.

test are clearly unsatisfactorily low. This is perhaps unsurprising because the test is one sided therefore the null hypothesis is only rejected for values of h large enough. To further explain this, consider Table 5.3 where values of h were calculated for both the normal K_{LX} function and the stabilised version. For each value of d, 1000 point patterns were generated on a unit square window, each with 500 points. The percentiles for each of these cases were calculated. As can be seen, the all percentiles for $d = 0.05$ and

n	Value of d under alternative	α	Power using K_{LX}	Power using stabilised K_{LX}
100	0.05	0.01	0.011	0.005
		0.05	0.047	0.023
		0.10	0.088	0.046
	0.10	0.01	0.004	0.002
		0.05	0.016	0.006
		0.10	0.039	0.009
500	0.05	0.01	0.011	$<$0.001
		0.05	0.060	0.006
		0.10	0.115	0.016
	0.10	0.01	0.001	0
		0.05	0.007	$<$0.001
		0.10	0.017	$<$0.001
1500	0.05	0.01	0.019	0
		0.05	0.077	$<$0.001
		0.10	0.134	0.002
	0.10	0.01	$<$0.001	$<$0.001
		0.05	$<$0.001	$<$0.001
		0.10	0.001	$<$0.001

Table 5.2: Power values for the MAD test when the null hypothesis is that the points are within 0.01 of the linear network and the alternative is the points shifted by a value of d from the linear network.

0.10, are less than the corresponding percentiles for $d = 0.01$, meaning if the null hypothesis is the points lying within distance 0.01 and the alternative is that they lie within 0.05, and an upper tailed test is used, we will rarely reject the null hypothesis when this alternative is true.

5.3.3 Two sided alternative

One way that this unsatisfactory performance can be improved is if we consider a two sided test, particularly for the cases where the null hypothesis is not that of CSR, that is, when we have specified a value for d under the null and alternative hypotheses as well. The difference then will be that the $\alpha/2$ and (1-$\alpha/2$) percentiles are selected for the critical values, not the $1-\alpha$ suggested in Section 5.3.1. The calculation of the power is also modified to take into account the lower critical values.

We will not consider the null hypothesis of CSR because the results are likely to be similar. Table 5.4 shows the power values for the two-sided approach for the MAD test where the null hypothesis was that

using K_{LX}			
Percentile	$d = 0.01$	$d = 0.05$	$d = 0.10$
50%	0.009812	0.009531	0.008473
90%	0.010791	0.010555	0.009477
95%	0.010997	0.010809	0.00971
97.50%	0.01131	0.011074	0.010048
using stabilised K_{LX}			
Percentile	$d = 0.01$	$d = 0.05$	$d = 0.10$
50%	0.016194	0.013808	0.010612
90%	0.016793	0.014679	0.011686
95%	0.016959	0.014914	0.011924
97.50%	0.017082	0.015123	0.012362

Table 5.3: Percentiles for the test statistics calculated for point-to-line configurations for different values of d. 1000 point patterns were simulated for each case, each with 500 points on a unit square window.

the points were within 0.01 units of the linear network and the alternative was that the points were within d units, where d is given in the first column. At first glance it is clear that there is an improvement in the power, especially with the variance-stabilised function. The powers using the ordinary K_{LX} are still quite low, and this is consistent with what has been said in the literature. The variance-stabilised function outperforms the K_{LX} function under both alternatives, and for all values of n considered. For both functions and for each value of d under the alternative, the powers increase as the number of the points in the point pattern, n, increase. For values of n greater than 500 the test using variance-stabilised function correctly rejects the null hypothesis when the alternative is true for both values of d. One possible reason for the poor performance of the usual K_{LX} function here is illustrated through Figure 5.12. For the normal K_{LX}, the MAD value occurs at nearly the last value of r, regardless of the value of d used. It would be visually difficult to distinguish these. For the variance-stabilised version though the MAD value occurs at different values of r, perhaps an indication that the underlying processes are different.

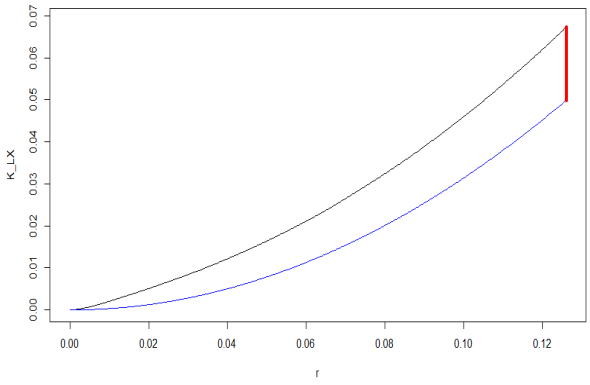
Table 5.5 shows the power values for the two-sided approach for the MAD test where the null hypothesis was that the points were within 0.05 units of the linear network and the alternative was that the points were within d units, where d is again given in the first column. For this null again, the MAD test based on the variance-stabilised function gives better results compared to the K_{LX}. Perhaps the concerning values are those for the alternative of $d = 0.10$ and $n = 100$ because they are relatively low. This problem is of course dealt with as n increases.

Finally, Table 5.6 shows the powers for the MAD test when the null hypothesis is $d = 0.10$ and the alternative is shown in the first column. Once again the concerning values are those for the case $n = 100$

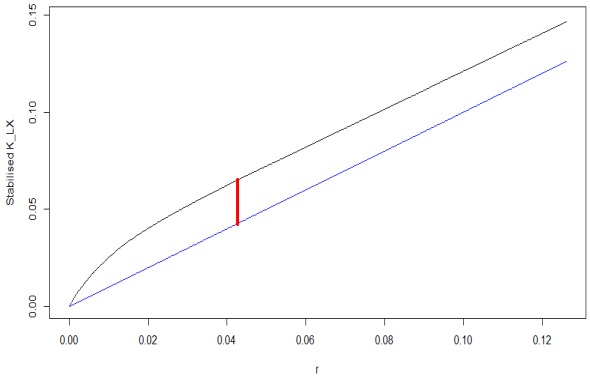
n	Value of d under alternative	α	Power using K_{LX}	Power using stabilised K_{LX}
100	0.05	0.01	0.031	0.916
		0.05	0.099	0.969
		0.10	0.167	0.978
	0.10	0.01	0.173	0.998
		0.05	0.280	1.000
		0.10	0.343	1.000
500	0.05	0.01	0.038	1.000
		0.05	0.141	1.000
		0.10	0.215	1.000
	0.10	0.01	0.176	1.000
		0.05	0.315	1.000
		0.10	0.389	1.000
1500	0.05	0.01	0.074	1.000
		0.05	0.225	1.000
		0.10	0.331	1.000
	0.10	0.01	0.304	1.000
		0.05	0.474	1.000
		0.10	0.565	1.000

Table 5.4: Power values for the two-sided MAD test when the null hypothesis is that the points are within 0.01 units of the linear network and the alternative is the points shifted by a value of d from the linear network.

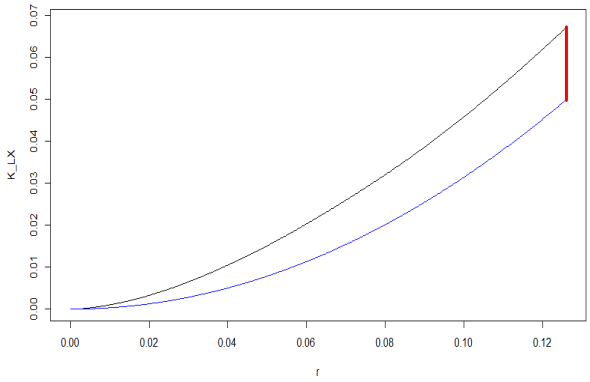
and $d = 0.05$, and similar to before the problem is alleviated as n increases. For all combinations of null and alternative hypotheses shown in Tables 5.4, 5.5 and 5.6, the test based on the variance stabilised function outperforms its counterpart, by quite a margin. The test is able to distinguish between different point and line configurations, and does so extremely well even at 100 points in some cases.



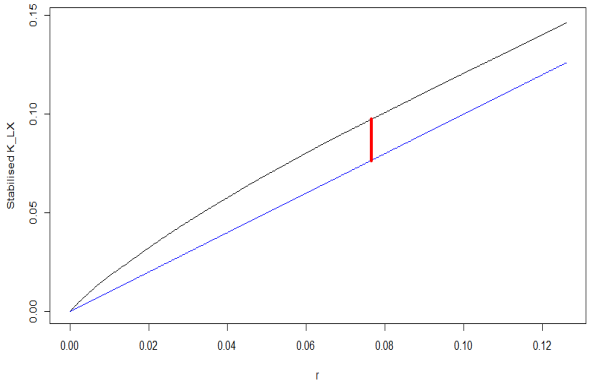
(a) $\hat{K}_{LX}$, d=0.01



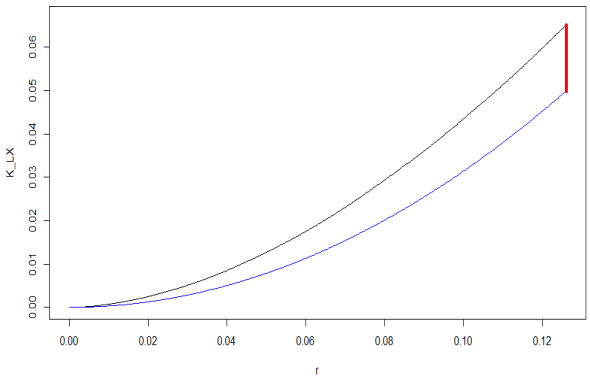
(b) Stabilised $\hat{K}_{LX}$, d=0.01



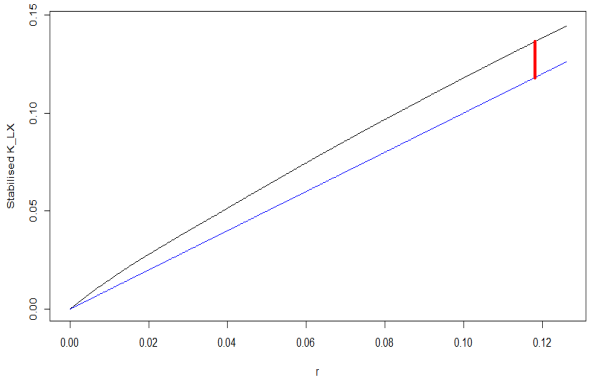
(c) $\hat{K}_{LX}$, d=0.05



(d) Stabilised $\hat{K}_{LX}$, d=0.05



(e) $\hat{K}_{LX}$, d=0.10



(f) Stabilised $\hat{K}_{LX}$, d=0.10

Figure 5.12: K_{LX} and stabilised K_{LX} estimates for different values of d. The point patterns used in each case had 1500 points and were on a unit square window. The curve in black is the estimate and the curve in blue is the estimate expected under CSR. The vertical red line indicates the point at which the estimate deviates the most from the independence curve, i.e. the MAD value.

n	Value of d under alternative	α	Power using K_{LX}	Power using stabilised K_{LX}
100	0.01	0.01	0.002	0.603
		0.05	0.012	0.940
		0.10	0.036	0.983
	0.10	0.01	0.064	0.416
		0.05	0.136	0.616
		0.10	0.204	0.700
500	0.01	0.01	0.002	1.000
		0.05	0.017	1.000
		0.10	0.042	1.000
	0.10	0.01	0.098	0.963
		0.05	0.225	0.994
		0.10	0.337	0.997
1500	0.01	0.01	0.006	1.000
		0.05	0.051	1.000
		0.10	0.122	1.000
	0.10	0.01	0.277	1.000
		0.05	0.531	1.000
		0.10	0.659	1.000

Table 5.5: Power values for the two-sided MAD test when the null hypothesis is that the points are within 0.05 units of the linear network and the alternative is the points shifted by a value of d from the linear network.

n	Value of d under alternative	α	Power using K_{LX}	Power using stabilised K_{LX}
100	0.01	0.01	0.006	0.984
		0.05	0.034	1.000
		0.10	0.070	1.000
	0.05	0.01	0.011	0.134
		0.05	0.059	0.406
		0.10	0.105	0.560
500	0.01	0.01	0.008	1.000
		0.05	0.064	1.000
		0.10	0.137	1.000
	0.05	0.01	0.057	0.915
		0.05	0.164	0.996
		0.10	0.270	0.999
1500	0.01	0.01	0.042	1.000
		0.05	0.164	1.000
		0.10	0.270	1.000
	0.05	0.01	0.249	1.000
		0.05	0.492	1.000
		0.10	0.607	1.000

Table 5.6: Power values for the two-sided MAD test when the null hypothesis is that the points are within 0.10 units of the linear network and the alternative is the points shifted by a value of d from the linear network.

5.4 Power of the second deviation test

In this section we consider the test presented in Section 4.5.3. The power of the test is now considered but slightly different to the approach used in the MAD test. The approach in [13] is used. Under the null hypothesis, two sets of point patterns are generated for a given value of d, giving a null sampling distribution as before. The critical values are determined from the distribution by considering the relevant percentiles. For the alternative, two distinct values of d are considered. The first value, d_1 is the value considered under the null hypothesis, and the second value is the value of d under the alternative i.e. d_2.

Table 5.7 shows the power values obtained for the test for $n = 100$, 500 and 1500 points on a unit square window where the linear network was simulated from a stationary Poisson line process. Starting with $n = 100$, the test performs particularly well for $d_1 = 0.01$ despite a small value of n. As one would expect, for $d_1 = 0.01$ and $d_2 = 0.10$, the power is high because of the relatively large difference between the two d values, so this is reflected in the number of times the null hypothesis is rejected given the alternative is true. For $d_1 = 0.05$ and $d_2 = 0.01$, the power values are not high, relative to the previous case. A possible reason being the values of d are close, so there is not much difference, at least for $n = 100$. For $n = 500$ and 1500, the test performs extremely well, as shown by the power. As the number of points in the point pattern increases, the power also increases. Table 5.8 shows the performance of the test when the linear network is generated from a non-stationary Poisson line process. The results for $n = 100$ are not particularly pleasing seeing that the power values are low. The power does increase as n increases however. The test also runs into similar problems for the cases where the values of d are not different enough, i.e. for $d_1 = 0.05$ and $d_2 = 0.10$, but again this problem diminishes as n increases. For $n = 1500$ the test has quite high power values as one would expect. This suggests that for large values of n, the non-stationarity of the Poisson line pattern does not hinder the performance of the best significantly. The test can in fact distinguish between K_{LX} estimates calculated from different values of d.

Finally, for completeness in Tables 5.9 and 5.10 we have the empirical and theoretical quantiles of the test statistics under the null hypothesis for the different values of d and n for comparison. These values are relatively close to each other which gives some confidence about the assumptions made about the distribution of the test statistics.

$n = 100$				$n = 500$				$n = 1500$			
d_1	d_2	α	Power	d_1	d_2	α	Power	d_1	d_2	α	Power
0.01	0.05	0.01	0.7926	0.01	0.05	0.01	0.9998	0.01	0.05	0.01	0.9998
0.01	0.05	0.05	0.9172	0.01	0.05	0.05	0.9998	0.01	0.05	0.05	0.9998
0.01	0.05	0.10	0.9558	0.01	0.05	0.10	0.9998	0.01	0.05	0.10	0.9998
0.01	0.10	0.01	0.9990	0.01	0.10	0.01	0.9998	0.01	0.10	0.01	0.9998
0.01	0.10	0.05	0.9998	0.01	0.10	0.05	0.9998	0.01	0.10	0.05	0.9998
0.01	0.10	0.10	0.9998	0.01	0.10	0.10	0.9998	0.01	0.10	0.10	0.9998
0.05	0.01	0.01	0.7494	0.05	0.01	0.01	0.9998	0.05	0.01	0.01	0.9998
0.05	0.01	0.05	0.9198	0.05	0.01	0.05	1.0000	0.05	0.01	0.05	0.9998
0.05	0.01	0.10	0.9580	0.05	0.01	0.10	1.0000	0.05	0.01	0.10	0.9998
0.05	0.10	0.01	0.4424	0.05	0.10	0.01	0.9960	0.05	0.10	0.01	0.9998
0.05	0.10	0.05	0.6806	0.05	0.10	0.05	0.9996	0.05	0.10	0.05	0.9998
0.05	0.10	0.10	0.7620	0.05	0.10	0.10	0.9996	0.05	0.10	0.10	0.9998
0.10	0.01	0.01	0.9990	0.10	0.01	0.01	0.9998	0.10	0.01	0.01	0.9998
0.10	0.01	0.05	0.9998	0.10	0.01	0.05	0.9998	0.10	0.01	0.05	0.9998
0.10	0.01	0.10	0.9998	0.10	0.01	0.10	0.9998	0.10	0.01	0.10	0.9998
0.10	0.05	0.01	0.4070	0.10	0.05	0.01	0.9940	0.10	0.05	0.01	0.9998
0.10	0.05	0.05	0.6540	0.10	0.05	0.05	0.9996	0.10	0.05	0.05	0.9998
0.10	0.05	0.10	0.7534	0.10	0.05	0.10	0.9996	0.10	0.05	0.10	0.9998

Table 5.7: Power values for the second deviation test when the null hypothesis is $d = d_1$ and the alternative is $d = d_2$, calculated for n=100, 500 and 1500 on a unit square window. The linear network was simulated using a stationary Poisson line process.

$n = 100$				$n = 500$				$n = 1500$			
d_1	d_1	α	Power	d_1	d_1	α	Power	d_1	d_1	α	Power
0.01	0.05	0.01	0.2238	0.01	0.05	0.01	0.8860	0.01	0.05	0.01	0.9998
0.01	0.05	0.05	0.4138	0.01	0.05	0.05	0.9702	0.01	0.05	0.05	0.9998
0.01	0.05	0.10	0.5278	0.01	0.05	0.10	0.9880	0.01	0.05	0.10	0.9998
0.01	0.10	0.01	0.5886	0.01	0.10	0.01	0.9996	0.01	0.10	0.01	0.9998
0.01	0.10	0.05	0.7792	0.01	0.10	0.05	0.9998	0.01	0.10	0.05	1.0000
0.01	0.10	0.10	0.8524	0.01	0.10	0.10	0.9998	0.01	0.10	0.10	1.0000
0.05	0.01	0.01	0.2010	0.05	0.01	0.01	0.8890	0.05	0.01	0.01	0.9998
0.05	0.01	0.05	0.3934	0.05	0.01	0.05	0.9692	0.05	0.01	0.05	0.9998
0.05	0.01	0.10	0.4988	0.05	0.01	0.10	0.9870	0.05	0.01	0.10	0.9998
0.05	0.10	0.01	0.0610	0.05	0.10	0.01	0.3354	0.05	0.10	0.01	0.9136
0.05	0.10	0.05	0.1818	0.05	0.10	0.05	0.5856	0.05	0.10	0.05	0.9772
0.05	0.10	0.10	0.2688	0.05	0.10	0.10	0.6998	0.05	0.10	0.10	0.9874
0.10	0.01	0.01	0.5204	0.10	0.01	0.01	0.9996	0.10	0.01	0.01	1.0000
0.10	0.01	0.05	0.7692	0.10	0.01	0.05	0.9996	0.10	0.01	0.05	1.0000
0.10	0.01	0.10	0.8472	0.10	0.01	0.10	0.9996	0.10	0.01	0.10	1.0000
0.10	0.05	0.01	0.0508	0.10	0.05	0.01	0.3424	0.10	0.05	0.01	0.8864
0.10	0.05	0.05	0.1764	0.10	0.05	0.05	0.5966	0.10	0.05	0.05	0.9674
0.10	0.05	0.10	0.2618	0.10	0.05	0.10	0.7164	0.10	0.05	0.10	0.9838

Table 5.8: Power values for the second deviation test when the null hypothesis is $d = d_1$ and the alternative is $d = d_2$, calculated for $n = 100$, 500 and 1500 on a unit square window. The linear network was simulated using a non-stationary Poisson line process.

Variance	d	n	Q	2.50%	5%	10%	25%	50%	75%	90%	95%	97.50%
1809.09	0.01	100	E	-82.52	-68.90	-53.38	-27.51	1.35	30.36	56.69	73.86	85.75
			T	-83.36	-69.96	-54.51	-28.69	0.00	28.69	54.51	69.96	83.36
1822.85	0.05	100	E	-86.24	-71.45	-55.45	-29.58	-0.06	29.08	55.99	69.42	83.40
			T	-83.68	-70.23	-54.72	-28.80	0.00	28.80	54.72	70.23	83.68
1868.18	0.10	100	E	-83.27	-71.30	-54.87	-30.35	-0.77	28.52	54.84	71.57	86.79
			T	-84.71	-71.09	-55.39	-29.15	0.00	29.15	55.39	71.09	84.71
1834.13	0.01	500	E	-85.24	-70.36	-55.25	-29.13	0.07	28.64	54.93	71.98	86.63
			T	-83.94	-70.44	-54.88	-28.89	0.00	28.89	54.88	70.44	83.94
1823.66	0.05	500	E	-86.57	-70.95	-53.90	-28.16	-0.85	27.97	54.44	69.07	81.62
			T	-83.70	-70.24	-54.73	-28.80	0.00	28.80	54.73	70.24	83.70
1882.53	0.10	500	E	-84.58	-72.31	-56.26	-30.38	0.34	29.37	55.78	70.84	84.52
			T	-85.04	-71.37	-55.60	-29.26	0.00	29.26	55.60	71.37	85.04
1817.76	0.01	1500	E	-87.84	-69.98	-53.75	-28.96	0.85	29.60	53.97	68.13	82.15
			T	-83.56	-70.13	-54.64	-28.76	0.00	28.76	54.64	70.13	83.56
1828.11	0.05	1500	E	-83.96	-70.00	-55.18	-28.83	0.75	29.28	54.73	69.00	82.50
			T	-83.80	-70.33	-54.79	-28.84	0.00	28.84	54.79	70.33	83.80
1873.81	0.10	1500	E	-86.03	-71.65	-56.43	-29.81	-0.02	29.28	55.72	71.02	83.90
			T	-84.84	-71.20	-55.48	-29.20	0.00	29.20	55.48	71.20	84.84

Table 5.9: The variance of the test statistic under different null hypothesis values of d and number of points n. The Quantiles (Q) are shown, E representing the empirical quantiles and T representing the theoretical quantiles. The linear network used was simulated from a stationary Poisson line process.

Var[H]	d	n	Q	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
1850.60	0.01	100	E	-86.54	-71.89	-55.83	-28.87	0.39	29.82	56.47	72.04	84.21
			T	-84.31	-70.76	-55.13	-29.02	0.00	29.02	55.13	70.76	84.31
1893.46	0.05	100	E	-84.85	-72.97	-56.92	-30.38	-1.12	27.80	55.14	72.01	85.57
			T	-85.29	-71.57	-55.77	-29.35	0.00	29.35	55.77	71.57	85.29
1950.37	0.10	100	E	-85.92	-72.76	-58.36	-30.61	-0.37	30.66	56.22	72.47	87.26
			T	-86.56	-72.64	-56.60	-29.79	0.00	29.79	56.60	72.64	86.56
1855.62	0.01	500	E	-85.05	-70.93	-55.92	-28.84	0.08	29.69	55.18	71.10	85.14
			T	-84.43	-70.86	-55.21	-29.05	0.00	29.05	55.21	70.86	84.43
1897.09	0.05	500	E	-83.34	-70.17	-54.91	-28.71	0.69	31.17	58.01	73.60	87.88
			T	-85.37	-71.64	-55.82	-29.38	0.00	29.38	55.82	71.64	85.37
1949.48	0.10	500	E	-86.20	-71.00	-55.24	-29.79	0.08	30.12	55.55	71.61	85.77
			T	-86.54	-72.63	-56.58	-29.78	0.00	29.78	56.58	72.63	86.54
1851.00	0.01	1500	E	-83.10	-70.30	-56.25	-29.45	-0.40	29.00	55.71	70.76	84.40
			T	-84.32	-70.77	-55.14	-29.02	0.00	29.02	55.14	70.77	84.32
1894.31	0.05	1500	E	-84.29	-71.77	-56.44	-29.43	-0.03	29.54	56.49	71.56	83.87
			T	-85.30	-71.59	-55.78	-29.36	0.00	29.36	55.78	71.59	85.30
1968.89	0.10	1500	E	-85.69	-73.53	-56.40	-28.85	1.35	30.15	57.58	73.66	85.86
			T	-86.97	-72.99	-56.87	-29.93	0.00	29.93	56.87	72.99	86.97

Table 5.10: The variance of the test statistic under different null hypothesis values of d and number of points n. The Quantiles (Q) are shown, E representing the empirical quantiles and T representing the theoretical quantiles. The linear network used was simulated from a non-stationary Poisson line process.

5.5 Conclusion

In this chapter we have considered simulations for various point-to-line configurations. We began with a graphical method where the K_{LX} estimates were plotted with confidence intervals. The function was able to pick up on attraction and repulsion for the cases where the linear network was a stationary Poisson line pattern and the case where either the orientation or perpendicular distance of the line segments were non-uniform i.e. non-stationary. The MAD test was investigated in Section 5.3 and as mentioned in the literature, the variance-stablised version of the K_{LX} estimate outperforms its counterpart. Finally, the second deviation test presented in Section 4.5.3 was investigated. The non-stationarity of the Poisson line process was of interest and had not been studied before in this context. The results from the simulation study suggest that the non-stationarity of the Poisson line process does not affect the ability of the function to detect attraction or repulsion of the points by the linear network. This of course agrees with the graphical methods in Sections 5.1 and 5.2 and the MAD test in Section 5.3. We now apply these methods to a crime dataset.

Chapter 6

Application

6.1 Data

In this chapter the graphical methods and hypothesis tests discussed so far are applied. The Gugulethu township located in the Western Cape province of South Africa is considered. The data analysed is the location of crimes in the area in relation to the roads[1]. The question at hand is if there is a positive association between the points and the linear network, that is, clustering of some sort about the roads. The crimes in the dataset are of 28 categories including robberies, carjacking and assault and so the crimes are analysed by category. Only a few of these categories are considered, with preference given to those with points larger than 200. This is in part because we have observed in Chapter 5 that the tests we shall apply perform better when there is a large number of points in the point pattern. These crimes were recorded from 2006 to 2016 but we limit analysis to 2 dimensions, leaving out the temporal aspect for future research.

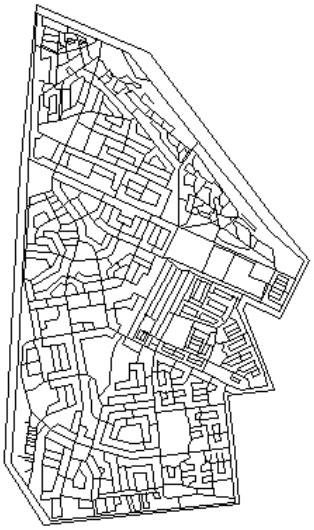
Figure 6.1 shows the Gugulethu area, crime locations and the linear network considered. The length of the linear network is 139 871.2m and the area of the window of observation is 8 487 $924m^2$. The dataset has 22 125 crime locations but only a subset of these are used. Figure 6.1(a) shows 16 074 of the crime locations we will analyse. Interestingly looking at Figure 6.1(a), even without the linear network imposed we can see points forming linear structures.

The following crime categories are considered:

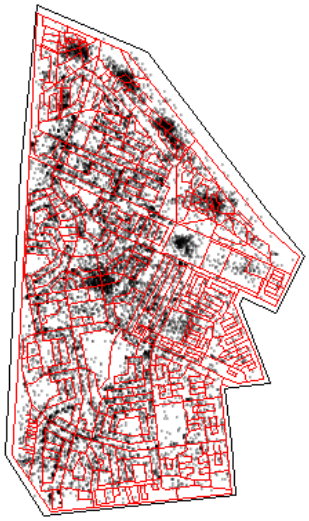
- Assault with the purpose to inflict grievous bodily harm, $n = 6497$
- Common assault, $n = 5543$

[1] Ethics approval: NAS404/2019

(a) Crime locations in the Gugulethu area



(b) The linear network of the Gugulethu area.



(c) Crime locations superimposed on the linear network.

Figure 6.1: The linear network of the Gugulethu area. The total length of roads is 139871.2m. The area of the observation widow is 8487924m^2. The total number of crime points is 16 074, and these were recorded from 2006 to 2016.

- Robbery with firearm, $n = 2037$
- Attempted murder, $n = 919$
- Carjacking, $n = 728$
- Business robbery, $n = 350$

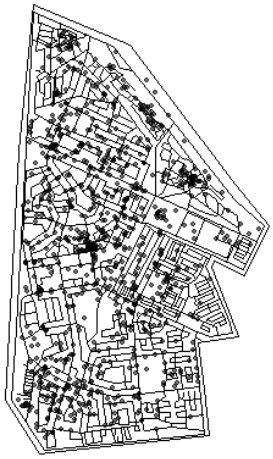
The crime locations specified above are shown in Figure 6.2. For each of these categories $\hat{K}_{LX}$ will be calculated along with confidence intervals and simulation envelopes. The MAD tests will also be applied,

but using the variance-stabilised $K_{LX}(r)$ for better performance. Then finally the second deviation test is applied. The p-values for the MAD test using the variance-stabilised estimate are shown in Table

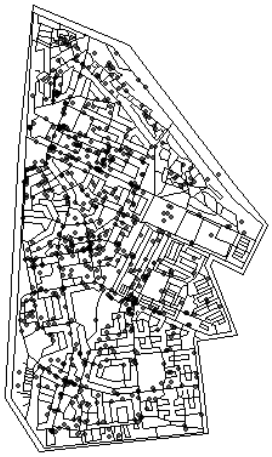
(a) Assault with the purpose To inflict grievous Bodily Harm

(b) Common assault

(c) Robbery with firearm



(d) Attempted murder



(e) Carjacking

(f) Business robbery

Figure 6.2: Locations of different types of crimes in the Gugulethu area.

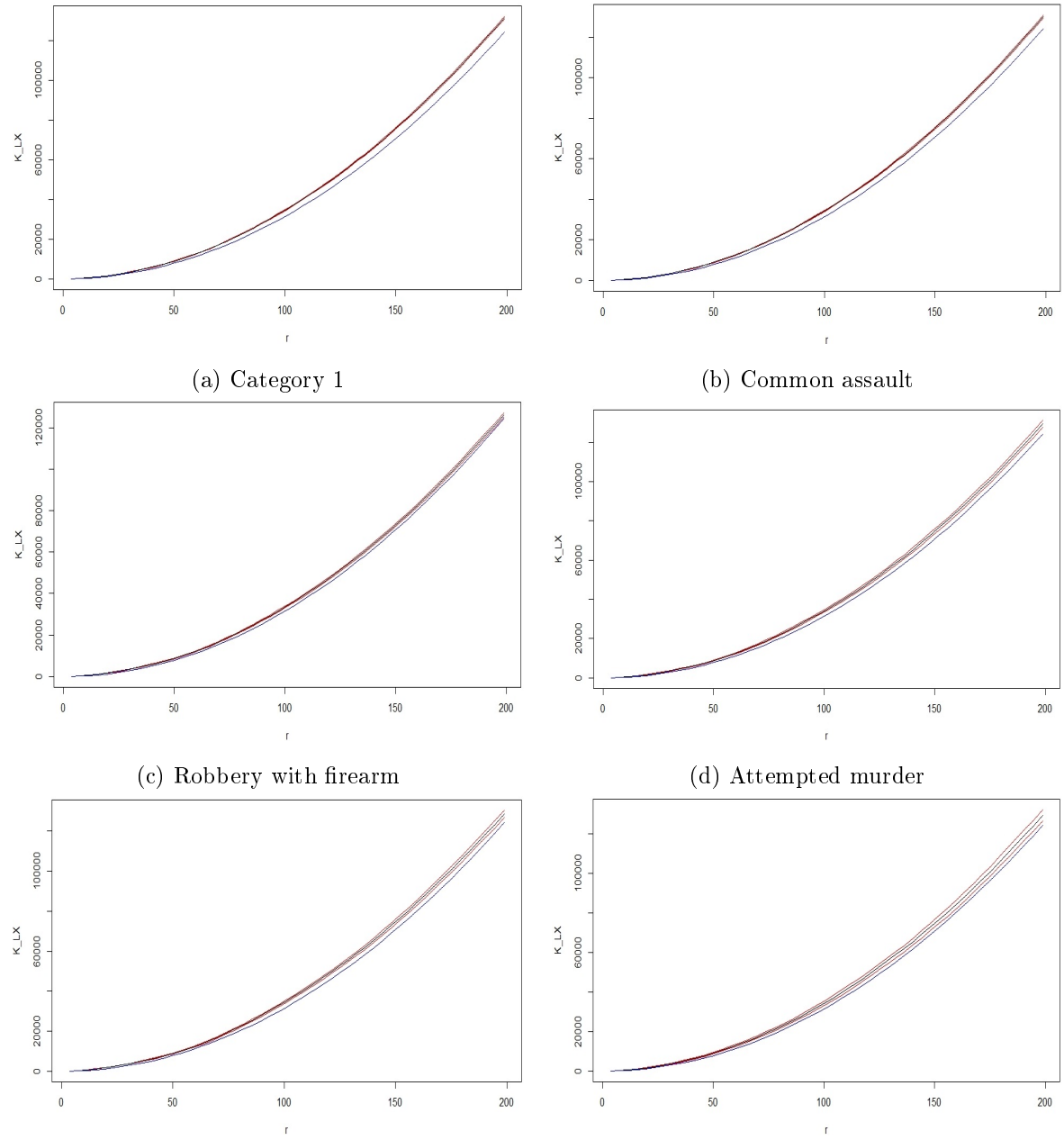
6.1. The p-values indicate that the points locations are significantly different from complete spatial randomness since they are all less than 0.10. Furthermore, Figure 6.3 shows the K_{LX} estimates for the different crime categories, with 95% confidence intervals (some of these are difficult to see because the confidence intervals are so thin). The K_{LX} estimates for all the crime categories lie above the expected curve under independence suggesting clustering of these crimes around the linear network. The curve under independence also lies outside these confidence intervals suggesting that the crime locations are significantly different from CSR. This agrees with the results of the MAD test.

Crime category	p-value for MAD test
Assault with the purpose to inflict grievous bodily harm	0.01
Common assault	0.01
Robbery with firearm	0.05
Attempted murder	0.04
Carjacking	0.04
Business robbery	0.01

Table 6.1: p-values obtained after applying the MAD test using the variance-stablised estimate. As can be seen all the p-values are less than or equal to $\alpha = 0.05$. The null hypothesis of complete spatial randomness can be rejected at a 5% level.

We now consider the second deviation test for the different crime categories and values of d under the null hypothesis. Tables 6.2-6.7 show the variance of the test statistics under H_0 for different values of d for the 6 crime categories. In addition, the theoretical and empirical quantiles are shown for comparison purposes. While these values do not match exactly, they are relatively close to each other. The main issue of course being the number of test statistics simulated under the null, which was 2000 in this case for each crime and value of d. If one were to increase these, we would expect the theoretical and empirical quantiles to be much close. This brings up a well known problem in the Monte Carlo setting, that is, the need of computational power as mentioned in [19] as well.

Table 6.8 shows the results of applying the test outlined in [13, 19]. The value of d under the null hypothesis is shown along with the test statistics calculated assuming this null hypothesis. The corresponding p-values are shown as well. First we observe that for values of $d = 50$ the null hypothesis is rejected for all the crime categories at a significance level of $\alpha = 0.10$. More clearly, we reject the null hypothesis that the crimes are within distance d of the linear network and that the attraction mechanism is appropriate. For the 'Business Robbery' category however we do not reject this null hypothesis at a significance level of 0.05. For the 'Robbery with firearm' crime category, the null hypothesis is rejected for all three values of d considered. This could be again because of the mechanism used to generate points that are within the specified distance of d is inappropriate for this crime category, or the value of d itself. This highlights



(a) Category 1

(b) Common assault

(c) Robbery with firearm

(d) Attempted murder

(e) Carjacking

(f) Business robbery

Figure 6.3: K_{LX} estimates for the different crime categories, the empirical estimate in black, the expected curve under complete spatial randomness in blue and 95% confidence intervals. (Note that category 1 is Assault with the purpose to inflict grievous bodily harm.)

the caveat of the test in that it is not clear which value of d to consider under the null hypothesis. The computational power available also has to be kept in mind when attempting to determine the values of d to use. The crime categories 'Assault with the purpose to inflict grievous bodily harm' and 'Common assault' reject the null hypothesis for $d = 150$ and $d = 100$ respectively. For the rest of the crime categories, the

d	var[H]	Quantiles	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
50	1562.06	Emp	-78.06	-64.84	-51.44	-26.34	0.59	26.52	52.51	64.70	77.27
		Theo	-77.46	-65.01	-50.65	-26.66	0.00	26.66	50.65	65.01	77.46
100	1678.51	Emp	-79.16	-66.42	-52.90	-27.08	0.93	27.76	54.14	69.62	82.85
		Theo	-80.30	-67.39	-52.50	-27.63	0.00	27.63	52.50	67.39	80.30
150	1777.68	Emp	-81.40	-68.69	-54.82	-27.04	0.65	29.17	53.76	69.45	85.94
		Theo	-82.64	-69.35	-54.03	-28.44	0.00	28.44	54.03	69.35	82.64

Table 6.2: Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Assault with the purpose to inflict grievous bodily harm' crime category for different values of d under H_0.

d	var[H]	Quantiles	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
50	1500.75	Emp	-76.35	-63.35	-50.26	-26.79	-1.02	24.70	48.88	61.68	74.76
		Theo	-75.93	-63.72	-49.65	-26.13	0.00	26.13	49.65	63.72	75.93
100	1775.19	Emp	-80.17	-67.85	-52.01	-25.92	1.30	27.88	51.91	66.74	80.44
		Theo	-82.58	-69.30	-54.00	-28.42	0.00	28.42	54.00	69.30	82.58
150	1819.4	Emp	-84.53	-71.19	-56.32	-29.40	-1.31	29.09	54.53	71.14	81.13
		Theo	-83.60	-70.16	-54.66	-28.77	0.00	28.77	54.66	70.16	83.60

Table 6.3: Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Common assault' crime category for different values of d under H_0.

d	var[H]	Quantiles	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
50	1588.45	Emp	-78.22	-65.25	-49.67	-26.70	-0.17	25.86	53.30	65.43	77.02
		Theo	-78.12	-65.56	-51.08	-26.88	0.00	26.88	51.08	65.56	78.12
100	1703.97	Emp	-82.23	-69.54	-53.30	-30.40	-1.82	26.67	53.88	67.02	81.22
		Theo	-80.91	-67.90	-52.90	-27.84	0.00	27.84	52.90	67.90	80.91
150	1762.21	Emp	-79.79	-69.27	-51.61	-27.08	1.41	29.41	54.37	68.42	79.19
		Theo	-82.28	-69.05	-53.80	-28.31	0.00	28.31	53.80	69.05	82.28

Table 6.4: Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Robbery with firearm' crime category for different values of d under H_0.

mechanism of clustering and the values of d around 100 or 150 seem to be appropriate since the p values are all greater than $\alpha = 0.10$.

Figures 6.4 and 6.5 show the K_{LX} estimates for the different crime categories, with simulation envelopes for values of d of interest. These figures show that particularly for $d = 50$ the attraction mechanism and the distance is not appropriate for the point locations for the different crimes. It appears that the

d	var[H]	Quantiles	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
50	1561.17	Emp	-77.15	-65.28	-51.63	-27.40	-2.18	24.24	50.20	65.73	76.87
		Theo	-77.44	-64.99	-50.64	-26.65	0.00	26.65	50.64	64.99	77.44
100	1706.32	Emp	-80.68	-69.42	-53.17	-27.70	0.00	29.11	54.37	69.04	81.48
		Theo	-80.96	-67.94	-52.94	-27.86	0.00	27.86	52.94	67.94	80.96
150	1741.95	Emp	-81.54	-66.32	-52.79	-26.64	0.73	28.08	54.98	68.23	79.67
		Theo	-81.80	-68.65	-53.49	-28.15	0.00	28.15	53.49	68.65	81.80

Table 6.5: Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Attempted murder' crime category for different values of d under H_0.

d	var[H]	Quantiles	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
50	1524.65	Emp	-75.81	-64.02	-50.98	-25.55	0.75	26.69	47.60	60.55	73.41
		Theo	-76.53	-64.23	-50.04	-26.34	0.00	26.34	50.04	64.23	76.53
100	1676.26	Emp	-79.89	-67.49	-52.87	-28.41	0.82	27.37	50.64	66.42	79.56
		Theo	-80.25	-67.34	-52.47	-27.62	0.00	27.62	52.47	67.34	80.25
150	1772.45	Emp	-84.57	-71.70	-52.15	-27.92	0.03	29.40	52.46	69.64	82.46
		Theo	-82.52	-69.25	-53.95	-28.40	0.00	28.40	53.95	69.25	82.52

Table 6.6: Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Car jacking' crime category for different values of d under H_0.

d	var[H]	Quantiles	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
50	1559.81	Emp	-75.94	-63.42	-49.26	-23.86	1.91	27.19	50.97	63.88	79.29
		Theo	-77.41	-64.96	-50.61	-26.64	0.00	26.64	50.61	64.96	77.41
100	1677.47	Emp	-79.48	-67.85	-51.87	-28.30	-2.29	27.09	53.72	67.98	77.65
		Theo	-80.27	-67.37	-52.49	-27.63	0.00	27.63	52.49	67.37	80.27
150	1750.06	Emp	-80.21	-69.65	-56.26	-26.74	2.24	28.94	53.64	68.30	82.77
		Theo	-81.99	-68.81	-53.61	-28.22	0.00	28.22	53.61	68.81	81.99

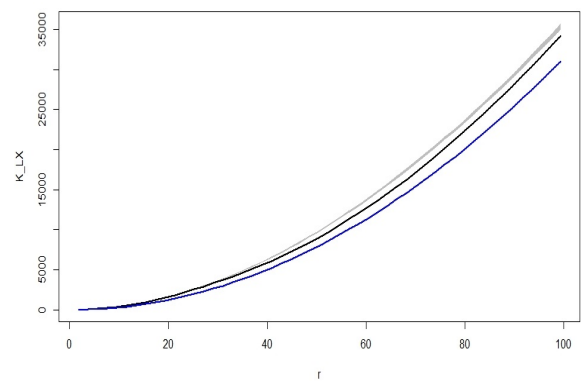
Table 6.7: Empirical (Emp) and theoretical (Theo) quantiles for the test statistic for the 'Business robbery' crime category for different values of d under H_0.

attraction mechanism is too strong, evidenced by the simulation envelopes lying above the K_{LX} estimates and the confidence intervals. The exception is the 'Business robbery' crime category, with the estimate lying within the simulation envelopes. The K_{LX} estimates for the crime categories for which the null hypothesis for the value of d was not rejected are shown as well. A common feature among these graphs is that the K_{LX} veers outside the simulation envelopes even though the null hypothesis was not rejected for the particular value of d. This is perhaps because of the difference in the two methods. The graphical

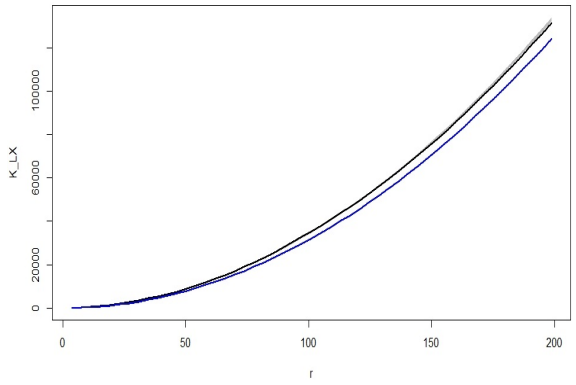
Crime	d	H	p-value
Assault with the purpose to inflict grievous bodily harm	50	-289.54	<0.001
Assault with the purpose to inflict grievous bodily harm	100	-14.64	0.3604
Assault with the purpose to inflict grievous bodily harm	150	82.97	0.0245
Common assault	50	334.67	<0.001
Common assault	100	-92.74	0.0118
Common assault	150	-5.53	0.4478
Robbery with firearm	50	-185.65	<0.001
Robbery with firearm	100	-94.12	0.0113
Robbery with firearm	150	-92.07	0.0141
Attempted murder	50	-110.88	0.0025
Attempted murder	100	-26.12	0.2636
Attempted murder	150	-3.28	0.4687
Carjacking	50	-66.64	0.0439
Carjacking	100	10.01	0.4035
Carjacking	150	20.44	0.3136
Business robbery	50	-59.09	0.0673
Business robbery	100	-2.09	0.4797
Business robbery	150	5.00	0.4524

Table 6.8: Test statistics obtained for the different crime categories assuming the value of d shown under H_0 and the corresponding p-value.

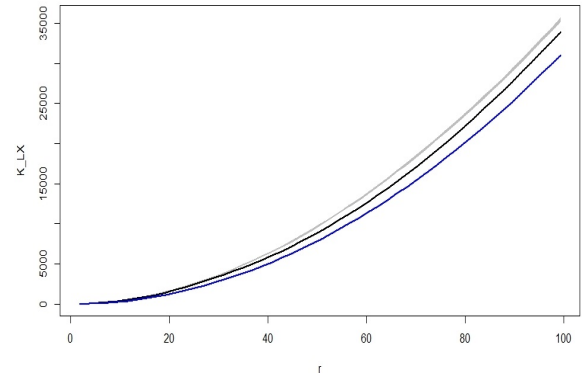
method involves simulation of point patterns under a null hypothesis assuming a given value of d and calculating the K_{LX} estimates, sorting by the largest and smallest and plotting the simulation envelopes. The method involving the test statistic also involves simulating points under the null hypothesis and using equation (4.26). Perhaps the second of the two is not that sensitive to deviations from the null hypothesis. Furthermore, the discussion in [5] concerning pointwise envelopes argues that we cannot attach a significance level of say 0.05 when the estimate veers out of the simulation envelopes. The fact that it lies outside the simulation envelope for some values of r indicates what the conclusion would have been if we had chosen to test a null hypothesis specific to that value of r prior to plotting the graphs.



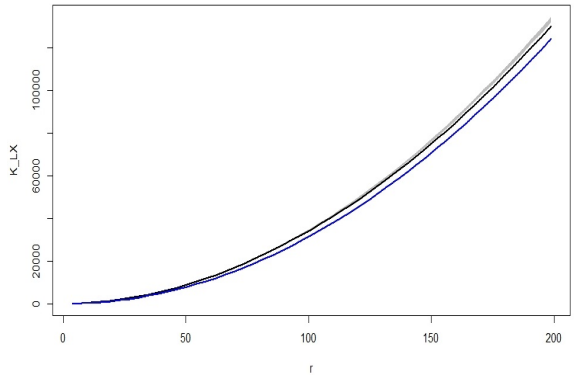
(a) Category 1, $d = 50$



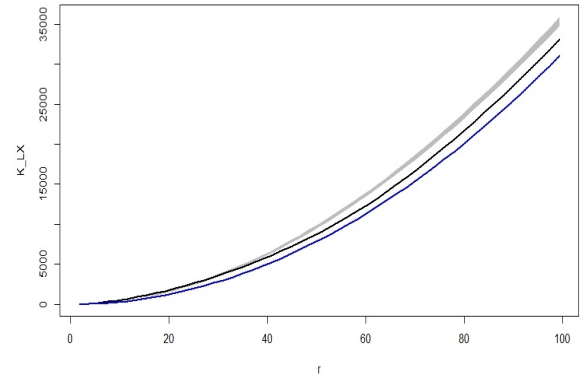
(b) Category 1, $d = 150$



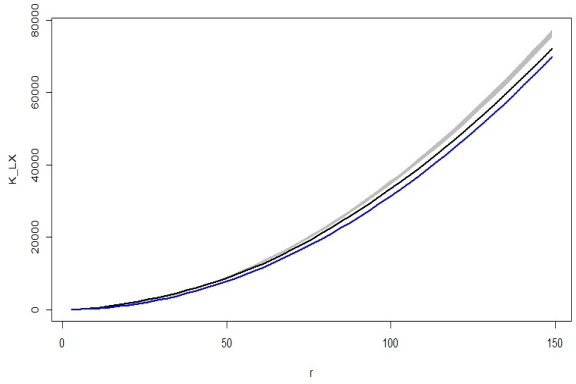
(c) Common assault, $d = 50$



(d) Common assault, $d = 100$

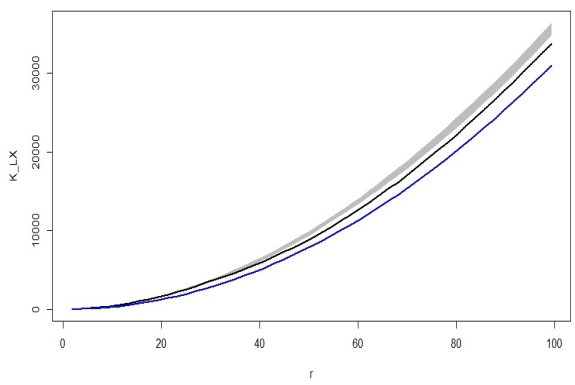


(e) Robbery with firearm, $d = 50$

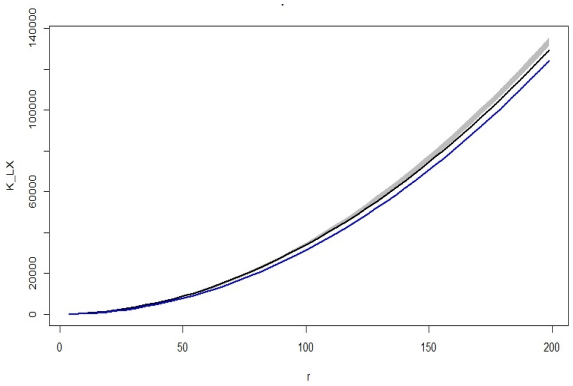


(f) Robbery with firearm, $d = 100$

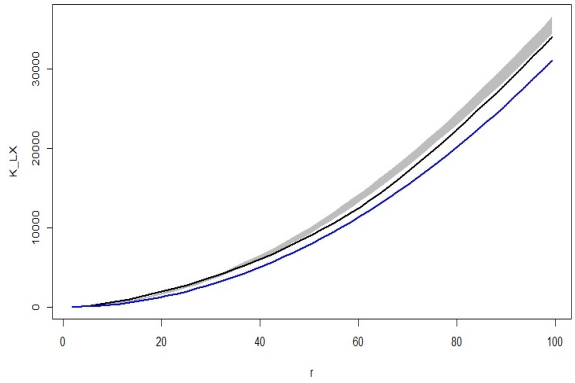
Figure 6.4: K_{LX} estimates for the different crime categories, the empirical estimate in black, the expected curve under complete spatial randomness in blue and simulation envelopes for values of d of interest in shaded grey. (Note that category 1 is Assault with the purpose to inflict grievous bodily harm.)



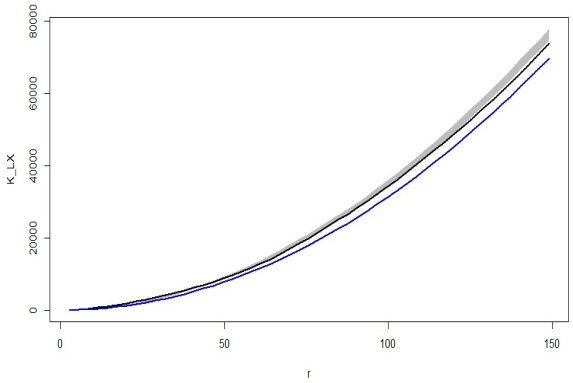
(a) Attempted murder, $d = 50$



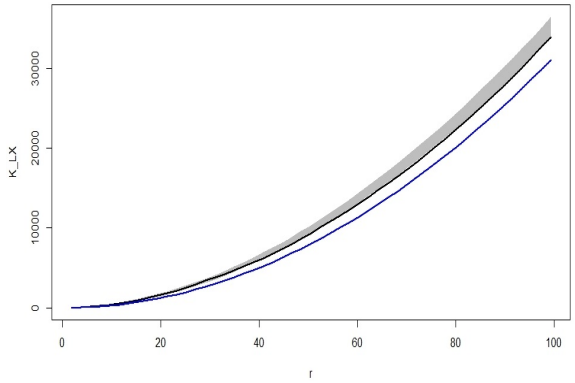
(b) Attempted murder, $d = 150$



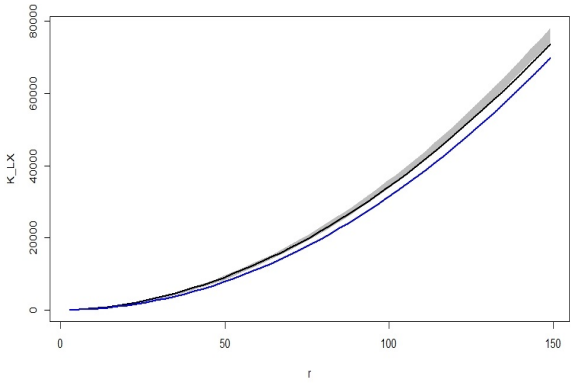
(c) Car jacking, $d = 50$



(d) Car jacking, $d = 100$



(e) Business robbery, $d = 50$



(f) Business robbery, $d = 100$

Figure 6.5: K_{LX} estimates for the different crime categories, the empirical estimate in black, the expected curve under complete spatial randomness in blue and simulation envelopes for values of d of interest in shaded grey.

6.2 Conclusion

In this chapter we applied the methods discussed in Chapter 4, particularly Sections 4.4 and 4.5, to a real dataset. The MAD tests for each of the crime categories considered indicated that the crime locations were significantly different from complete spatial randomness. The next step was then to determine the appropriate value of d for the attraction mechanism proposed. For all the crime categories, except for 'Robbery with firearm', the attraction mechanism was found to be valid for $d = 100$. For the 'Robbery with firearm' category the null hypothesis was rejected for all the values of d considered which could be because of the value of d chosen or the attraction mechanism. The simulation envelopes were also computed for the different crime categories for two values of d. While the estimates did not lie entirely in the simulation envelopes for the values of d which were not rejected, the estimates were closer to the simulation envelopes than for the values of d rejected. This confirmed that out of the values of d considered, the values not rejected under H_0 in Table 6.8 were most appropriate.

Chapter 7

Conclusion

In this mini-dissertation we have

- discussed the K- and cross-K functions for unmarked and marked point patterns respectively,
- discussed the extension to these proposed in [13] along with confidence intervals for the estimate,
- discussed Monte Carlo methods including simulation envelopes and the maximum absolute deviation test considered in [5] and the deviation test considered in [19],
- investigated the powers of the tests considered including the case where the non-stationary Poisson line process was used to simulate a linear network,
- applied the methodology to a real dataset successfully, showing that roads could be modelled as crime generators.

Non-stationary line patterns, to our knowledge, had not been discussed in this context before. The simulations involving non-stationary line patterns suggest the non-stationarity of the lines does not affect the ability of the function defined in equation (4.17) to pick up the attraction or repulsion if any. So if non-stationarity of linear networks is defined as non-stationarity of the line pattern that is made up of line segments, then this non-stationarity does not influence the performance of the function. The same was suggested in the original paper in [13] where they claimed that because we are only concerned with the point to line distances not point to point or line to line distances the 'internal structure' of the linear network does not affect the performance of the function. This has been confirmed herein.

The Poisson line process is a 'random' collection of lines as defined in [10]. It is clear then that this may not be the best way to model linear networks since these are unlikely to be random. These arise as needed

by the inhabitants of the area and far more complex than a Poisson line process and serve to fulfill some function like transportation. As mentioned in [7] there are other things to consider like over-head roads and tunnels when discussing the validity of planarity in street networks. One could also draw from graph theory to model road networks as discussed in [6].

One could argue that the clustering mechanism used to generate attraction was simplistic so we could attempt a more complex clustering mechanism to better capture the distribution of the points. This approach could also be valid because as was shown, the K_{LX} estimates did not lie entirely in the simulation envelopes of K_{LX} estimates calculated on point patterns assuming the value of d not rejected by the tests shown in Table 6.8. To gain more information about the point to line relationship we could consider a myriad of other spatial descriptive summary functions such as the pair correlation function, and the G and F functions.

As mentioned before, using just this approach it is difficult to select the value of d to consider under the null hypothesis. This and the computational power required for Monte Carlo requires the researcher to think carefully about the setting of their study. At the end for our purposes when we chose the values we had in mind distances that human beings are likely to travel so we limited our range to 150 metres. Of course the researcher has to keep in mind the edge effects issue because some edge effects are valid up to a maximum value of r, like the isotropic edge correction we used. This could also be circumvented by fitting a point process model to the data.

Minimum contrast estimation, [5], could be performed for both this simplistic model and for the more complex clustering mechanism. As described in [5], the aim in this approach is to minimise the sum of squared errors between the K_{LX} estimate of theoretical model and that of our data. This is similar to the test statistic used in Section 4.5.3, equations (4.24)-(4.26), where the idea was to sum the deviations of the K_{LX} estimate from the proposed theoretical model. However, if the theoretical K_{LX} function is unknown, as is the case in the simplistic model, one would have to estimate it through Monte Carlo, which again brings the problem of computational power. The test described in Section 4.5.3 involved 2000 test statistics under H_0. So to achieve this one had to simulate 4000 point patterns and hence calculate 4000 K_{LX} estimates. However it may be less computationally expensive if the minimum contrast estimation is used. One would have to estimate the theoretical K_{LX} model for the three values of d, but 99 simulations would be sufficient and the average of these would be taken as the estimate. Having got the three estimated K_{LX} functions for the three values of d, we could pick the value of d that minimises the squared differences. Then finally use this value in the test as described in Section 4.5.3. In this case we would only need to perform one set of simulations for to get an empirical distribution (instead of 3 corresponding to the three values of d) to confirm our findings from the minimum contrast estimation results.

A spatio-temporal analysis could also be conducted to investigate how the crime locations change over

the years that the data is available. This would be similar to the question raised in [18] about locations of farms that tested positive for bovine tuberculosis in the period 1989 to 2002 in Cornwall, United Kingdom. A test of similarity between the point patterns over the years could be conducted as well as suggested in [1] to again determine if and how the crime locations differ over time. Furthermore, in township areas such as Gugulethu, we could further investigate the differences between formal roads and informal roads and pathways and determine if there is more clustering around the one type of road versus the other.

We end by mentioning that it is clear that there may be other factors that influence the locations of crimes as this is likely to be a multi-dimensional problem. Therefore future analyses could consider other factors that could influence crime locations such as the types of businesses that operate in the area. Clearly with the growth of the field of spatial statistics, there are multiple analyses that could be considered given the researcher's interest.

Bibliography

[1] Martin A Andresen. An area-based nonparametric spatial point pattern test: The test, its applications, and the future. *Methodological Innovations*, 9:1–11, 2016.

[2] Qi Wei Ang, Adrian Baddeley, and Gopalan Nair. Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4):591–617, 2012.

[3] S Bryn Austin, Steven J Melly, Brisa N Sanchez, Aarti Patel, Stephen Buka, and Steven L Gortmaker. Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments. *American Journal of Public Health*, 95(9):1575–1581, 2005.

[4] Adrian Baddeley, Peter J Diggle, Andrew Hardegen, Thomas Lawrence, Robin K Milne, and Gopalan Nair. On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, 84(3):477–489, 2014.

[5] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. CRC Press, 2015.

[6] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[7] Geoff Boeing. Planarity and street network representation in urban form analysis. *Environment and Planning B: Urban Analytics and City Science*, 47(5):855–869, 2020.

[8] Gregory D Breetzke, Inger Fabris-Rotelli, Jacob Modiba, and Ian S Edelstein. The proximity of sexual violence to schools: evidence from a township in south africa. *GeoJournal*, pages 1–12, 2019.

[9] Vishnu Vardhan Chetlur and Harpreet S Dhillon. Coverage analysis of a vehicular network modeled as Cox process driven by Poisson line process. *IEEE Transactions on Wireless Communications*, 17(7):4401–4416, 2018.

[10] Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic Geometry and its Applications*. John Wiley & Sons, 2013.

[11] Sung Nok Chiu and Ling Wang. Homogeneity tests for several poisson populations. *Computational Statistics & Data Analysis*, 53(12):4266–4278, 2009.

[12] Philip J Clark and Francis C Evans. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453, 1954.

[13] C Comas, S Costafreda-Aumedes, N López, and C Vega-Garcia. On the correlation structure between point patterns and linear networks. *Spatial Statistics*, 29:192–203, 2019.

[14] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 2015.

[15] Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes, Volume 1: Elementary Theory and Methods*. Verlag New York Berlin Heidelberg: Springer, 2003.

[16] Peter L Day and Jamie Pearce. Obesity-promoting food environments and the spatial clustering of food outlets around schools. *American Journal of Preventive Medicine*, 40(2):113–121, 2011.

[17] Peter J Diggle. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, pages 87–101, 1979.

[18] Peter J Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press, 2013.

[19] Peter J Diggle and Amanda G Chetwynd. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, pages 1155–1163, 1991.

[20] Philip M Dixon. Ripley's K function. *Encyclopedia of Environmetrics*, 2002.

[21] Philip M Dixon. Nearest neighbor methods: Overview with examples. *Wiley StatsRef: Statistics Reference Online*, 2014.

[22] Peter Haase. Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction. *Journal of Vegetation Science*, 6(4):575–582, 1995.

[23] J Bryan Kinney, Patricia L Brantingham, Kathryn Wuschke, Michael G Kirk, and Paul J Brantingham. Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment*, 34(1):62–74, 2008.

[24] John Kornak, Mark E Irwin, and Noel Cressie. Spatial point process models of defensive strategies: detecting changes. *Statistical Inference for Stochastic Processes*, 9(1):31–46, 2006.

[25] Christine Kraamwinkel. Describing heterogeneity in spatial point processes. Master's thesis, University of Pretoria, 2017.

[26] Christine Kraamwinkel, Inger Fabris-Rotelli, and Alfred Stein. Bootstrap testing for first-order stationarity on irregular windows in spatial point patterns. *Spatial Statistics*, 28:194–215, 2018.

[27] Naa Oyo A Kwate and Ji Meng Loh. Separate and unequal: the influence of neighborhood and school characteristics on spatial proximity between fast food and schools. *Preventive Medicine*, 51(2):153–156, 2010.

[28] Ji Meng Loh. A valid and fast spatial bootstrap for correlation functions. *The Astrophysical Journal*, 681(1):726, 2008.

[29] HW Lotwick and BW Silverman. Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 406–413, 1982.

[30] Atsuyuki Okabe and Ikuho Yamada. The K-function method on a network and its computational implementation. *Geographical Analysis*, 33(3):271–290, 2001.

[31] Richard F Potthoff and Maurice Whittinghill. Testing for homogeneity: II. The Poisson distribution. *Biometrika*, 53(1/2):183–190, 1966.

[32] Brian D Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212, 1977.

[33] S Samuel and Richard C Larson. Bertrand's paradox revisited: more lessons about that ambiguous word, random. *Journal of Industrial and Systems Engineering (JISE)*, 2009.

[34] Shino Shiode. Analysis of a distribution of point events using the network-based quadrat method. *Geographical Analysis*, 40(4):380–400, 2008.

[35] Peter G Spooner, Ian D Lunt, Atsuyuki Okabe, and Shino Shiode. Spatial analysis of roadside acacia populations on a road network using the network K-function. *Landscape Ecology*, 19(5):491–499, 2004.

[36] Ikuho Yamada and Jean-Claude Thill. Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2):149–158, 2004.