Spatial point pattern analysis with application to confocal microscopy data

by

Philip Pembe Umande
Department of Mathematics

A thesis submitted for the degree of *Doctor of Philosophy of the University of London* and for the *Diploma of Imperial College*

Department of Mathematics
Section of Statistics
Imperial College London
University Of London
180 Queen's Gate, London SW7 2BZ.
England

December 2007

RSL LONDON UNIV.

IMPERIAL COLLEGE LONDON

Spatial point pattern analysis with application to confocal microscopy data

Philip Pembe Umande

Abstract

Widespread interest in the nuclear body Promyelocytic leukemia (PML) has emerged because of its link to several human disorders, including acute Promyelocytic leukemia and AIDS (Borden, 2002). After numerous and wide-ranging studies, the functions of PML-body components remains elusive (Hodges et al, 1998). Advances in microscopy have made it possible to capture images of PML and other nuclear bodies. Using recently developed imaging software, it is also possible to obtain coordinates of the centroid of these nuclear bodies from their microscopy images. In this thesis we focus on the statistical analysis of datasets comprising of three-dimensional PML centroid point coordinates which have been obtained in such a manner. We treat the data as a spatial point pattern so that we are able to carry out analysis within a spatial point process framework.

As is the generally adopted approach in point pattern exploration, we begin our analysis by testing for Complete Spatial Randomness (CSR). CSR is the "white noise" of spatial point processes, and although unattainable in practice, the findings from the test can be highly informative. Results of CSR tests on the PML-body centroid locations reveal that, the level of evidence for rejecting the CSR null hypothesis varies across cell nuclei. Furthermore, the tests disclose some interesting relationships between PML bodies and their distance to the boundary.

We adopt distance-based functions for carrying out CSR tests and other spatial point process inference procedures. These functions, or 'characteristics' are used as descriptive tools for spatial point patterns and they include the commonly used K-function. We introduce a new point process characteristic, the Q-function, which has advantages over the K-function. The closed form expression for the Q-function of some common point processes including the homogeneous Poisson process and Gauss-Poisson process is provided for the two- and three-dimensional cases.

Estimation of distance-based characteristics in a bounded sampling window introduces estimation bias resulting from not accounting for unobserved objects lying outside of the sampling region. This can provide misleading results when comparing the theoretical form of a distance-based characteristic with that estimated from spatial point process data. Edge-correction techniques are methods for dealing with this issue. By applying an edge-correction, we can reduce estimation bias and variance. We introduce two new edge-corrections and demonstrate through simulation studies that in some cases, they have a lower bias and variance compared to some of the currently used edge-correction methods.

CSR tests on datasets that provide details of the nuclear boundary in the form of a convex hull provide evidence for rejecting the CSR null hypothesis. As a result, we attempt to fit a more suitable model to the data. Using current ideas regarding nuclear architecture and results from the CSR tests, we consider several competing alternative models for the spatial distribution of PML bodies. These include an inhomogeneous Poisson process, log-Gaussian Cox process and Markov point process. Various model-fitting routines are explored, including those that utilise mininmum contrast techniques, pseudo maximum likelihood optimisation, and Markov Chain Monte Carlo (MCMC) optimisation.

Acknowledgement

I am forever grateful towards Dr. Niall Adams and Dr. David Stephens for their support throughout the whole of this research. I would like to thank my mother, Cecile Umande for her care and support throughout all of my academic career. I am also very thankful towards my fiancée Kathryn Cooke for Figures 3.1 and 4.1 and her limitless patience and support throughout this research. I would like to express my gratitude towards EPSRC for funding this research.

I wish to also thank the Imperial College London centre for structural biology. In particular, I wish to express my gratitude towards Paul Freemont for allowing the reproduction of the cell images in this thesis. Also, Carol Shiels, Richard Russell and Elizabeth Batty for their role in producing the datasets and providing some insight into the imaging process and general cell biology theory.

Contents

1 Introduction **13**

1.1 Overview 13

1.2 Spatial point processes and point patterns 16

1.3 Nuclear architecture and PML nuclear bodies 18

1.4 Producing spatial point patterns by use of confocal microscopy 20

1.5 Initial tests for CSR on PPDS1 21

2 Spatial point processes: Theory, models and statistics **26**

2.1 Spatial point process definition and characterisation 27

2.2 Moments and moment measures 29

2.3 Palm distributions 30

2.4 The Binomial point process 31

2.5 Homogeneous Poisson point process 33

2.6 Inhomogeneous Poisson point process 33

2.7 Estimating intensity 34

2.8 Edge effects 35

2.9 The empty space function $F(r)$ 36

2.10 The nearest neighbour distribution function $G(r)$ 37

2.11 The pair correlation function $g(r)$ 38

2.12 The $K(r)$-function 39

2.13 Relationships between spatial point process functions 39

2.14 Third-order characteristics 40

2.15 More complicated point process models 41

2.15.1 Gauss-Poisson process 42

2.15.2 Matérn cluster process 42

2.15.3 Markov point processes 43

2.15.4 Cox processes 44

3 The Q-function **45**

3.1 The Q-function 46

3.2 The Q-function of some point processes 48

3.2.1 The Q-function for a spatial homogeneous Poisson process 48
3.2.2 The Gauss-Poisson process Q-function 52
3.2.3 The Q-function of a p-thinned process 53
3.3 Estimating Q 53
3.4 Practical applications 54
3.4.1 Comparing the Q-function to other characteristics 60
3.4.2 Q-function based CSR tests on PPDS1 64

4 Edge-correcting in three-dimensions **68**
4.1 Initial illustrations 69
4.2 Comparing the severity of edge effects in three-dimensions with two-dimensions 77
4.3 Techniques for edge-correcting 88
4.3.1 Edge-corrections via weighting factors 89
4.3.2 Kaplan-Meier estimation 93
4.3.3 Minus sampling (the "border" method) 93
4.3.4 Plus sampling 94
4.3.5 Probabilistic edge-correction 95
4.3.6 The probabilistic border edge-correction 100
4.4 Edge-correction for irregular windows 102
4.5 Occurrences where edge-correcting may not be advantageous 103

5 Modelling replicated PML confocal microscopy data **105**
5.1 Testing for similarity 107
5.2 Model 1: Stationary Poisson process with exclusion zones 111
5.2.1 Model 1 characteristics 111
5.2.2 Fitting Model 1 111
5.2.3 Model 1 diagnostics 112
5.3 Model 2: Boundary effects model 116
5.3.1 Simulation and theoretical properties of Model 2 120
5.3.2 Fitting Model 2: parameter estimation and diagnostics 121
5.3.3 Model 2 diagnostic tests on PPDS2 127
5.4 Model 3: Boundary effects and PML NB interaction 134

5.4.1 Estimating the parameters of Model 3 using MLE 136
5.4.2 Perfect simulation of Markov models through coupling from the past (CFTP) 138
5.4.3 Estimating the parameters of Model 3 via path sampling 140
5.5 A log-Gaussian Cox process 145
5.6 Model selection 149
5.6.1 Model selection through pattern recognition methods 150
5.6.2 Model selection via partial Bayes factors 150

6 Marked spatial point pattern analysis **153**
6.1 Some preliminaries for marked spatial point processes 155
6.1.1 Definitions 155
6.1.2 Stationarity and isotropy 156
6.1.3 Moment measures and summary statistics for marked spatial point patterns 157
6.2 Bivariate spatial point process analysis of PPDS3 158
6.3 Marked spatial point process analysis of PPDS2 160
6.4 A marked inhomogeneous Poisson process model for PPDS2 161

7 Conclusions **167**

A Basic overview on measure theory, topological and metric spaces **173**

B Datasets PPDS1, PPDS2 and PPDS3 **177**
B.1 Distance units for the data 177
B.2 PPDS1 177
B.3 PPDS2 177
B.4 PPDS3 178
B.5 Data simulated using perfect simulation 178

C Look up Tables **182**

D Derivation of the Q-function for a Gauss-Poisson process **183**

E $BIAS$ curves and simulation envelopes **185**

List of Tables

3.1 Values of σ for Japanese black pine, Redwood seedlings and Cell centres . 59

4.1 Experimental results for applying a weighting factor edge-correction . 91
4.2 Results for 500 simulations from a homogeneous Poisson process using the weighted edge-correction (4.23, labelled 'A') and the probabilistic edge-correction (labelled 'B') for estimating the $K(r)$-function 98
4.3 Results for 500 simulations from a simple inhibition process using the weighted edge-correction (labelled 'A') and probabilistic edge-correction (labelled 'B') for the $K(r)$-function 98
4.4 Results for 500 simulations from a Matérn cluster process using the weighted edge-correction (labelled 'A') and the probabilistic edge-correction (labelled 'B') for estimating the $K(r)$-function 99

5.1 Simulation assessment for fitting κ to Model 2 122
5.2 Simulation results for assessing the accuracy of the quadrature scheme 126
5.3 Simulation results for fitting Model 2 to 1000 simulated realisations of Model 2 126
5.4 p-values for KS tests for the appropriateness of Model 2 132
5.5 Model 2 parameter estimates for nuclei in PPDS2 132
5.6 p-values for KS tests for the appropriateness of Model 2 for PPDS2(b) PML NB data 135
5.7 Model selection using partial Bayes factors 152

C.1 Lookup table for values of σ for 100 simulations of a homogeneous Poisson process in the unit square 182

List of Figures

1.1 Scatter plots of the Japanese black pine (a), redwood seedlings (b) and cell centres data (c) 17

1.2 Organisation of the mammalian cell nucleus (Lanctot et al, 2007) . . 19

1.3 Confocal 3D images of MRC5 cell nuclei with three colour staining (PML bodies (green), nucleoli (red), and nuclear boundary (blue)), projected onto the XY plane. Images obtained by Imperial College London CSB and reproduced with permission. These are the 8 cell nuclei that make up PPDS2 21

1.4 Ellipsoid sampling regions for PPDS1. The image labeled '1' corresponds to the cell nucleus 1 data and so fourth. PML NB centroids are located at the red dots 23

1.5 PPDS1 CSR EDF simulation envelope plots. The top left graph is for cell nucleus 1. 24

3.1 An $\tilde{r}$-close point pair within distance r of an arbitrary point 48

3.2 The geometrical interpretation of $r^2 A_1(t, r, \tilde{r}) + A_2(t, r, \tilde{r})$ 51

3.3 Theoretical Q-function of a homogeneous Poisson process (a) and Gauss-Poisson process (b) 55

3.4 (a) Theoretical Q-function of a homogeneous Poisson process (white surface) and Gauss-Poisson process (Grey surface) (b) Q-function of a Gauss-Poisson process-$Q_{\text{poi}}(r, \tilde{r})$ 56

3.5 Simulated homogeneous Poisson process (a), Gauss-Poisson process (b) and p-thinned Poisson process (c) 57

3.6 Border estimate of the Q-function of a simulated homogeneous Poisson process (white surface) and simulated Gauss-Poisson process (grey surface) 58

3.7 Border estimate of the Q-function of a simulated homogeneous Poisson process (white surface) and simulated p-thinned Poisson process (grey surface) 59

3.8 Q-function of the theoretical homogeneous Poisson process (white surfaces) with estimates for the Japanese black pine (a), Redwood seedlings (b) and cell centres data (b) (grey surfaces) 60

3.9 Graph showing the percentage of times that CSR is rejected at the 5% significance level for a simulated mixed Matérn cluster and Strauss process with $n_p = 17$, $n_d = 3$, $n_s = 50$ and $R = 0.06$. The points represented by a circle are the results for the $Q(r, \tilde{r})$-function and those represented by a square are for the $L(r)$-function. 62

3.10 A sample of the NGSPP Model test data 64

3.11 Typical realisations of the NGSPP null Model 65

3.12 Results for the percentage of times that a hypothesis test based on using the Q-, K-, or T-function results in correctly rejecting the null model at particular significance levels 66

3.13 Results for the percentage of times that a hypothesis test based on using the Q-function results in correctly rejecting the null model at a particular level of significance compared to the K- and T-functions 66

3.14 $Q(r, \tilde{r})$-functions of cells 1-5 (white surfaces) of PPDS1 with simulation envelopes (grey surfaces). Results for cell 1 are on the top left. 67

4.1 A demonstration of edge effects when estimating a distance-based summary statistic in the nuclear interior 68

4.2 $K(r)$-function adjusted $MISE$ for the homogeneous Poisson process in three-dimensions (for $t = \frac{1}{4}\sqrt{2}$ (a), $\frac{1}{2}\sqrt{2}$ (b), 1 (c)) 70

4.3 $Q(r,\tilde{r})$-function adjusted $MISE$ for the homogeneous Poisson process in three-dimensions (with $t = \frac{1}{2}\sqrt{2}$) 71

4.4 $K(r)$-function adjusted $MISE$ for the homogeneous Poisson process in two-dimensions for $t = \frac{1}{4}\sqrt{2}$ (a), $\frac{1}{2}\sqrt{2}$ (b), 1 (c) 72

4.5 $K(r)$-function adjusted $MISE$ for the simple inhibition process in two-dimensions with various underlying intensities and $\delta = 0.01$ (a). 0.1 (b), 0.25 (c), with $t = \frac{1}{4}\sqrt{2}$ 73

4.6 $K(r)$-function adjusted $MISE$ for the Matérn cluster process in three-dimensions for R = 0.01 (a), 0.1 (b), 0.25 (c) and $t = \frac{1}{4}\sqrt{2}$ 74

4.7 $E(\widehat{K}(r)) \times 10^2$ (red line) obtained by simulating 100 events uniformly 500 times inside the unit square. The envelopes for $\widehat{K}(r) \times 10^2$ produced from these simulations (red dashed line) and the theoretical curve (black line) for a homogeneous Poisson process with intensity 100 are also displayed. 77

4.8 $E(\widehat{K}(r)) \times 10^2$ (red line) obtained by simulating 100 events uniformly 500 times inside the unit cube. The envelopes for $\widehat{K}(r) \times 10^2$ produced from these simulations (red dashed line) and the theoretical curve (black line) for a homogeneous Poisson process with intensity 100 are also displayed. 78

4.9 $BIAS(r)$ for the unedge-corrected L_1-function for the two-dimensional (blue) and three-dimensional case (red) 79

4.10 $MSE(r)$ for the two-dimensional (blue) and three-dimensional case (red) 81

4.11 $RELBIAS(r)$ (the bias ratio) for the two-dimensional (blue) and three-dimensional (red) case produced through simulation studies (the standard scenario). 82

4.12 The geometric meanings of A_1, A_2, and A_3 85

4.13 A plot of the theoretical $BIAS(r)$ for 100 events in a square window of length 3 and eroded window of length 1 85

4.14 A plot of the theoretical $RELBIAS(r)$ for 100 events in a square window of length 3 and eroded window of length 1 86

4.15 The empirical CDF, $\widehat{F}_G(r)$ of the distance between two events dropped in a cube of lengths 1 (blue), 2 (green), and 3 (red) 87

4.16 Theoretical $BIAS(r)$ for 100 events uniformly placed in the square (blue) and cube (red) of length 3 with eroded window of length 1 . . 88

4.17 Theoretical $RELBIAS(r)$ for 100 events uniformly placed in the square (blue) and cube (red) of length 3 with eroded window of length 1 88

4.18 A plot of $\widehat{K}(r)$ (using estimator (4.23) (red)) with simulation envelopes based on 100 simulations (red dashed curves) of 100 events uniformly placed in a sphere of radius 0.62. The theoretical $K(r)$-function curve (black line) for a homogeneous Poisson process is also provided for comparison 91

4.19 A brick window and eroded window (grey) 94

4.20 Geometrical illustration of estimator (4.29). The crosses represent events of the spatial point process 96

4.21 A plot of $E(\widehat{K}_{\text{Prob}}(r))$ (red) with simulation envelopes (red dashed curves) for $\widehat{K}_{\text{Prob}}(r)$ compared with the weighted edge-correction estimate (4.23) (blue) and the theoretical curve for a homogeneous Poisson process (black) 97

4.22 The mean probabilistic border estimate of the $K(r)$-function (red) compared to the mean of the usual border method (blue) and theoretical curve (black) using the standard scenario. The respective upper simulation envelopes for the border estimates are given by the red and blue dashed curves 101

4.23 Plot of an estimate of the $K(r)$-function (estimated using the standard border method) for 100 events uniformly placed in an ellipsoid (approximated by a convex hull) with semi-axis lengths 5, 4 and 2 and theoretical curve (red) 102

4.24 Plot of an estimate of the $K(r)$-function ($\times 324$), using equation (4.23) with $\omega(.)$ calculated using Monte Carlo integration (blue curve) for 100 events uniformly placed in an ellipsoid with semi-axis lengths 5, 4 and 2, and theoretical curve (red curve) 103

5.1 $Q(r, \tilde{r})$ function ($\times 10^{-10}$) of cells 1-4 of PPDS2 (grey surfaces) with simulation envelopes (transparent surfaces) 113

5.2 $Q(r, \tilde{r})$ function ($\times 10^{-10}$) of cells 5-8 of PPDS2 (grey surfaces) with simulation envelopes (transparent surfaces) 114

5.3 Cells 1 - 5 superimposed f_i and g_i function significance test results . . 115

5.4 Cells 6 - 8 superimposed f_i and g_i function significance test results . . 115

5.5 Cell 8 f_i and g_i function significance test results 116

5.6 Cell 6 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1 117

5.7 Cell 7 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1 117

5.8 Cell 8 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1 118

5.9 Cells 1-5 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1. Cell 1 is the top left and cell 2 the top right. 119

5.10 Histogram of cells 1 - 8 F-function significance levels against simulations from Model 2 128

5.11 EDF plot of u_i for cell 7 (black curve) with the theoretical CDF of a uniform distribution on (0,1) (dashed curve) 129

5.12 EDF plot of u_i for cells 1-4 (black curve) with the theoretical CDF of a uniform distribution on (0,1) (dashed curve) 130

5.13 EDF plot of u_i for cells 5,6, and 8 (black curve) with the theoretical CDF of a uniform distribution on (0,1) (dashed curve) 131

5.14 Cells 1-6 of PPDS2(b) estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1. 134

5.15 The graph on the left shows the value of α at a particular simulation (the mth transition in the Metropolis-Hastings procedure) and similarly for the parameter κ on the right. 143

5.16 Histograms of the values plotted in Figure 5.15. 144

5.17 A scatter plot of the values plotted in Figure 5.15. The correlation between the accepted values of α and κ was found to being -0.0077 . 144

5.18 Simulation envelopes for the $K(r)$-function of 100 simulated LGCPs with parameters $m = 0.0045, \kappa = 1.2$ and $\sigma = 0.3$ (dashed curves) with the average of the simulations (black curve) and the theoretical curve (red). 149

5.19 Simulation results for testing the accuracy of the partial Bayes factor (PBF) model selection method with $K = 1000$. On the left is a barchart of PBFs when testing the evidence for Model 1 against simulated realisations of Models 1 and 2. The barchart on the right is for the case when testing the evidence for Model 2. The interpretations of the PBFs are based on those given by Jeffreys (1961) and as also stated by Walsh & Raftery (2005) for Bayes Factors. 151

6.1 A two-dimensional projection of cell 4 of PPDS3 (see also Appendix B) produced by the Imperial College centre for structural biology. The bright green dots are PML NBs. The bright red dots are RNA Polymerase II 154

6.2 The $K_{12}(r)$-function plot for the PML (type 1) and RNA-Polymerase II (type 2) in cell nucleus 4 of PPDS3. The blocked curve is K_{12} and dashed curve is K_{21} 159

6.3 Histogram, density and CDF of z_π for all PML in PPDS2 162

6.4 Q-Q plot of the PML NB marks (left). The points are approximately linear, suggesting possible normality. The graph on the right shows the empirical CDF of the PML NB marks (dashed curve) with the CDF of a normal distribution with mean 0.045 and standard deviation 0.019 (black line). 163

6.5 The graph on the bottom right shows the empirical CDF of the PML NB marks (dashed curve) with CDF of a truncated normal distribution with mean 0.045 and standard deviation 0.019 (black line). The scatter plots are of PML NB centroid x, y, z coordinates (divided by nuclear length) against simulated realisations of a truncated (0.1) normal distribution with mean 0.045 and standard deviation 0.019 (the modeled marks distribution). The red crosses represent the data and the black crosses are for the simulated marks. 164

6.6 Marked (Black curve) and unmarked (dashed line) $J(r)$-functions for PPDS3 cells 1-4. The mark set $B = [0.01, 0.04]$ 165

6.7 Marked (Black curve) and unmarked (dashed line) $J(r)$-functions for PPDS3 cells 5-8. The mark set $B = [0.01, 0.04]$ 166

B.1 The details for PPDS1 178

B.2 PPDS2 details. U is the number of interior points that are used for simulating inside the nucleus and estimating integrals, as done in Chapter 5. This is a subset of the number of interior points provided by the CSB, which is generally in excess of 100,000 points. We take a random subset so that computational burden is reduced and we are not faced with computer memory issues when running programs. . . . 178

B.3 The different processes for obtaining the data that make up PPDS2 (cell 7 is used as an example). First the confocal image is produced (top left). The image data is analysed by the CSB to form datasets consisting of points that are labeled PML (bottom left), nucleoli, nuclear boundary, or empty space (top right) inside the nucleus. We then run this large dataset through a program that forms a point pattern by converting the PML bodies into PML points, by calculating centroids. Note that the nuclear boundary is obtained by calculating a convex hull for the points that have been classified from the image data (see Russell, 2006 for details) 179

B.4 PPDS3 details. U is the number of interior points that are used for simulating inside the nucleus and estimating integrals, as done in Chapter 5. This is a subset of the number of interior points provided by the CSB, which is generally in excess of 100,000 points. We take a random subset so that computational burden is reduced and we are not faced with computer memory issues when running programs. . . . 180

B.5 Two-dimensional projections of simulated realisations of Model 3 (using CFTP) with parameters $\alpha = 100, \kappa = 0.004$ and $R = 1.5$ inside the sphere of radius 10. The horizontal and vertical axis are respectively the $\eta^{(2)}$ and $\eta^{(2)}$ coordinates 181

E.1 $BIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 5$ (top left), 10 (top right), 30(bottom left). 100 (bottom right) conditional number of events) inside a cube (black line) and square (dashed line) of length 1 186

E.2 $BIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 5$, 10, 30, 100 conditional number of events) inside a cube (black line) and square (dashed line) of length 5 187

E.3 $BIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 5$, 10, 30, 100 conditional number of events) inside a cube (black line) and square (dashed line) of length 20 188

E.4 $BIAS$ of the unedge-corrected K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) inside a cube of length 1 189

E.5 $BIAS$ of the unedge-corrected K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) inside a cube of length 5 190

E.6 $BIAS$ of the unedge-corrected K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) inside a cube of length 20 191

E.7 $BIAS$ of the unedge-corrected K-function of a simple inhibition process (with underlying intensity = 5, 10, 30, 100 and $\delta = 0.1$) inside a cube of length 1 192

E.8 $BIAS$ of the unedge-corrected Q-function of a homogeneous Poisson process inside a cube of length 1 193

E.9 $BIAS$ of the unedge-corrected Q-function of a homogeneous Poisson process inside a cube of length 5 194

E.10 $BIAS$ of the unedge-corrected Q-function of a homogeneous Poisson process inside a cube of length 20 195

E.11 $RELBIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 90$, 150, 300 conditional number of events) inside a cube (black line) and square (dashed line) of length 5 196

E.12 $RELBIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 90, 150, 300$ conditional number of events) inside a cube (black line) and square (dashed line) of length 20 197
E.13 $BIAS$ of the K-function of a homogeneous Poisson process (with $n = 5, 10, 30, 100$ conditional number of events) estimated using the border method (black line) and probabilistic border method (dashed curve) inside a cube of length 1 198
E.14 $BIAS$ of the K-function of a Matérn cluster process (with $n = 5, 10, 30, 100$ conditional number of parent events and $R = 0.1$) estimated using the border method (black line) and probabilistic border method (dashed curve) inside a cube of length 1 199
E.15 $BIAS$ of the K-function of a simple inhibition process (with underlying intensity $= 5, 10, 30, 100$ and $\delta = 0.1$) estimated using the border method (black line) and probabilistic border method (dashed curve) inside a cube of length 1 200
E.16 The MISE for the Q-function using a border edge correction (grey surface) and no edge correction (red non-filled surface) for a homogeneous Poisson process with various expected number of events n in a cube window of various lengths. 201
E.17 Plots of $K(r) - \frac{4}{3}\pi r^3$ for the PML NBs in the cells that make up PPDS1, estimated using the volume weighting edge correction (black line) and probabilistic K-function estimate (dashed curve). 202
E.18 $K_{12}(r)$-function plot for the PML NBs (type 1) and RNA Polymerase II (type 2) in PPDS3 cells (black curve) with simulation envelopes of the $K_{12}(r)$-function of PML (type 1) with uniformly distributed events (type 2). 203

Chapter 1

Introduction

1.1 Overview

This thesis focuses on the statistical analysis of spatial data arising from coordinates of the centroids of objects found in animal tissue cell (Fibroblast) nuclei. In this case, we assume that the data is a realisation of a stochastic mechanism for placing objects in real coordinate space.

Analysis of spatial point patterns begins with a test for Complete Spatial Randomness (CSR). If a point pattern is asserted to exhibiting CSR in some bounded region $W \subset \mathbb{R}^d$, then it is a realisation of a homogeneous Poisson process. This implies that the number of objects in W has been generated by a Poisson process. Furthermore, these objects are uniformly scattered within W. Whether or not we reject the hypothesis of CSR is usually not of intrinsic value. Practitioners tend to express a greater interest in other information obtained from such tests. As an example, the extent to which departure from CSR occurs, may influence the choice of any models that are to be subsequently formulated for the data.

In this research, three point pattern datasets are analysed (PPDS1, PPDS2, and PPDS3 (see Appendix B)). All of the datasets are equipped with information on the Cartesian coordinates of Promyelocytic leukemia nuclear body (PML NB) centroids in a mammalian cell nucleus and have been obtained by structural biologists at the centre for structural biology (CSB) at Imperial College London, using confocal microscopy and image analysis. CSR tests are carried out on PPDS1 and PPDS2. The boundary of the entire region on which the point pattern is observed (in this case the region enclosed by the nuclear membrane) is unknown for PPDS1. In order

to carry out point process analysis that requires knowledge of the sampling region boundary, an ellipsoidal sampling region is constructed. We criticise this initial approach on the basis that inspection of confocal microscopy images provides evidence for Fibroblast nuclei being somewhat flat and ellipsoidal in shape (Bolzer et al. 2005).

The outcome of CSR tests on PPDS1 and PPDS2 may support introducing new biological theory on the function of certain subnuclear bodies. As an example, Mahy et al (2002) report that chromosomes with the highest gene density are preferentially disposed toward the nuclear interior, suggesting that the nuclear interior may facilitate, or create a permissive environment for, transcription. This may support a conjecture that the subnuclear body PML, would generally be located closer to the centre of the nucleus as it is believed to be involved in regulating transcription.

Testing for CSR usually involves estimating one of several quantities. These quantities (which, we will interchangeably refer to as functions, characteristics or summary statistics, as is the accepted practice) include the widely-used $K(r)$-function. Furthermore, and also as is the practice, we shall often abbreviate this as the K-function (and similarly for other such functions). The K-function's merits include ease of estimation and the ability to detect second-order features. However, higher-order features can be very informative but fail to be captured by the K-function. To address this shortcoming, we introduce a new third-order function, the Q-function. The function is a generalisation of Schladitz and Baddeley's T-function (Schladitz & Baddeley, 2000). The closed form Q-function for some well-known point processes is obtained. These include the homogeneous Poisson process, Gauss-Poisson process and a thinned Poisson process. We argue that the Q-function is able to provide a more detailed description of point patterns compared to popular lower-order functions. Furthermore, the Q-function is a more powerful tool in some hypothesis tests compared to some of the established summary statistics. However, the Q-function, (like all other point process summary statistics) suffers from some disadvantages. One of the most notable being that the Q-function estimator can be unstable and difficult to compute.

Estimation of summary statistics that rely on measuring inter-event distances, like the K- and Q-functions, is hampered by edge effects. We explore edge-correction in various three-dimensional settings. Although it appears intuitive, and has been remarked by several authors (for example Baddeley et al, 1993 and Cressie, 1991) that edge effects are more severe in three-dimensions compared to two: to the best

of our knowledge, there has not been any analysis carried out to demonstrate this. Furthermore, to what extent the severity is greater in three-dimensions compared to two is yet to be reported. We suspect that the reason for this is that finding a tool for comparing edge effects in different dimensions is less trivial than we would anticipate.

We provide theoretical results and simulation studies to confirm that edge effects are more severe in certain three-dimensional regions compared to their two-dimensional counterparts. More importantly, we illustrate to what extent the severity is greater. As we will see from this document, the region W and spatial point process are both factors that impact the severity of edge effects.

We introduce two new edge-corrections: a probabilistic border method and a probabilistic edge-correction for the K-function. We demonstrate through simulation studies that in certain cases these new procedures produce a smaller bias and variance compared to standard edge-correction techniques.

The datasets PPDS2 and PPDS3 are equipped with details of the nuclear boundary, albeit, in the form of a convex hull. As is often the approach, on rejecting the null hypothesis of CSR, we seek a more appropriate spatial point process model for the data. One of the models that we put forward is an inhomogeneous Poisson process with an intensity function that is constructed in such a way that it captures possible interaction between the nuclear boundary and the PML bodies. We fit this model to PPDS2 by maximising an adapted pseudo log-likelihood function via a grid search followed by a local stochastic optimisation near the log-likelihood surface extrema. We also look into the possibility of using Markov Chain Monte Carlo (MCMC) methods, coupled with perfect simulation and thermodynamic integration to fit a more complicated model to PPDS2. Following some insight gained whilst attempting to fit an inhomogeneous Poisson process to PPDS2, we are motivated to exploring a class of doubly stochastic models as a plausible model for the spatial distribution of PML nuclear bodies. Specifically, we study fitting a log-Gaussian Cox process (LGCP) to PPDS2 using a popular model fitting procedure known as minimum contrast.

We find that throughout this research, we are heavily reliant on simulation based inference. As a result we are regularly faced with the task of simulating spatial point processes inside a three-dimensional convex hull. It becomes apparent that

simulation of the LGCP and other sophisticated spatial point processes in a convex hull is non-trivial. Moreover, we also learn from this research that from the vast literature and our own experience, spatial point process analysis techniques can be very powerful tools for spatial data exploration. However, when analysing spatial point patterns such as those provided in PPDS1, PPDS2 and PPDS3, certain practical difficulties can arise. In such cases, other approaches may perhaps be more appropriate. A key example of a re-occurring complication is a small number of objects. This can be problematic in situation such as when estimating some of the point process functions using edge-corrections such as the border method, that become increasingly unusable as the number of objects decreases.

1.2 Spatial point processes and point patterns

Spatial point processes are mathematical models for random or irregular spatial point patterns. They play an important role in the field of spatial statistics (Cressie, 1991). A point pattern, as described by Diggle (2003), is data in the form of a set of points, irregularly distributed within a region of space. Examples of spatial point patterns include the location of trees in a forest, or of ants nests in a region inhabited by an ant colony. We refer to the objects of interest (for example the trees or ants nests) as events, so as not to confuse them with an arbitrary point in space. One of the most simple point processes is the homogeneous Poisson process. If a spatial point pattern is asserted to exhibit CSR in some region, then the statement is equivalent to 'the point pattern is a realisation of a homogeneous Poisson process in that region'. Most of the more complicated point processes such as the inhomogeneous Poisson process stem from the homogeneous Poisson process. These models can be divided into two broad categories; those which capture regularity, for example the simple inhibition process (see for example Cressie, 1991, section 8.5.4) and those which are best suited to clustered data, for example Cox processes (see for example Møller & Waagepetersen, 2004).

In the spatial point process literature, one of the most analysed point patterns include the Japanese black pine data introduced by Numata (1961) (cited by Diggle, 2003). The data was later tested for CSR by Bartlett (1964) using spectral analysis, then by Besag & Diggle (1977) and more recently again by Diggle (2003) using plots of the empirical distribution function (EDF) of point-to-nearest event distances. Datasets concerning the locations of redwood seedlings (Strauss, 1975, Ripley, 1977, and Diggle, 2003) and the centres of the locations of biological cells

(Crick & Lawrence, 1975, Ripley, 1977, and Diggle, 2003) have also received attention in the spatial point process literature. These three point patterns (displayed in Figure 1.1) have been of interest as they provide good examples of spatial data which evidently fall into one of the three basic categories of point patterns: CSR, regular and clustered data. These standard datasets are all examples of planar point

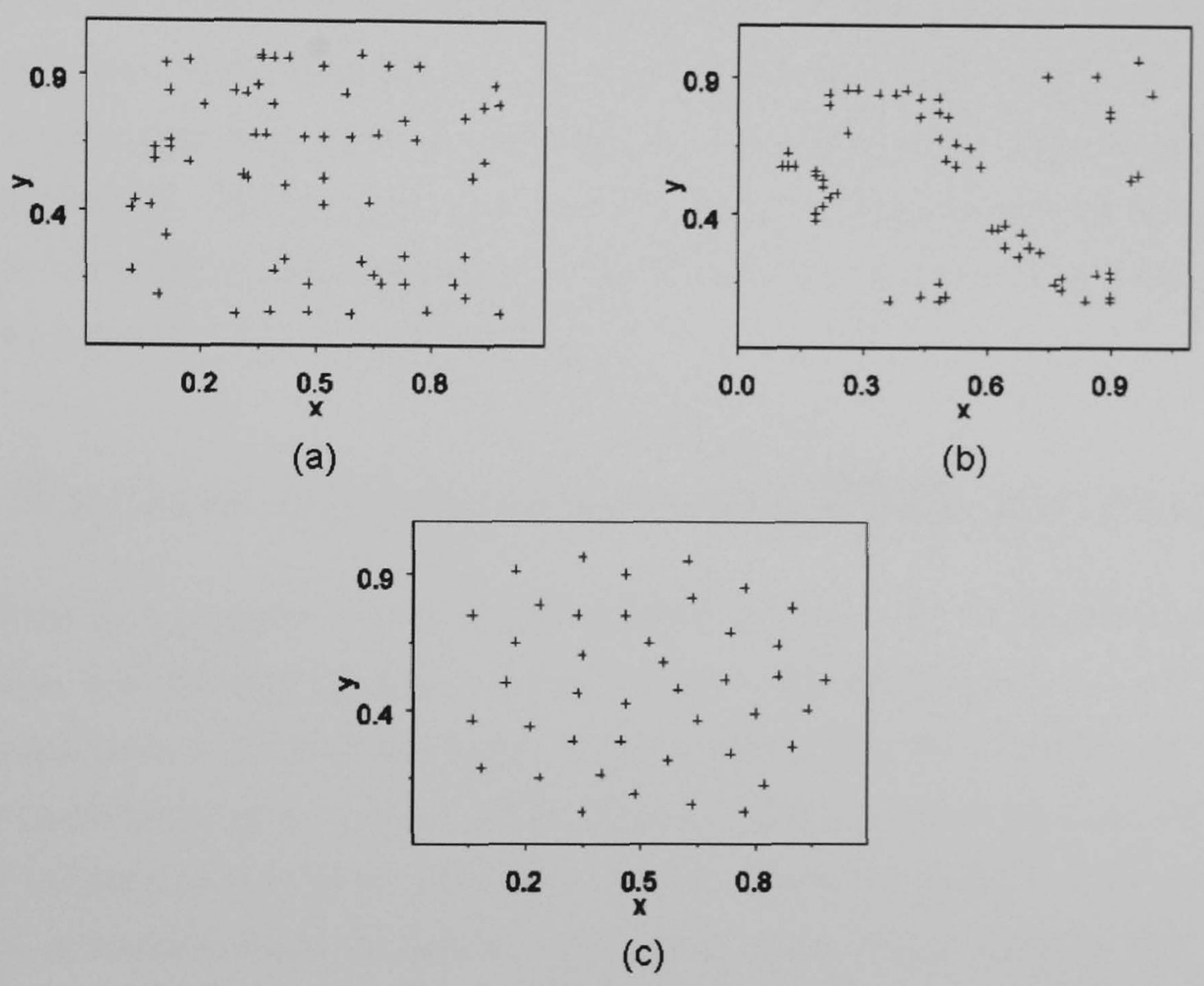


Figure 1.1: Scatter plots of the Japanese black pine (a), redwood seedlings (b) and cell centres data (c)

patterns. The locations of the events are specified in two-dimensional space. Data consisting of events with three-dimensional coordinates have gained less attention in the spatial point pattern literature. The reasons for this include three-dimensional data being generally more difficult to obtain. Furthermore, three-dimensional point patterns provide difficult statistical problems (Stoyan et al, 1995). Diggle et al (1991), Konig et al (1991), and Baddeley et al (1993) are important in the development of three-dimensional spatial point pattern analysis, using the three-dimensional versions of the standard distance-based summary statistics, such as the $K(r)$-, $F(r)$-, and $G(r)$-functions.

Sometimes it might be of interest to consider spatial point patterns which contain

multiple event types. An example of such a spatial point pattern is the displaced amacrine cells in the retina of a rabbit, analysed by Diggle (1985). Diggle (2003). and Baddeley et al (2006) and the forestry data analysed by Diggle & Cox (1983). These spatial point patterns are referred to as multivariate spatial point patterns and are special cases of a marked spatial point process (Stoyan et al, 1995). Analysis of marked spatial point patterns in literature is much less common than unmarked. Nevertheless, examples of analysis of marked spatial point patterns can be found in Baddeley et al (2006) and VanLieshout (2004). Using the information in PPDS2 we can estimate the diameter of the PML nuclear bodies (and we find that our estimates are consistent with the biological literature) at particular locations in the Fibroblast nuclei. We use this information to fit a marked inhomogeneous Poisson process to the data. The information in PPDS3 is used for carrying out bivariate analysis in Chapter 6.

1.3 Nuclear architecture and PML nuclear bodies

The nucleus is a complex (membrane enclosed) organelle (a discrete structure of a mammalian cell having specialised functions) with an internal structure and component organisation that is not fully characterised (Fox et al, 2002). It is the largest cellular organelle (varying from 11 to 22μm), and occupies around ten per cent of the total volume of a typical animal cell (Alberts et al, 2002). It is a site of major metabolic activities, such as DNA replication, gene transcription, RNA processing and ribosome subunit maturation and assembly (Lamond & Sleeman, 2003). These attributes make the nucleus the most prominent feature of eukaryotic cells. The main structural elements of the nucleus are a double membrane (nuclear envelope) that encloses the entire organelle, keeping its contents separated from the cellular cytoplasm (Paine et al, 1975). The nuclear envelope contains pores and rests on a meshwork of intermediate filaments called the nuclear lamina (Lactot et al, 2007) with the internal structure being characterized by a compartmentalised distribution of functional components (Lactot et al, 2007) as depicted in figure 1.2. These compartments, or, distinct classes of subnuclear bodies, include nucleoli, splicing speckles, Cajal bodies, gems and PML bodies (Dundr & Misteli, 2001, Fox et al. 2002). Although little is known about several of these compartments, their existence suggests that the nucleoplasm is not a uniform mixture, but rather, the compartmentalisation of nuclear components gives rise to a highly organised and tightly controlled environment (Hodges et al, 1998).

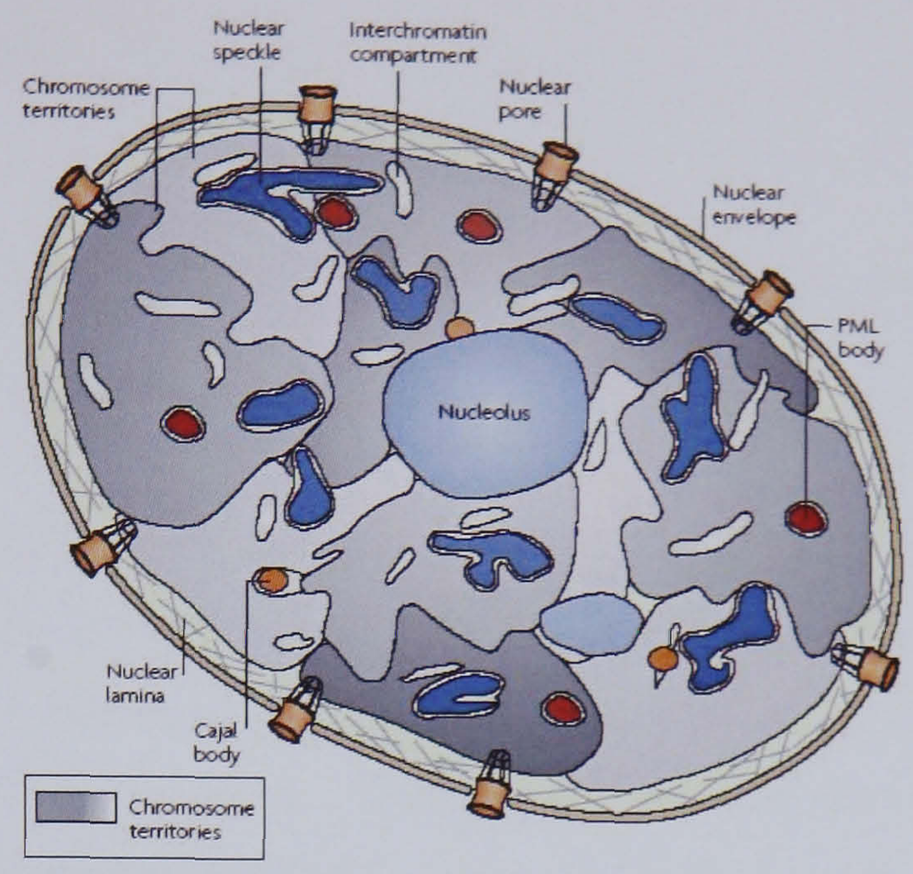


Figure 1.2: Organisation of the mammalian cell nucleus (Lanctot et al, 2007)

PML nuclear bodies (PML NBs), also referred to as nuclear domain 10s (ND10s), Kremer bodies (Kr bodies), and PML oncogenic domains (PODs) (Hodges et al, 1998, Bernardi & Pandolfi, 2003) have attracted great research interest. They have been reported as being "doughnut-like" in shape, with a dense fibrillar ring (Hodges et al, 1998). They vary in number, between 5 and 30 per nucleus (Wang et al, 2004), and typically have a diameter of between 0.2 and 1μm (Hodges et al, 1998, Melnick & Licht, 1999). However, their morphology and size alters during the cell cycle (Hodges et al, 1998). At least 15 constituent components have been identified as part of the PML bodies, including the PML protein itself, CBP and RNA (Hodges et al, 1998). PML and by inference PML NBs have been attributed apparently disparate roles in transcription, DNA repair, DNA replication, and RNA transportation. Despite the evident physiological importance and hence criticality of defining a set of discrete biochemical functions, the molecular and biochemical bases for physiological phenomena associated with PML NBs are yet to be well understood (Borden, 2002). For such reasons, biologists are very interested in the PML body in its own right, and in terms of how it relates to other nuclear components.

Throughout this thesis we make several references to the nucleoli. There is currently debate as to how these compartments influence the spatial organisation of PML NBs within the nucleus. Most mammalian cells contain 1-5 nucleoli, each 0.5-5.0μm in diameter (Spector, 2001). The nucleolus is differentiated into three regions; The fibrillar centres, fibrillar component, and granular region (Spector, 2001). The nucleolus is a the site of rRNA synthesis, processing and ribosomal subunits

assembly (Spector, 1993).

1.4 Producing spatial point patterns by use of confocal microscopy

Microscopic examination of cells has changed dramatically since its origin in the mid-seventeenth century (Eils & Chaitanya, 2003). Confocal laser scanning microscopy is a technique that enables the capture of sharp three-dimensional images (that would appear blurred when viewed with a conventional microscope) of specimens such as the cell nucleus and its compartments. The technique was pioneered by Marvin Minsky in 1955 (Semwogerere & Weeks, 2005) and is becoming a popular method for imaging specimens (Kyan et al, 2001). A laser scanning microscope incorporates two principal ideas: point by point illumination of a specimen and rejection of out of focus light. At any particular instant only one layer of the specimen is observed. The two-dimensional image plane is reconstructed one pixel at a time by a computer. A three-dimensional reconstruction of the specimen can be performed by combining a series of such slices at different depths (Prasad et al, 2007).

Images for producing the datasets PPDS1, PPDS2, and PPDS3 detailed in Appendix B, were captured as optical sections using a Zeiss confocal laser scanning microscope equipped with a Plan Apo 63×/NA 1.4 objective. Sections were collected at $0.4\mu m$ intervals (and these $0.4\mu m$ thick slices overlap by $0.2\mu m$ so that all areas of the nuclei are captured) through each nuclei and processed using imaging software, which produced three-dimensional coordinates of centroid positions for PML bodies and each genomic locus (Shiels et al, 2001). To produce PPDS2 (the dataset consisting of the cell nuclei shown in Figure 1.3) and PPDS3, cells are immunofluorescently labelled (see for example Xie & Pombo, 2006) by biologists so that fluorescent markers target specific objects within the nucleus. An unclear Red, Green and Blue (RGB) image is then obtained, with each colour associated with a different channel. A popular segmentation technique known as thresholding is applied to the image data so that we are able to distinguish between the four objects: PML NBs, other subnuclear body. empty space in the nucleus interior, and nuclear boundary. Once this method has been adopted, the nuclear boundary can consequently be presented in the form of a convex hull using methods described in Russell (2006). These datasets have the advantage that we are not required to make assumptions about the sampling region.

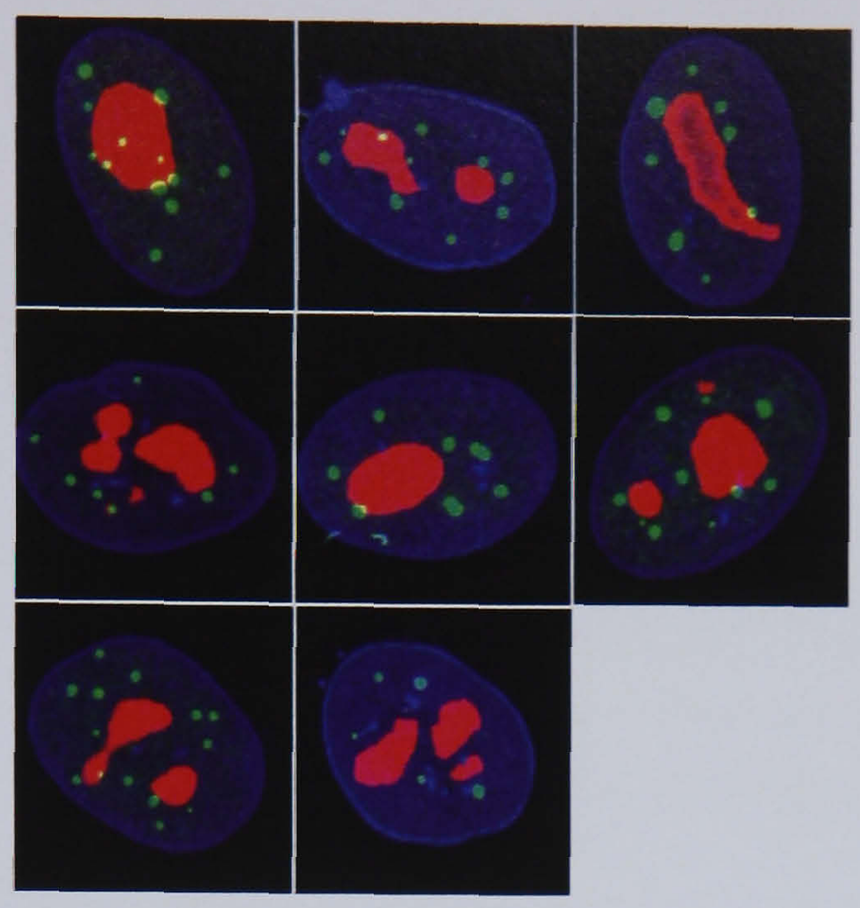


Figure 1.3: Confocal 3D images of MRC5 cell nuclei with three colour staining (PML bodies (green), nucleoli (red), and nuclear boundary (blue)), projected onto the XY plane. Images obtained by Imperial College London CSB and reproduced with permission. These are the 8 cell nuclei that make up PPDS2

1.5 Initial tests for CSR on PPDS1

The notion of subnuclear bodies being non-randomly distributed within the cell nucleus dates back to the early 1900s (Agard & Sedat, 1983). Several techniques have been adopted for assessing the spatial distribution of nuclear bodies. A popular and relatively fast approach for assessing whether nuclear compartments exhibit spatial positioning preference is known as erosion analysis or "nuclear peeling" (Shiels et al, 2007). This entails some form of radial analysis, in which the nucleus is subdivided into concentric rings or shells from the periphery to the centre. The location of the nuclear compartment relative to the geometric centre is then calculated and usually expressed as a fraction of the distance to the nuclear lamina. This approach has been adopted by several authors including Wiblin et al (2005), Zink et al (2004), and Cremer et al (2001).

Other techniques for investigating subnuclear body spatial preference have included those adopted by Bolzer et al (2005). They used the mean of inter-body distances and Kolmogorov-Smirnov (KS) tests to assess the spatial distribution of subnuclear bodies. Wang et al (2004) analysed the correlation between the minimum locus-PML distances against their local transcriptional activity to show that PML

associate with transcriptionally active genomic regions.

Recently, practitioners have began adopting a spatial point process framework for assessing the spatial preference of subnuclear bodies. Perhaps one of the main reasons for the interest in the application of spatial statistics to biological point patterns is the rigorous mathematical theory accrediting the methods and techniques used in spatial point pattern analysis. Nordmans et al (1998) make use of histograms coupled with the pair-correlation function to assess the "randomness" of two proteins in the nucleus of a cultured human cell line. More recently, Fleischer et al (2006) have used several point process statistics to describe the topology of chromocentres during differentiation of NB4 cells.

Information on spatial preference is usually obtained from a CSR test. This is usually carried out via a Monte Carlo approach (see for example Besag & Diggle, 1977, Besag & Clifford, 1989, and Diggle, 2003 (section 1.7) for discussions on the advantages and disadvantages of Monte Carlo tests). Simulation approaches are required because closed form expressions for the probability that a point pattern is CSR is in general difficult to obtain. This procedure generally involves taking the common approach for assessing uniformity, by simulating (the null model) inside a theoretical nucleus and comparing theoretical quantities with the observed. Models used for the nucleus include a sphere or ellipsoid (Weimer et al, 1992). N simulations (where N depends on the the level of significance that we are interested in) are conducted and then simulation envelopes, obtained from the EDFs of the N simulations are compared with that of the observed data (see for example Diggle, 2003, Chapter 2).

The approach we take in this chapter, for PPDS1 (where no information on the nuclear boundary is provided), is not to attempt to model the nuclear boundary, but to form a sampling region which is assumed to be representative of the whole nucleus interior. Diggle (2003) discusses the different types of sampling. The method we adopt entails selecting a sampling region as a subset of a larger region (the larger region is the entire cell nucleus). Diggle (2003) refers to this sampling method as intensive mapping. For this approach, the sampling region is usually chosen according to some probability sampling scheme, or as is in our case, in a way that reflects the experimenters' view that it is in some sense representative of the larger region (Diggle, 2003). In developing tests for CSR for mapped patterns, it is more usual to analyse the pattern conditional on the observed number of events (Diggle, 2003).

Further, the use of replicated data, or in this case, multiple cells provides a stronger argument to support the outcome of CSR tests (Diggle, 2003).

An ellipsoid sampling region, as shown in Figure 1.4 for the five cells in PPDS1 is constructed as follows for an individual cell. We carry out rigid body transformations on the PML NB points within the individual cell, so that the line passing through the two PML bodies furthest apart is a subset of $\{a + \beta b | \beta \in \mathbb{R}\}$, where $a = (0, 0, 0)^T$ and $b = (1, 0, 0)^T$. We then use an optimisation scheme to compute the lengths of the semi-axis of the ellipsoid with smallest volume that all the PML NB points are contained in. We shall refer to this approach as the 'smallest ellipsoid method'. We are aware that there are issues surrounding the three-dimensional coordinate system of the confocal microscope. Specifically, a three-dimensional image is produced by stacking two-dimensional sections (250×250 pixel images) at $0.4\mu m$ intervals (Russell, 2006). Until stated otherwise, let $F(r)$ represent the observed

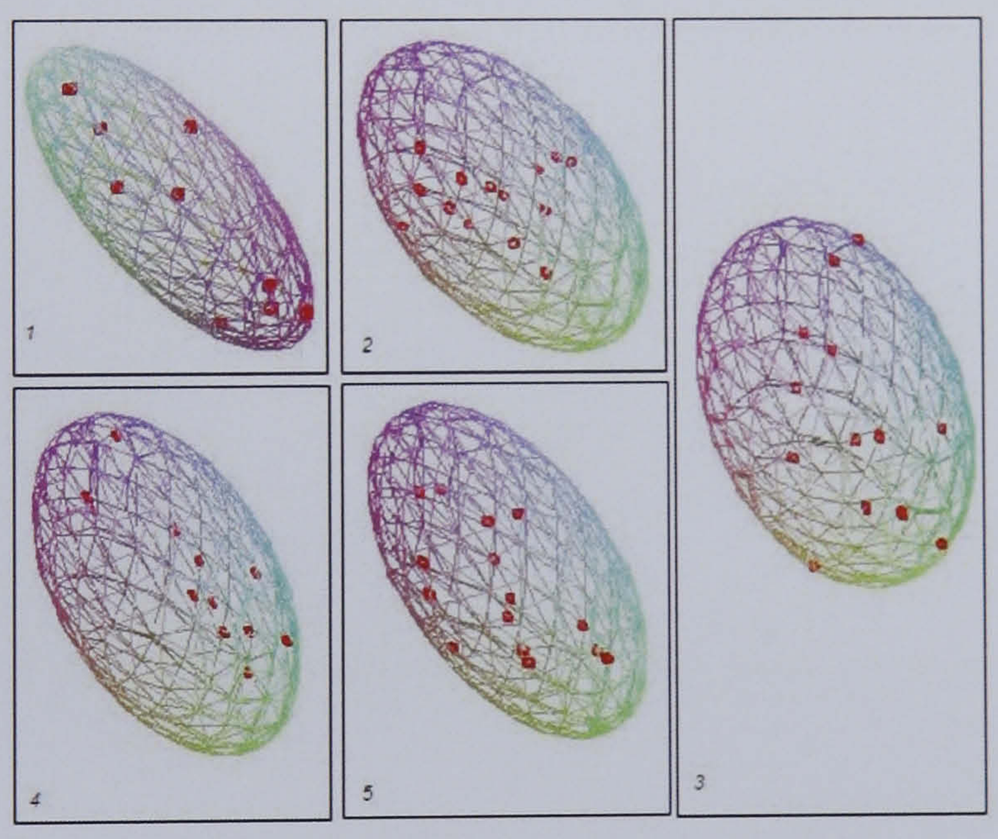


Figure 1.4: Ellipsoid sampling regions for PPDS1. The image labeled '1' corresponds to the cell nucleus 1 data and so fourth. PML NB centroids are located at the red dots

proportion of inter-event (inter-PML NB) distances which are at most r; that is

$$F(r) = \frac{\sum_{\eta} I_{(\|\eta - \zeta\| \leq r)}}{2n(n-1)} \tag{1.1}$$

where $n \in \mathbb{R}^+ \backslash \{1\}$ is the total number of events (PML NBs) and $\eta, \zeta \in \mathbb{R}^3$ are the spatial event locations. Furthermore $I_{(.)}$ is the indicator function and $\|.\|$ is the Euclidean norm. Let n_i be the number of PML bodies in W_i, where W_i denotes the ellipsoid sampling region for cell i, obtained by using the smallest ellipsoid method.

CSR tests on PPDS1, for each of the five cells in PPDS1, were carried out by simulating n_i events 99 times inside W_i and plotting $\inf\{F(r)_{ij} : j = 1 \ldots 99\}$ and $\sup\{F(r)_{ij} : j = 1 \ldots 99\}$, where $F(r)_{ij}$ is the calculated $F(r)$ for the jth simulation of n_i events *uniformly* and independently inside W_i, against the calculated $F(r)$ for the PML bodies locations in cell nucleus i $(i = 1 \ldots 5)$. Similar to Baddeley et al (1993), we ignore edge effects in these CSR hypothesis tests. The simulation envelopes and observed data EDF plots are shown in Figure 1.5. Noting that 100

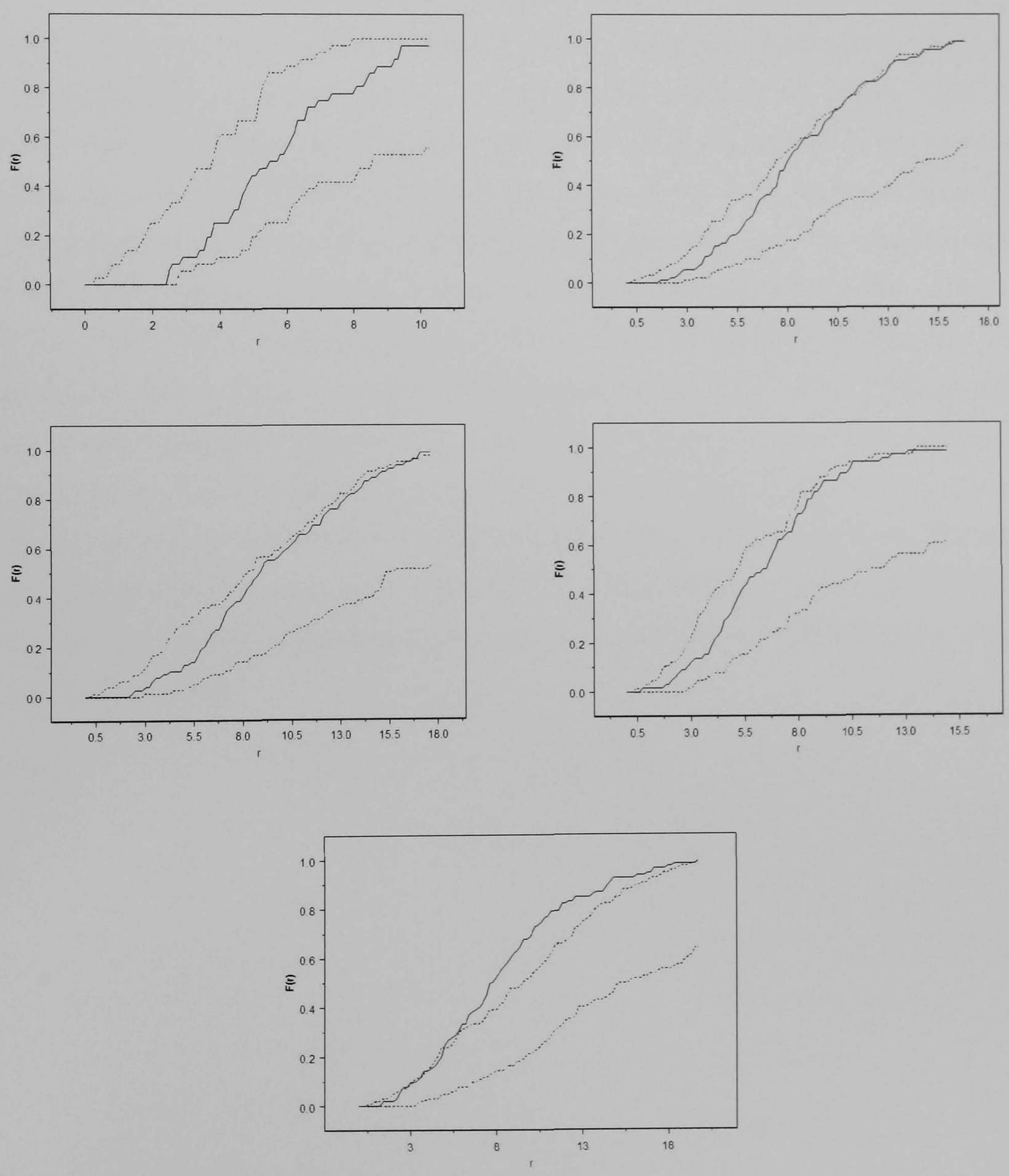


Figure 1.5: PPDS1 CSR EDF simulation envelope plots. The top left graph is for cell nucleus 1.

observations is sufficient for a test at the 5% level (Diggle, 2003, Hope, 1968), we

see that at this level, for nuclei 2-5, there is slight evidence to reject the null hypothesis of CSR. Furthermore, there appears to be a general indication of clustering compared to CSR, for large values of r. The results for cell 1 are particularly interesting since for this particular cell nucleus, compared to events that are CSR, the PML bodies appear to be distributed more regularly for $r \leq 3$ units (which is approximately $0.25\mu m$. See section B.1 of Appendix B), CSR for $3 < r \leq 9$ units and slightly clustered for $r \geq 9$ units.

Up until now it has been believed that the spatial organisation of PML NBs is influenced by other subnuclear bodies such as nucleoli. Biological ideas that suggests that nucleoli act as exclusion, or favourable regions for PML NBs, would be consistent with rejecting the null hypothesis of CSR provided that the nucleoli do not themselves exhibit CSR within the nucleus. Note that, had we been certain of the theory regarding the spatial distribution of nucleoli and its possible interaction with PML NBs, then we could deduce (without the need for CSR tests) whether or not the PML NB locations exhibit CSR in the cell nucleus. However, as demonstrated in our tests, there is more to be gained from CSR tests. Interestingly, the results for cell 1 suggest that there is not evidence to reject the CSR null hypothesis. This result does not necessarily contradict an idea that PML NB locations in the nucleus are possibly influenced by other subnuclear bodies such as nucleoli, since for example, perhaps no nucleoli exist in W_1. This demonstrates the importance of careful selection of the sampling region and the additional information gained from using replicated data.

Chapter 2

Spatial point processes: Theory, models and statistics

Neyman (1939) was an important early contribution to the general theory of point processes. The earliest discussions on spatial point processes date back to the early 1950s when used by Skellam (1952) and Thompson (1955) in statistical ecology. The work of Matérn in 1960, later re-published in 1986, has been viewed as pioneering (Cox & Isham 1980, Kemp, 1988). More recently, Møller & Waagepetersen (2004) and Møller & Waagepetersen (2006) have provided more detailed accounts of modern spatial point process theory, statistics and models, while Stoyan (2006) has provided a good and relatively concise background on important point process theory.

In this chapter we provide formal background on important spatial point process theory and concepts used up to Chapter 6 (where the theory is then extended to marked spatial point patterns). Although we focus on *spatial point processes* as these are specific to analysing PPDS1, PPDS2, and PPDS3, there is vast literature on the theory of point processes in various domains. For point processes along a time axis one can refer for example to Cox & Isham (1980). Theory on space-time point processes can be found in Fishman & Snyder (1976) with applications in Rathbun & Cressie (1994).

We begin by introducing notation, definitions and some important concepts. These initial definitions and concepts are as those given by Cressie (1991), Karr (1991), Stoyan & Stoyan (1994), Stoyan et al (1995), and Stoyan (2006). The precise definition of a spatial point process requires some measure theoretic concepts. Some of

the measure theory required throughout this thesis is included in Appendix A and is owed to the work of Kelley (1955), Hausdorff (1957), Halmos (1974), Cohn (1980), and Reid & Szendrői (2005). We will move onto discuss various spatial point process statistics. We shall only briefly mention edge effects, as this is an important issue in practical applications of spatial point processes and hence part of our reason for providing a more detailed treatment on the subject in Chapter 4. We will then go on to look at some important spatial point process models.

Throughout we denote locations in $\mathbb{R}^d$ by η, ζ, ξ, etc. and at times use subscripts (for example $\eta_1, \ldots, \eta_n$) to mean n locations. The notation

$$\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(d)})$$

will be used for the vector form of $\xi \in \mathbb{R}^d$. Our discussions will often be biased towards the case $d = 3$ in line with the datasets analysed and hence for instance, we shall denote the d-dimensional ball, of radius r centred at ξ by $b^d(\xi, r)$ and an omission of the superscript d will sometimes occur when it is clear what value d takes.

2.1 Spatial point process definition and characterisation

Let E be a locally compact Hausdorff space (every point of E has a compact, closed, and bounded neighbourhood) whose topology has a countable base, with Borel σ-algebra $\mathcal{E}$. Also, a measure μ on E is a Radon measure if $\mu(A) < \infty$ (it is locally finite) for all A in the ring of bounded Borel sets, $\mathcal{B}$. Let M be the set of Radon measures and $M \supseteq M_p = \{\mu \in M : \mu(A) \in \mathbb{N}, \forall A \in \mathcal{B}\}$. Furthermore, define $\mathcal{M}$ as the σ-algebra generated by the coordinate mappings $\mu \to \mu(f) = \int f d\mu$ where f ranges over the set of continuous functions on E whose support is compact. Also, denote the trace σ-algebra on M_p by $\mathcal{M}_p$. Then as stated by Karr (1991), we have the following definitions:

Definition 1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space.*
a) A random measure on E is a measurable mapping X of $(\Omega, \mathcal{F})$ into $(M, \mathcal{M})$.
b) A point process on E is a measurable mapping X of $(\Omega, \mathcal{F})$ into $(M_p, \mathcal{M}_p)$.

A *spatial* point process is a point process with $E \equiv V \subset \mathbb{R}^d$ and $M_p = N$ is the set of locally finite counting measures on V. Furthermore, $\mathcal{M}_p = \mathcal{N}$ is the smallest

σ-algebra generated by sets of the form $\{x \in N : x(B) = n\}$ for all B in the Borel σ-algebra of V and all $n \in \mathbb{Z}^+$. For spatial point process X, we write

$$x(B) = \sum_{\eta \in X} I_{(\eta \in B)}$$

where $I_{(.)}$ denotes the indicator function. Furthermore, for set S, the notation $X(S) = n$ means that S contains n points of X. An intuitive way of thinking about a spatial point process provided by Stoyan et al (1995) is that it is a random choice of one of the x in N.

The natural way to define a *spatial point pattern* (a realisation of a spatial point process) is through locations of events in V (Cressie, 1991). For mathematical convenience (Cressie, 1991), we define a spatial point pattern through a locally finite counting measure x on V. Conversely, ξ is an event of the point pattern if $x(\{\xi\}) > 0$.

X generates a distribution P on $[N, \mathcal{N}]$, the distribution of X, and this distribution is determined by the probabilities

$$\begin{aligned} P(Y) &= \mathrm{P}(X \in Y) \\ &= \mathrm{P}(\{\omega \in \Omega : X(\omega) \in Y\}) \end{aligned}$$

for $Y \in \mathcal{N}$. Moreover, we say that two spatial point processes X_1 and X_2 are identically distributed if $P(X_1 \in Y) = P(X_2 \in Y)$ for all $Y \in \mathcal{N}$. Throughout we shall assume that our spatial point process is *simple*. We say that a spatial point process is *simple* if $x(\{\xi\}) \in \{0, 1\}$ for all $\xi \in V$ and almost all $x \in N$. If the point process X is simple, then its distribution P is determined by the system of *void probabilities* v_c as c ranges through the compact sets and

$$\begin{aligned} v_B &= \mathrm{P}(X \cap B = \emptyset) \\ &= P(\{x \in N : x(B) = 0\}) \end{aligned}$$

for Borel sets B.

We will at times make the assumption that the spatial point process X is *stationary* or *isotropic* or both. We say that the point process X is *stationary* (or equivalently, *homogeneous*) if it has the property that $X_\xi = \{\xi + \eta : \eta \in X\}$ has the same distribution for all $\xi \in \mathbb{R}^d$. Also, the point process X is said to be *isotropic* if it is invariant under rotation. That is, applying (the same) rotation about the

origin to every $\eta \in X$ does not alter the distribution of X. Stationarity and isotropy can be very important assumptions in spatial point process analysis. Assuming stationarity or isotropy often simplifies calculation. Also, practitioners will at times (often implicitly) assume that stationarity holds (without carrying out formal tests for stationarity), or be content that it holds approximately so that certain point process techniques can be readily adopted (see Baddeley et al, 1993 and Glasbey & Roberts, 1997, as examples). *Ergodicity* is also an important property for spatial point processes. The property ensures that one sample of the point process is sufficient to obtain statistically secure results (Stoyan & Stoyan, 1994). We say that the stationary distribution P on $[N, \mathcal{N}]$ is *ergodic* if for all *invariant* sets $Y \in \mathcal{N}$ we have

$$P(Y) \in \{0, 1\}$$

where by *invariant* set $Y \in \mathcal{N}$ we mean one for which

$$P(Y \backslash Y_\xi \cup Y_\xi \backslash Y) = 0 \quad \forall \xi \in \mathbb{R}^d$$

where $Y_\xi = \{x \in N : x_\xi \in Y\}$. The stationary Poisson process is an example of an ergodic spatial point process (Stoyan et al, 1995).

2.2 Moments and moment measures

The mean is an important characteristic of real-valued random variables. The *intensity measure* Λ of X is a point process characteristic analogous to the mean, that is defined as

$$\Lambda(B) \;=\; E(X(B)) = \int x(B) P(dx) \quad \left(= E\left(\sum_{\eta \in X} I_{(\eta \in B)} \right) \right)$$

In the *homogeneous* case it suffices to consider an *intensity* λ since then

$$\Lambda(B) \;=\; \lambda \nu(B)$$

where we use ν to denote the Lebesgue measure on $(\mathbb{R}^d, B)$ for Borel set B of $\mathbb{R}^d$.

In general, we define the kth moment measure $\mu^{(k)}$ by

$$\mu^{(k)}\left(B_1 \times \ldots \times B_k\right) \;=\; E\left(X(B_1) \ldots X(B_k)\right)$$

for Borel sets $B_1, \ldots, B_k$, and $\times$ denoting the Cartesian product. Furthermore, for any positive measurable function $f(\eta_1, \ldots . \eta_k)$ we have

$$E\left(\sum_{\eta_1, \ldots, \eta_k \in X} f(\eta_1, \ldots, \eta_k) \right) \;=\; \int f(\eta_1, \ldots . \eta_k) \mu^{(k)}(d(\eta_1, \ldots . \eta_k)).$$

The *factorial* kth *moment measure* $\alpha^{(k)}$ is similar to $\mu^{(k)}$, formally defined as

$$E\left(\sum_{\eta_1,\ldots,\eta_k\in X}^{\neq} f(\eta_1,\ldots,\eta_k)\right) = \int f(\eta_1,\ldots,\eta_k)\alpha^{(k)}(d(\eta_1,\ldots,\eta_k)) \quad (2.1)$$

The only difference between $\alpha^{(k)}$ and $\mu^{(k)}$ is that the summation on the left hand side of (2.1) does not include *n-tuples* containing two or more equal points, and the $\neq$ symbol is used to show this. The *kth product density*, $\varrho^{(k)}(\eta_1,\ldots,\eta_k)$ is a density function which exists whenever $\alpha^{(k)}$ is suitably continuous. Moreover, $\varrho^{(k)}(\eta_1,\ldots,\eta_k)$ is defined as

$$\alpha^{(k)}(B_1\times\ldots\times B_k) = \int_{B_1}\ldots\int_{B_k}\varrho(\eta_1,\ldots,\eta_k)d\eta_1\ldots d\eta_k$$

(2.1) can be rewritten in terms of $\varrho^{(k)}(\eta_1,\ldots,\eta_k)$ as

$$E\left(\sum_{\eta_1,\ldots,\eta_k\in X}^{\neq} f(\eta_1,\ldots,\eta_k)\right) = \int f(\eta_1,\ldots,\eta_k)\varrho^{(k)}(\eta_1,\ldots,\eta_k)d\eta_1,\ldots,d\eta_k \quad (2.2)$$

Stoyan & Stoyan (1994) provide a more detailed account on higher-order moments, including some geometrical interpretations. We shall see later that third and higher-order moments, although not commonly used, may also be of great importance.

2.3 Palm distributions

The idea of *Palm distributions* is one of the most fundamental to the theory of point processes. Later on we will see that it may be of interest to condition on there being a point at some fixed location such as the origin. Heuristically, the *Palm distributions* P_o (with respect to the origin o) can be thought of as the conditional distribution of X given that there is a point at the origin. Formally, suppose that X is a stationary point process, with finite and non-zero intensity λ. Then the *Palm distribution* at o of P is a distribution defined on $[N,\mathcal{N}]$ by

$$P_o(Y) = \int\sum_{\eta\in x\cap B} I_{(x_{-\eta}\in Y)}P(dx)(\lambda\nu(B))^{-1} \quad (2.3)$$

for $Y\in\mathcal{N}$ and $x_{-\eta}=\{\xi-\eta:\xi\in x\}$. The *reduced Palm distribution* (at ξ) is even more important as it is used in the definition of important point process functions such as the $K(r)$-function. It can be interpreted as the conditional distribution of X on the reduced set $N\backslash\{\xi\}$, given that there is a point at ξ. The definition of the *reduced Palm distribution*, $P_o^!$ (at the origin) is obtained by replacing $I_{(x_{-\eta}\in Y)}$

with $I_{(x_{-\eta}\setminus\{o\}\in Y)}$ in (2.3). A more modern definition of the Palm distributions is presented in terms of the reduced and unreduced *Campbell measure.* The *reduced Campbell measure* of a point process X is a measure on $\mathcal{X}\times\mathcal{N}$ defined by

$$C^!(B\times Y) = \int\int I_{(x\setminus\{\xi\}\in Y)}x(d\xi)P(dx), \tag{2.4}$$

for $Y\in\mathcal{N}$ and B in $\mathcal{X}$, the Borel σ-algebra of $V\subset\mathbb{R}^d$. For $Y\in\mathcal{N}$, $C^!$ is absolutely continuous with respect to Λ (See Cressie (1991) section 8.3.4). By the Radon-Nikodym theorem (see Appendix A), there exist uniquely determined measures $P^!_\xi$ on $(N,\mathcal{N})$ such that

$$C^!(B\times Y) = \int_B P^!_\xi(Y)\Lambda(d\xi) \tag{2.5}$$

2.4 The Binomial point process

The points $\xi_1\ldots\xi_n$ form a *Binomial point process,* $X_{\text{bin}(W,n)}$ in the compact set $W\subset\mathbb{R}^d$ if they are independently and uniformly distributed inside W, where a random point ξ is uniformly distributed in W if

$$P(\xi\in U) = \frac{\nu(U)}{\nu(W)} \tag{2.6}$$

for all Borel sets $U\subset W$. We therefore have from (2.6) that

$$\begin{aligned}\mathrm{P}(\xi_1\in B_1,\ldots,\xi_n\in B_n) &= \mathrm{P}(\xi\in B_1)\cdot\ldots\cdot\mathrm{P}(\xi\in B_n)\\ &= \frac{\nu(B_1)\cdot\ldots\cdot\nu(B_n)}{\nu(W)^n}\end{aligned}$$

for $B_1,\ldots,B_n$ Borel subsets of W. The intensity λ of the Binomial point process is given by

$$\lambda = np(A)(\nu(B))^{-1}$$

where $p(B)=\nu(B)/\nu(W)$. The void probabilities of $X_{\text{bin}(W,n)}$ are given by

$$\mathrm{P}(X_{\text{bin}(W,n)}(V)=0) = \frac{(\nu(W)-\nu(V))^n}{\nu(W)^n}$$

for arbitrary compact $V\subset W$. If the events $\xi_1,\ldots,\xi_n$ form a binomial point process in W then the random point pattern formed by these events can be thought of as a random set. The point pattern can also be thought of as a random counting measure (Stoyan et al, 1995).

The simulation of virtually all spatial point process in W requires simulating a binomial point process (n points uniformly inside W). Simulating n events uniformly inside W, a unit square or cube is straightforward; for a unit cube, one simply superimposes n independent uniform random points, $\eta_1, \ldots, \eta_n$ where $\eta_i = (\widetilde{u}_{i1}, \widetilde{u}_{i2}, \widetilde{u}_{i3})$, and $\widetilde{u}_{ij} \sim$Uniform(0,1). Once simulated inside the unit cube, we can then apply scaling and translation in order to obtain a simulation inside any fixed cuboid. Coordinate transformation can be used for simulating uniformly inside a sphere. That is, if $\xi = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$ is a uniform point in the unit cube, then $\overrightarrow{\xi} = (\overrightarrow{\xi^{(1)}}, \overrightarrow{\xi^{(2)}}, \overrightarrow{\xi^{(3)}})$ where

$$\begin{aligned} \overrightarrow{\xi^{(1)}} &= R\sin(\theta)\cos(\phi) \\ \overrightarrow{\xi^{(2)}} &= R\sin(\theta)\sin(\phi) \\ \overrightarrow{\xi^{(3)}} &= R\cos(\theta) \end{aligned}$$

and

$$\begin{aligned} R &= \xi^{(1)\frac{1}{3}}, \quad 0 < R < 1 \\ \theta &= \arccos(1 - 2\xi^{(2)}), \quad 0 < \theta < \pi \\ \phi &= 2\pi\xi^{(3)}, \quad 0 < \phi < 2\pi \end{aligned}$$

is uniform inside the unit sphere. Once again, the relevant scaling can be applied for the case where simulation inside an ellipsoid is required. This is the approach used in all subsequent chapters. For non-ellipsoid and cuboid sets $W_0 \subset \mathbb{R}^3$, rejection sampling (see for example Ripley, 1987) can be used. This involves, for example, simulating uniformly inside $W \supset W_0$ and retaining the points $\xi \in W_0$. Simulation is repeated until the desired number of points are obtained.

Datasets PPDS2 and PPDS3 provide the locations of points that constitute empty space inside the nucleus. We exploit this information as a means of simulating n points uniformly inside the nucleus. Specifically, for the u points $\xi_1, \ldots \xi_u$ that are classified as empty space, n such points are chosen at random (without replacement). Random selection is made by equipping each of the $\xi_1, \ldots \xi_u$ with a unique integer $k \in \{1, \ldots, u\}$. The point ξ is selected if the randomly chosen $y \in \{1, 2, \ldots u\}$ is its assigned integer. Occasionally, we attempt to simulate uniformly inside the nuclear interior by adopting methods for simulating uniformly inside a convex hull. Such methods are detailed by Fishman (1996). The choice of random number generator is of importance. Our choice and preference is the Mersenne Twister algorithm (Matsumoto & Nishimura, 1998).

2.5 Homogeneous Poisson point process

The homogeneous (stationary) Poisson process X_{poi} is defined by the following postulates

(i) For some constant $\lambda > 0$, and bounded Borel set B, $X_{\text{poi}}(B)$ follows a Poisson distribution with mean $\lambda\nu(B)$. The parameter λ is the intensity. For the three-dimensional case, this can be interpreted as the number of events per unit volume.
(ii) Given $X_{\text{poi}}(B) = n$, the n events in B form an independent sample from the uniform distribution on B.
(iii) If $B_1, \ldots, B_s$ are disjoint bounded Borel sets then $X_{\text{poi}}(B_1), \ldots . X_{\text{poi}}(B_s)$ are independent Poisson random variables with mean $\lambda\nu(B_1), \ldots, \lambda\nu(B_s)$. Therefore

$$\begin{aligned} \mathrm{P}(X_{\text{poi}}(B_1) &= n_1, \ldots, X_{\text{poi}}(B_s) = n_s) \\ &= \frac{\lambda^{\sum_{i=1}^{s} n_i} \nu(B_1)^{n_1} \cdot \ldots \cdot \nu(B_s)^{n_s}}{n_1! \cdot \ldots \cdot n_s!} \exp\left(-\sum_{i=1}^{s} \lambda\nu(B_i)\right) \end{aligned}$$

These postulates are simultaneously the definition of CSR. The void probabilities of X_{poi} are given by

$$\mathrm{P}(X(B) = 0) = \exp\left(-\lambda\nu(B)\right)$$

The first-order moment Λ follows from property (i) and is given by

$$\Lambda(B) = E(X(B)) = \lambda\nu(B)$$

for all Borel sets B.

The homogeneous Poisson spatial point process can be simulated inside W directly from the first two postulates. First we simulates $n \sim$Poisson$(\lambda\nu(B))$. Second, simulate n points uniformly inside W (that is, simulate $\tilde{x} \sim X_{\text{bin}(W,n)}$). Methods for simulating a Poisson random variate can be found in Stoyan et al (1995). The approach we adopt is the inverse transform method (see for example Fishman, 1996. p155) coupled with the approximation $n \approx [\lambda\nu(W) + \text{Normal}(0, \sqrt{\lambda\nu(W)})]$ when $\lambda\nu(W) > 20$.

2.6 Inhomogeneous Poisson point process

One of the most simple alternatives to the homogeneous Poisson point process is the *inhomogeneous Poisson point process*. The *inhomogeneous Poisson point process* is

obtained by replacing the intensity λ by a spatially varying density $\lambda(\zeta)$. Let Λ be a diffuse Radon measure on $\mathbb{R}^d$. An *inhomogeneous Poisson point process* is a point process possessing the following two properties

(i) The number of events in a bounded Borel set B has a Poisson distribution with mean $\Lambda(B)$

$$\mathrm{P}(X(B) = n) = \frac{(\Lambda(B))^n}{n!}\exp\left(-\Lambda(B)\right), \quad \text{for} \quad n \in \{0, 1, 2, \ldots\}$$

(ii) The number of points in k disjoint Borel sets form k independent random variables.

If the Radon measure Λ has a density with respect to Lebesgue measure then it can be written as

$$\Lambda(B) = \int_B \lambda(\eta) d\eta$$

for Borel sets B. We call the density $\lambda(.)$ the intensity function of the inhomogeneous Point process.

An inhomogeneous Poisson spatial point process in a compact set W can be simulated by using the rejection algorithm of Lewis & Shedler (1979). The routine for an inhomogeneous Poisson process with intensity function $\lambda(\xi)$ is as follows: Simulate a homogeneous spatial Poisson point process of intensity $\lambda_{\max}$, where $\lambda_{\max} = \sup\{\lambda(\xi) : \xi \in W\}$. Then independently delete each point ξ with probability

$$1 - \frac{\lambda(\xi)}{\lambda_{\max}} \tag{2.7}$$

The retained points form a realisation of events from an inhomogeneous Poisson process with intensity function $\lambda(\xi)$.

2.7 Estimating intensity

Estimation of many of the point process functions discussed in this document rely on the estimation of the intensity of a stationary point process. Given a sampling region W, a natural unbiased estimator for the intensity of a stationary point process is

$$\hat{\lambda} = \frac{x(W)}{\nu(W)}.$$

The intensity function $\lambda(.)$ of an inhomogeneous Poisson process can also be estimated using parametric or non-parametric methods such as kernel methods (see for example Cressie, 1991, section 8.5.1). An example of an edge-corrected kernel estimator for $\lambda(\eta)$ is

$$\hat{\lambda}(\eta) = \frac{1}{p_h(\eta)} \sum_{\zeta \in X \cap W} h^{-d} \kappa_d \left(\frac{\eta - \zeta}{h} \right)$$

for sampling window $W \subset \mathbb{R}^d$ and bandwidth h. The d-dimensional kernel function $\kappa_d(.)$ is any probability density function symmetric about the origin. The function

$$p_h(\eta) \equiv \int_W h^{-d} \kappa_d \left(\frac{\eta - \xi}{h} \right) \nu(d\xi) \tag{2.8}$$

is an edge-correction. A small investigation involving simulating an inhomogeneous Poisson point process and attempting to estimate the intensity function using this technique only confirmed that choosing an appropriate value for h is a non-trivial task (see Chapter 5). Indeed, this is a general problem with kernel density estimation (Wand & Jones, 1995). There are few examples where the choice of h is discussed for spatial point process applications. The work of Beil et al (2005) is one of the few that we have encountered, which explicitly states an appropriate bandwidth that was used for analysing cell biology data.

The proceeding sections provide a background on the spatial point process functions currently in use and some of their estimators. All the functions considered are 'distance-based' functions (they are functions of the distance between points) and most of their estimators, as is the case for many estimators in spatial statistics, are not unbiased but instead are ratios of two unbiased consistent estimators (Baddeley & Gill 1997). A higher-order spatial point process function is developed in the next chapter. This function has uses in general point process analysis. In Chapter 5 we define a point process function that is designed specifically for a particular biological problem. In that case, however, the function may be of particular use in situations where the position of events is known or suspected to be influenced by the position of the boundary of W.

2.8 Edge effects

Estimating point process functions of interest in some bounded compact region W, which we refer to as a *sampling* or *observation window*, is not trouble-free. Problems

generally encountered are those arising from *edge effects*. *Edge effects* is a term given to estimation problems created by not being able to observe data outside the edges of the observation region. They play a big (and hindering) role in the analysis of point patterns. They are usually encountered when the region W on which the point pattern is observed is a subset of a larger region on which the process is defined. Therefore estimation of summary statistics is biased by having censored events which may be interacting with events in the observation window. The methods of dealing with edge effects can be split into three categories, (the simplest being) the use of border methods (see for example Diggle, 2003), using estimators that explicitly account for edge effects, and wrapping W into a torus by identifying opposite edges. Our interest is mainly in the first two approaches, as the toroidal wrapping technique does not generally apply to the confocal microscopy data analysed throughout.

2.9 The empty space function $F(r)$

Let X be an almost surely stationary and isotropic point process. That is, all probability distributions associated with X are invariant under rotation and translation (Baddeley et al, 1993). The *empty space function* of X, denoted $F(r)$, or F from now onwards, is the probability distribution of the distance from an arbitrary point to the nearest event. That is for $r \geq 0$

$$\begin{aligned} F(r) &= \mathrm{P}(\rho(o, X) \leq r) \\ &= \mathrm{P}(X(b^d(o, r)) > 0) \end{aligned}$$

where

$$\rho(\eta, A) = \inf\{\|\eta - \zeta\| : \zeta \in A\} \tag{2.9}$$

is the shortest (Euclidean) distance from η to A. The empty space function of a homogeneous Poisson process with intensity λ is given by

$$F(r) = 1 - \exp\left(-\nu(b^d(o, r))\lambda\right) \tag{2.10}$$

Using (2.10) and by estimating the empty space function of a point pattern we can spot whether there is regularity or aggregation (clustering) in a point pattern. Estimated values of $F(r)$ greater than that given by (2.10) suggests that there is regularity, while lower values suggest aggregation (Baddeley et al, 1993).

Baddeley et al (1993) state that the empty space function is typically estimated

by taking a fine grid in the sampling region W and computing the distance from each grid point to the nearest event. This technique results in edge effects as we are unable to search for points outside W. The only approach currently in use is the border method (Baddeley et al, 1993), detailed in Chapter 4. When adopting this technique, only events that are at least a distance r from the boundary of W are considered. That is

$$\widehat{F}(r) = \frac{\nu\left(W_{\ominus r} \cap \{\cup_{\eta \in X} b^d(\eta, r)\}\right)}{\nu(W_{\ominus r})}$$

where

$$W_{\ominus r} = \{\zeta \in W : b^d(\zeta, r) \subset W\}$$

is the erosion of W. Kaplan-Meier estimation of distance distributions is discussed by Gill (1994) and Baddeley & Gill (1997). For a stationary point process X, the Kaplan-Meier estimator, $\widehat{F}_{Kap}(r)$ for the empty space function of X is given by

$$\begin{aligned} \hat{\Lambda}(t) &= \int_0^t \frac{\mathcal{H}_{d-1}\left(\partial(X_{\ominus r}) \cap W_{\ominus r}\right)}{\nu\left(W_{\ominus r} \backslash X_{\oplus r}\right)} dr \\ 1 - \widehat{F}_{Kap}(t) &= \exp(-\hat{\Lambda}(t)) \end{aligned}$$

where $\mathcal{H}_d(.)$ is the d-dimensional Hausdorff measure (see Appendix A). Furthermore, ∂A denotes the boundary of A and $A_{\ominus r} = \{\eta \in \mathbb{R}^d : \rho(\eta, A) \leq r\}$ is the dilation of A. The geometrical interpretation of the Kaplan-Meier estimator for the empty space function is, within the eroded window $W_{\ominus r}$, we consider the region outside the union of balls of radius r centred at points of the process X, and the surface of this union of balls. By considering a range of r, the quantity $\hat{\Lambda}(t)$ is an estimate of the probability that an event falls within distance t of another event.

2.10 The nearest neighbour distribution function $G(r)$

The $G(r)$-function is the distribution of the distance from a *typical event* of the process to the nearest other point of the process. As done by Stoyan & Stoyan (1994) and Stoyan et al (1995) we often use the term *typical event* in place of "randomly chosen event". For stationary point process X, the $G(r)$-function associated with X is given by

$$\begin{aligned} G(r) &= \mathrm{P}(\rho(o, X \backslash \{o\}) \leq r | o \in X) \\ &= \mathrm{P}(X(b^d(o, r)) > 1 | o \in X) \quad r \geq 0 \end{aligned}$$

By stationarity the point o can be replaced by any arbitrary point η. An alternative definition of the $G(r)$-function using the Campbell-Mecke theorem (see Stoyan et al (1995), section 4.4) is

$$G(r) = \frac{E\left(\sum_{\eta \in X \cap W} I_{(\rho(\eta, X\setminus\{\eta\}) \leq r)}\right)}{E(X(W))}$$

for arbitrary W with $0 < \nu(W) < \infty$. For a homogeneous Poisson process with intensity λ the $G(r)$-function is given by

$$G(r) = 1 - \exp\left(-\lambda \nu(b^d(o, r))\right)$$

A border estimate for the nearest neighbour distribution function is given by

$$\hat{G} = \frac{\sum_{\eta \in X} I_{(\rho(\eta, X\setminus\{\eta\}) \leq r)} I_{(\eta \in W_{\ominus \rho(\eta, X\setminus\{\eta\})})}}{\sum_{\eta \in X} I_{(\eta \in W_{\ominus \rho(\eta, X\setminus\{\eta\})})}}$$

2.11 The pair correlation function $g(r)$

The *pair correlation function*, $g(r)$, is the frequency of event pairs within distance r. The *pair correlation function* is widely used in spatial statistics and particularly in astronomy and astrophysics, for example Kerscher (1998). Provided that the second-order product density exists, then in the stationary and isotropic case we can write $\varrho^{(2)}(\eta, \zeta) \equiv \varrho^{(2)}(r)$ for $r = \|\eta - \zeta\|$. Also assuming that the intensity function exists, then the *pair correlation function* is defined for a stationary point process by

$$g(r) = \frac{\varrho^{(2)}(r)}{\lambda^2} \tag{2.11}$$

For a Poisson process, we have $g(r) = 1$. Furthermore, $g(r) > 1$ indicates clustering while $g(r) < 1$ is a sign of regularity.

The pair correlation function can be estimated using

$$\hat{g}(r) = \frac{\hat{\varrho}^{(2)}(r)}{\hat{\lambda}^2}$$

and an estimator $\hat{\varrho}^{(2)}(r)$ for the second-order product density is

$$\hat{\varrho}^{(2)}(r) = \sum_{\eta, \zeta \in X \cap W} \frac{\kappa_1(\|\eta - \zeta\| - r)}{\mathcal{H}_d(b^d(o, r))\nu(W_\eta \cap W_\zeta)}$$

where κ_1 is some suitable kernel such as the Epanechnikov kernel $\kappa_{1,h}$ given by

$$\kappa_{1,h}(s) = \frac{3}{4h}\left(1 - \frac{s^2}{h^2}\right) I_{(s\in(-h,h))}$$

Stoyan (2007) has stated that the best kernel function is the box kernel given by

$$\kappa_h(s) = \frac{I_{(s\in[-h,h])}}{2h}$$

2.12 The $K(r)$-function

The $K(r)$-function appears at present to be the most popular second-order characteristic used in point process analysis. Heuristically, for a stationary point process with intensity λ, $\lambda K(r)$ is the mean number of events that are within distance r of the typical event,

$$K(r) = \frac{1}{\lambda} E(X(b^d(o,r))), \tag{2.12}$$

where the expectation is with respect to the reduced Palm distribution. The *Campbell-Mecke theorem* yields the alternative definition

$$K(r) = \frac{E\left(\sum_{\eta\in X\cap B} X(b^d(\eta,r)\backslash\{\eta\})\right)}{E(X(B))}$$

for arbitrary B with $0 < \nu(B) < \infty$. For a homogeneous Poisson process with intensity λ the $K(r)$-function of X is given by

$$K(r) = \nu(b^d(o,r))$$

A border correction estimate of $K(r)$ is

$$\hat{K}(r) = \frac{\nu(W_{\ominus r})}{X(W_{\ominus r})^2} \sum_{\eta\in W_{\ominus r}} \sum_{\zeta\in W} I_{(0<\|\eta-\zeta\|\leq r)}.$$

Other edge-corrected estimators for the $K(r)$-function are detailed in Chapter 4.

2.13 Relationships between spatial point process functions

The K- and g-functions are closely related. K can be expressed in terms of g by the equation

$$K(r) = \mathcal{H}_d(b^d(o,1)) \int_0^r u^{d-1} g(u) du \tag{2.13}$$

Some other characteristics have been defined as combinations and variants of those discussed. Of particular importance is the J-function, suggested by Lieshout & Baddeley (1999). For a stationary point process the $J(r)$-function (J-function) is defined as

$$J(r) = \frac{1 - G(r)}{1 - F(r)}$$

for $F(r) < 1$. Besides some desirable properties, the $J(r)$-function is useful because $J(r) = 1$ for a homogeneous Poisson process. However, $J(r) = 1$ does not imply that the point process is a homogeneous Poisson. $J(r)$ can be estimated by using

$$\hat{J}(r) = \frac{1 - \hat{G}(r)}{1 - \hat{F}(r)}$$

In general, for $r > 0$, $J(r) < 1$ indicates clustering and $J(r) > 1$ is a sign of regularity. An alternative to the K-function which is frequently used is the $L(r)$-function which is defined as

$$L(r) = \left(\frac{K(r)}{\nu(b^d(o,1))} \right)^{\frac{1}{d}}.$$

A natural estimator for $L(r)$ is

$$\hat{L}(r) = \left(\frac{\hat{K}(r)}{\nu(b^d(o,1))} \right)^{\frac{1}{d}}.$$

The $L(r)$-function is popular amongst statisticians because $L(r) - r = 0$ for a homogeneous Poisson process.

2.14 Third-order characteristics

Third-order functions are relatively uncommon as they tend to be computationally expensive to estimate. Analytical expressions for third-order characteristics of common processes can also be mathematically difficult to obtain. Schladitz & Baddeley (2000) introduced the $T(r)$-function, which they define as the expected number of r-close event pairs within a ball of radius r centred at the typical event. Formally, let X be a point process with intensity λ. The $T(r)$-function of X is defined as

$$T(r) = \frac{E\left(\sum_{\eta,\zeta \in X \cap b^d(o,r)} I_{(0<\|\eta-\zeta\|\leq r)}\right)}{2\lambda^2} \quad r \geq 0$$

where the expectation is with respect to the reduced Palm distribution. Stoyan & Stoyan (1994, section 14.5) give a good account on characteristics of third-order and higher. Particular examples of third-order characteristics that they mention include the $z_B(r)$-function defined for a stationary process with intensity λ by

$$z_B(r) = \frac{1}{\lambda^3 \nu(B)} \int \frac{I_{(\xi \in B)} \varrho^{(3)}(r, \xi)}{\varrho^{(2)}(r)} d\xi \quad r \geq 0$$

and an extended version of the $z_B(r)$ given by

$$Z_B(r_1, r_2) = \frac{1}{\lambda^3 \nu(B)} \int_{r_1}^{r_2} \int \frac{I_{(\xi \in B)} \varrho^{(3)}(r, \xi)}{K(r_2) - K(r_1)} d\xi dr,$$

where instead of just considering event pairs of distance r, an average over a certain range of values is taken. An estimator for $z_B(r)$ is given in Stoyan & Stoyan (1994, section 15.5). Estimation of the $T(r)$-function is discussed by Schladitz & Baddeley (2000, section 4).

The pair correlation and K-functions can be defined for the non-stationary case, see Møller & Waagepetersen (2004). The anisotropic versions of these functions are defined by Stoyan & Stoyan (1994). Baddeley et al (2000) propose definitions for the non-stationary versions of the F- and G-function in their concluding discussions on open problems for the statistical analysis of inhomogeneous point patterns.

2.15 More complicated point process models

Earlier we discussed the simplest point process, the homogeneous Poisson process. We can divide some of the more complicated point process models that we discuss throughout this thesis into three categories, inhomogeneous Poisson models, and models that allow for interaction between events and are designed to either capture regularity, or alternatively, mimic the features of a spatial point pattern that exhibits clustering. The exception to this classification are Cox processes, an important class of models that can be used to model both clustering and regularity (see for example Møller & Waagepetersen, 2004, chapter 5).

Our initial tests for CSR in Chapter 1 provided some possible evidence for clustering and hence our discussions here are favoured towards models for clustered data. The cluster models we discuss briefly include the Matérn cluster process (see for example Cressie, 1991). This model has, for example, been used for modelling

tree roots data (Fleischer et al, 2006). We also consider the Gauss-Poisson process (Stoyan et al, 1995). Point process models tend to be generalisations of other point process models. For example the homogeneous Poisson process is a special case of the inhomogeneous one, which can be generalised to a Cox process.

Models can also be formed by the three fundamental operations discussed by Stoyan et al (1995, section 5.1). These operations include superposition, thinning and clustering. In a clustering operation the events of a point process are replaced by clusters of points, X_0. The clusters (X_0s) themselves are spatial point processes. The union of these clusters form a cluster point process. It is common practice to refer to the events as "parents" and the events of the clusters as "daughters". The two cluster processes we discuss here are members of a group of processes called Neyman-Scott processes (see for example Stoyan et al, 1995). Neyman-Scott processes result from homogeneous independent clustering applied to a stationary Poisson process. Some Neyman-Scott process such as the Matérn cluster process are also Cox processes.

2.15.1 Gauss-Poisson process

A *Gauss-Poisson* process (Newman, 1970) is an example of a Poisson cluster process and is often used in spatial statistics (Stoyan et al (1995)). The parent points have a homogeneous Poisson distribution with intensity λ and the number of daughters of each parent is zero, one or two with probability q_0, q_1, and q_2 respectively. If the parent has one daughter then the daughter is placed at the parent location. If the parent has two daughters then one is placed at the parent and the other is placed randomly at (fixed) distance s from the first daughter. The resulting pattern only includes daughter points (and hence the parent points are deleted). Some further results for Gauss-Poisson processes can be found in Milne & Westcott (1972).

2.15.2 Matérn cluster process

Matérn's cluster process consists of parents that come from a homogeneous Poisson point process with intensity λ_p. Each parent has m daughters which are uniformly distributed inside $b^d(o, R)$ (with the parent point being regarded as the origin). The parameter m comes from a Poisson distribution with intensity λ_m. Implicit expressions for the K- and g-function for a Matérn cluster process can be found in Stoyan et al (1995). In Chapter 4 we derive an explicit expression for the K-functions of a Matérn cluster process.

The Matérn cluster process and Gauss-Poisson process can be simulated in the compact window W directly via the model definitions, although care should be taken with regards to edge effects. A simple way to account for edge effects is to simulate the parent points inside the dilated window $W_{\oplus R}$ where R is such that for the $P(X_0 \supset b^d(o, R))$ is very small or zero (Stoyan et al, 1995). Brix & Kendall (2002) discuss the simulation of cluster point processes without edge effects.

2.15.3 Markov point processes

Markov or Gibbs point processes have been intensively used in spatial statistics since 1970 (Stoyan & Stoyan, 1994). Although they are models for various types of point patterns, they are usually recognised for their ability to provide a more flexible framework for modeling spatial point patterns that exhibit inhibition (compared to a homogeneous Poisson distribution) (Cressie, 1995). Markov point processes were first defined by Ripley & Kelly (1977). As redefined by Cressie (1995), a spatial point process on bounded set $V \subset \mathbb{R}^d$ is said to be *Markov of range* ρ if it is a spatial point process that has conditional intensity at $\xi \in V$ given the realisation of the process in $A \backslash \xi$ that depends only on the events in $b^d(\xi, \rho) \backslash \{\xi\}$. Each Markov process is characterised by a likelihood ratio $f(.)$ with respect to a unit intensity Poisson process. Furthermore, $f(.)$ is usually defined up to a normalising constant that cannot be evaluated in closed form (Diggle, 2003). A popular example of a Markov point process is the Strauss process (Strauss, 1975). In this case, for a configuration of $n < \infty$ points x, we have

$$f(x) = \alpha \beta^n \gamma^{\phi(R)}, \quad \beta > 0, 0 \leq \gamma \leq 1, R > 0$$

Where $\phi(R)$ is the number of distinct pair of events within distance r. The *Papangelou conditional intensity* defined by

$$\lambda^*(x, \eta) = \frac{f(x \cup \eta)}{f(x)}, \quad \eta \in V \backslash x$$

where we take $a/0 = 0$ for $a \geq 0$ (Kallenberg, 1984) is a fundamental characteristic (Møller & Waagepetersen, 2001). If f is hereditary (that is $f(x) > 0 \Rightarrow f(y) > 0$ for $y \subset x$), then there is a one-to-one correspondence between f and λ^*. Distribution characteristics (such as the summary statistics introduced earlier) for Markov models are difficult to calculate (Stoyan & Stoyan, 1994). Further theory on Markov point process can be found in Stoyan et al (1995) while a good exposition on simulating Markov point processes can be found in Moller & Waagepetersen (2003, 2004).

2.15.4 Cox processes

A Cox process is a natural approach for generalising the definition of Poisson point process (Møller & Waagepetersen, 2004). A Cox Process on $V \subset \mathbb{R}^d$ is often referred to as a 'doubly stochastic' Poisson point process as the intensity measure is replaced by a random locally finite measure Z_Λ. More formally, we say that a point process X is a Cox process driven by Z_Λ if $X|Z_\Lambda = \Lambda$ is an inhomogeneous Poisson process with mean measure Λ. Due to their generality and associated manageable closed form calculations, Cox processes tend to find important applications as stochastic models (Stoyan et al, 1995). Examples of Cox processes include the Matérn cluster process. A particular class of Cox processes that is appealing from both an applied and theoretical perspective is the wealthy class of log-Gaussian Cox processes (see Chapter 5). A detailed account of Cox processes can be found in Moller & Waagepetersen (2002).

Chapter 3

The Q-function

An important part of point pattern analysis involves estimating one or more of the summary statistics based on inter-point distances, such as the $F(r)$- or $G(r)$-function, or second-order characteristics such as the $K(r)$-function (see Chapter 2). There are numerous examples, including König et al (1991), Baddeley et al (1993) and Diggle (1986) where these functions have been used to analyse real data. These functions share similar limitations including the requirement of stationarity and isotropy. One of the most recognisable problems involves estimation issues arising from edge effects.

New characteristics such as the $J(r)$-function have been introduced as improved alternatives. Van Lieshout & Baddeley (1999), and Baddeley et al (2000) have shown that amongst other desirable properties, the $J(r)$-function is less sensitive to edge effects. Definitions of some of the currently used point process functions have also been generalised in order to increase their utility. For example, Møller & Waagepetersen (2004) describe how a directional version of the $K(r)$-function can be constructed. This extended version can be used to investigate for possible anisotropy. Baddeley et al (2000) have gone further by defining a version of the $K(r)$-function for analysing inhomogeneous point patterns.

The $K(r)$-function has gained much attention for its ability to capture second-order behaviour. However, Baddeley & Silverman (1984) constructed a point process different from a homogeneous Poisson process but having the same $K(r)$-function as a homogeneous Poisson process. In response, third-order characteristics such as Schladitz and Baddeley's (2000) $T(r)$-function have been defined. These third-order characteristics are able to reveal features of a point pattern which the $K(r)$-function

would fail to reveal.

In order to provide a more complete description of the distribution of a point pattern, we introduce a new characteristic, the $Q(r, \tilde{r})$-function which is a generalisation of Schladitz & Baddeley's $T(r)$-function. Instead of considering the mean number of r-close pairs in a ball of radius r centred at the typical point, the Q-function examines the expected number of $\tilde{r}$-close pairs in a ball of radius r centred at the typical point, and by so doing represents a point process characteristic depending on two inter-point distance variables.

In the next section we provide a formal definition of the $Q(r, \tilde{r})$-function. Section 3.2 provides the form of the $Q(r, \tilde{r})$-function for the homogeneous Poisson process. We also provide the closed form solution for the $Q(r, \tilde{r})$-function for a Gauss-Poisson process and provide a brief discussion on thinned processes. In section 3.3 we discuss estimation. In the final section of this chapter we include practical applications and simulation studies. We show that the $Q(r, \tilde{r})$-function can be a more powerful tool compared to the $K(r)$- and $T(r)$-functions in certain hypothesis tests. Furthermore, Grabarnik & Chiu (2002) have introduced a goodness-of-fit test statistic, Q^2, that they show through simulation studies to be more sensitive to mixtures of regular and clustered spatial point processes compared to tests that utilise the nearest neighbour distance distribution, the second- and certain third-order characteristics. Unfortunately, however, practical applications of tests based on the Q^2 statistic are not straightforward and some parameters need to be chosen arbitrarily (Grabarnik & Chiu, 2002). By carrying out similar experiments to Grabarnik & Chiu (2002), we show that similar tests for CSR on mixed regular and clustered spatial point patterns based on using the $Q(r, \tilde{r})$-function provides results that are competitive with the Q^2 statistic. In Chapter 5 we analyse the spatial distribution of PML NBs using the Q-function.

3.1 The Q-function

In the definition of the $Q(r, \tilde{r})$-function, as for the $T(r)$- and $K(r)$-functions, we condition on there being a point at the origin, o. As we are conditioning on an event of zero probability, we adopt the idea of a Palm distribution, or in our case, as we do not count the point at the origin, the reduced Palm distribution (see Chapter 2); we do this so that the conditional probability is unequivocal. We formally define the $Q(r, \tilde{r})$-function in terms of an expectation with respect to the reduced Palm

distribution as follows:

Definition 2 *Let X be a stationary point process with finite intensity $\lambda > 0$. The $Q(r, \tilde{r})$ function of X is given by*

$$Q(r, \tilde{r}) = (2\lambda^2)^{-1} E_o^! \left(\sum_{\eta,\zeta \in X \cap b^d(o,r)} d_{\tilde{r}}(\eta, \zeta) \right) \tag{3.1}$$

for $r, \tilde{r} \geq 0$.

Here $E_o^!$ denotes expectation with respect to $P_o^!$ and $d_{\tilde{r}}(\eta, \zeta)$ represents the function taking values 1 if $0 < \|\eta - \zeta\| \leq \tilde{r}$ and 0 otherwise.

By the Campbell-Mecke theorem (see Stoyan et al (1995), section 4.4), for any non-negative measurable function f on $\mathbb{R}^d \times N$

$$E\left(\sum_{\zeta \in X} f(\zeta, X) \right) = \lambda \int \int f(\zeta, x_\zeta) P_o(dx) d\zeta. \tag{3.2}$$

We can use (3.2) and stationarity assumptions to show that $Q(r, \tilde{r})$ can be written in the form

$$Q(r, \tilde{r}) = (2\nu(B)\lambda^3)^{-1} E\left(\sum_{\zeta \in X \cap B} \sum_{\eta,\xi \in X} d_r(\eta, \zeta) d_{\tilde{r}}(\eta, \xi) d_r(\xi, \zeta) \right) \tag{3.3}$$

for any Borel set B with $0 < \nu(B) < \infty$. A particular justification is as follows:

$$\begin{aligned}
& E\left(\sum_{\zeta \in X \cap B} \sum_{\eta,\xi \in X} d_r(\eta, \zeta) d_{\tilde{r}}(\eta, \xi) d_r(\xi, \zeta) \right) \\
= \; & E\left(\sum_{\zeta \in X \cap B} \sum_{\eta,\xi \in X \setminus \{\zeta\} \cap b^d(\zeta, r)} d_{\tilde{r}}(\eta, \xi) \right) \\
= \; & E\left(\sum_{\zeta \in X \cap B} \sum_{\eta,\xi \in X \setminus \{o\} \cap b^d(o,r)} d_{\tilde{r}}(\eta - \zeta, \xi - \zeta) \right) \\
& \text{(using the Campbell-Mecke theorem)} \\
= \; & \lambda \int \int \sum_{\eta,\xi \in x_\zeta \setminus \{o\} \cap b^d(o,r)} d_{\tilde{r}}(\eta - \zeta, \xi - \zeta) P_o(dx) d\zeta \\
= \; & \lambda \nu(B) \int \sum_{\eta,\xi \in x_\zeta \cap b^d(o,r)} d_{\tilde{r}}(\eta - \zeta, \xi - \zeta) P_o^!(dx)
\end{aligned}$$

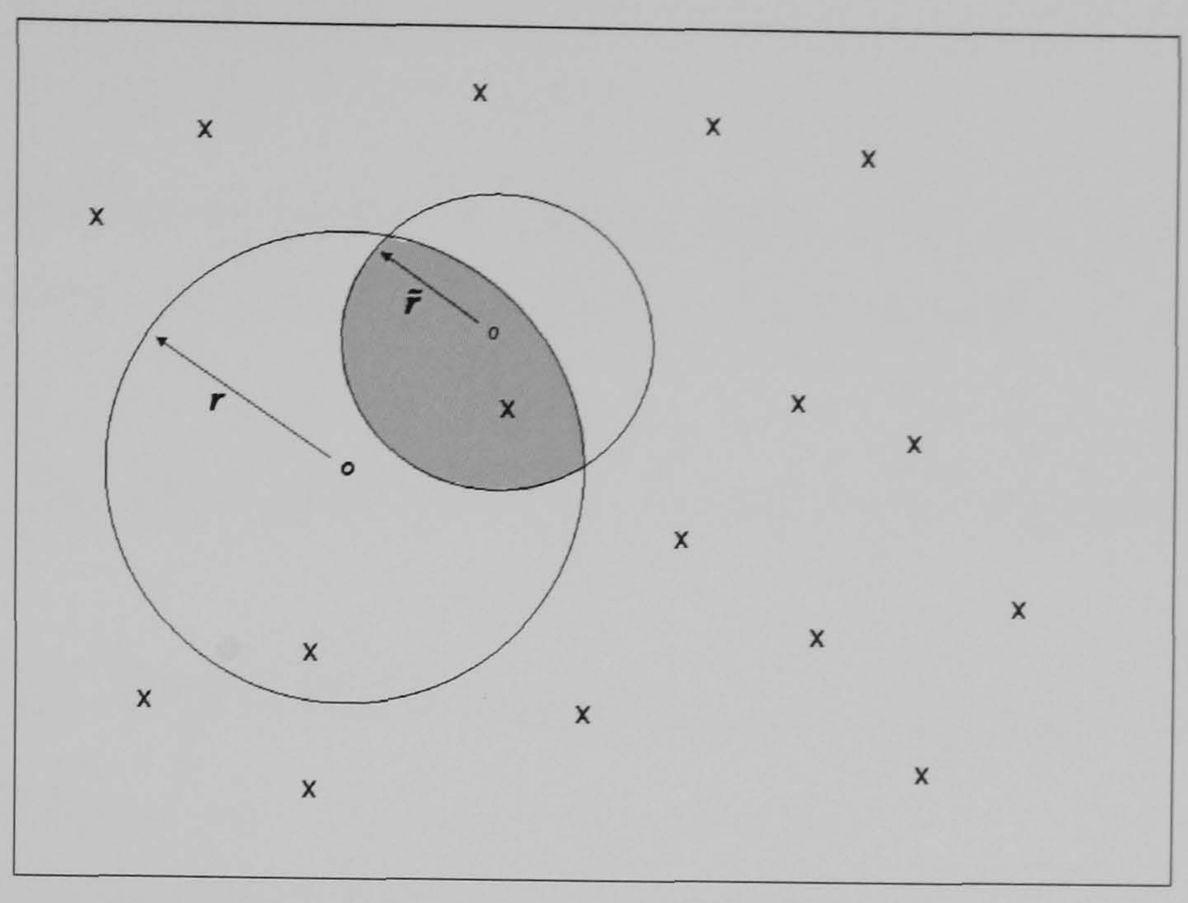


Figure 3.1: An $\tilde{r}$-close point pair within distance r of an arbitrary point

$$= \lambda\nu(B)E_o^! \left(\sum_{\eta,\xi \in X \cap b^d(o,r)} d_{\tilde{r}}(\eta,\xi) \right)$$

$$= 2\lambda^3\nu(B)Q(r,\tilde{r})$$

Figure 3.1 illustrates what is meant by an $\tilde{r}$-close point pair in the r neighbourhood of an arbitrary point. The grey region emphasises the points that are candidates to form a triple with two points within distance r of each other.

3.2 The Q-function of some point processes

Having formally defined the $Q(r,\tilde{r})$-function, we can consider its theoretical value for some standard point processes. Here we provide the theoretical $Q(r,\tilde{r})$-function, $Q_{\text{poi}}(r,\tilde{r})$, for a homogeneous Poisson process in the planar ($d = 2$) and spatial ($d = 3$) cases. The theoretical value for the $Q(r,\tilde{r})$-function of the Gauss-Poisson process and *p-thinned* process are also given.

3.2.1 The Q-function for a spatial homogeneous Poisson process

Making use of Slivnyak's theorem (see Stoyan et al (1995), section 4.4.7), which states that for a Poisson process with distribution P and reduced Palm distribution

$P_\varsigma^!$, we have

$$P_\varsigma^! = P, \tag{3.4}$$

and following Schladitz & Baddeley (2000), using (3.4) and the definition of the factorial kth moment measure $\alpha^{(k)}$, with $k = 2$ (see Chapter 2), we can show the following result:

Proposition 1 *The $Q(r,\tilde{r})$-function of a homogeneous Poisson process in $\mathbb{R}^2$ is given by*

$$\begin{aligned}
& \frac{r^2{}_{min}\pi^2(r_{max} - r_{min})^2}{2} + \pi r^2\{A_1(r,r,\tilde{r}) - A_1(r_{max} - r_{min}, r, \tilde{r})\} \\
+ \quad & \pi\tilde{r}^2\{A_1(r,\tilde{r},r) - A_1(r_{max} - r_{min}, \tilde{r}, r)\} \\
+ \quad & \pi r^2\{A_2(r,r,\tilde{r}) - A_2(r_{max} - r_{min}, r, \tilde{r})\} \\
+ \quad & \pi\tilde{r}^2\{A_2(r,\tilde{r},r) - A_2(r_{max} - r_{min}, \tilde{r}, r)\} \qquad \tilde{r} < 2r,
\end{aligned} \tag{3.5}$$

and

$$\frac{\pi^2 r^4}{2} \qquad \tilde{r} \geq 2r.$$

where

$$\begin{aligned}
A_1(t,q,p) &= \frac{t^2}{2} cos^{-1}\left(\frac{t^2+q^2-p^2}{2qt}\right) - \frac{\sqrt{u(t,p,q)}}{4} \\
&\quad + \frac{p^2 tan^{-1}\left(\frac{t^2-p^2-q^2}{\sqrt{u(t,p,q)}}\right)}{2}
\end{aligned}$$

$$\begin{aligned}
A_2(t,q,p) &= \frac{\sqrt{u(t,q,p)}}{16}\left(\frac{3p^2}{q^2} - \frac{t^2}{q^2} - 1\right) \\
&\quad + \frac{1}{8} tan^{-1}\left(\frac{t^2-q^2-p^2}{\sqrt{u(t,q,p)}}\right)\left(-2p^2 - q^2 + \frac{p^4}{q^2}\right) \\
&\quad + \frac{ln\left(\frac{v(t,q,p)}{t^2}\right)}{8k(q,p)}\left(3q^2p^2 - 3p^4 - q^4 + \frac{p^6}{q^2}\right)
\end{aligned}$$

with

$$k(q,p) = \sqrt{-(q^2-p^2)^2}$$

$$u(t,p,q) = 4t^2q^2 - (t^2+q^2-p^2)^2$$

$$v(t,q,p) = 2(-(q^2-p^2)^2 + t^2(p^2+q^2) + k(q,p)\sqrt{u(t,q,p)}).$$

Furthermore, $r_{max} = sup\{r,\tilde{r}\}$ and $r_{min} = inf\{r,\tilde{r}\}$.

BIBL. LONDON UNIV.

Proof. for $\tilde{r} < 2r$ we have

$$\begin{aligned} Q(r,\tilde{r}) &= (2\lambda^2)^{-1} E_o^! \left(\sum_{\eta,\zeta \in X \cap b(o,r)} d_{\tilde{r}}(\eta,\zeta) \right) \\ &= (2\lambda^2)^{-1} \int_{b(o,r)} \int_{b(o,r)} d_{\tilde{r}}(\eta,\zeta) \alpha^{(2)}(d(\eta,\zeta)) \\ &= \frac{1}{2} \int_{b(o,r)} \int_{b(o,r)} d_{\tilde{r}}(\eta,\zeta) d\zeta d\eta \\ &= \frac{1}{2} \int_{b(o,r)} \nu(b(o,r) \cap b(\eta,\tilde{r})) d\eta \end{aligned}$$

(after switching to polar coordinates)

$$\begin{aligned} &= \frac{1}{2} \int_0^{2\pi} \int_0^{r_{max}-r_{min}} t\pi {r_{min}}^2 dt d\varphi \\ &+ \frac{1}{2} \int_0^{2\pi} \int_{r_{max}-r_{min}}^{r} t \left(\tilde{r}^2 \left[\cos^{-1} \left(\frac{t^2+\tilde{r}^2-r^2}{2\tilde{r}t} \right) - B(t,\tilde{r},r) \right] \right. \\ &+ \left. r^2 \left[\cos^{-1} \left(\frac{t^2+r^2-\tilde{r}^2}{2rt} \right) - B(t,r,\tilde{r}) \right] \right) dt d\varphi \end{aligned} \tag{3.6}$$

where

$$B(u,v,w) = \frac{(u^2+v^2-w^2)\sqrt{4u^2v^2-(u^2+v^2-w^2)^2}}{4u^2v^2}.$$

Using

$$\int_{r_{max}-r_{min}}^{r} t\cos^{-1}\left(\frac{t^2+q^2-p^2}{2qt}\right) dt = A_1(r,q,p) - A_1(r_{max}-r_{min},q,p) \tag{3.7}$$

and

$$\begin{aligned} &- \int_{r_{max}-r_{min}}^{r} \frac{(t^2+q^2-p^2)\sqrt{4t^2q^2-(t^2+q^2-p^2)^2}}{4tq^2} dt \\ &= A_2(r,q,p) - A_2(r_{max}-r_{min},q,p) \end{aligned} \tag{3.8}$$

yields the assertion. For $\tilde{r} \geq 2r$ the proof is the same except $\nu(b(o,r) \cap b(\eta,\tilde{r})) = \pi r^2$ and the result follows readily. ■

Consider two two-dimensional balls of radii r and $\tilde{r}$ with centres separated by distance t. Then $(r^2A_1(t_1, r, \tilde{r}) + A_2(t_1, r, \tilde{r})) - (r^2A_1(t_0, r, \tilde{r}) + A_2(t_0, r, \tilde{r}))$ is the total area of the circular segments (the shaded region in Figure 3.2) of the ball of radius r that intersects the ball of radius $\tilde{r}$, as t ranges from t_0 to t_1. Although

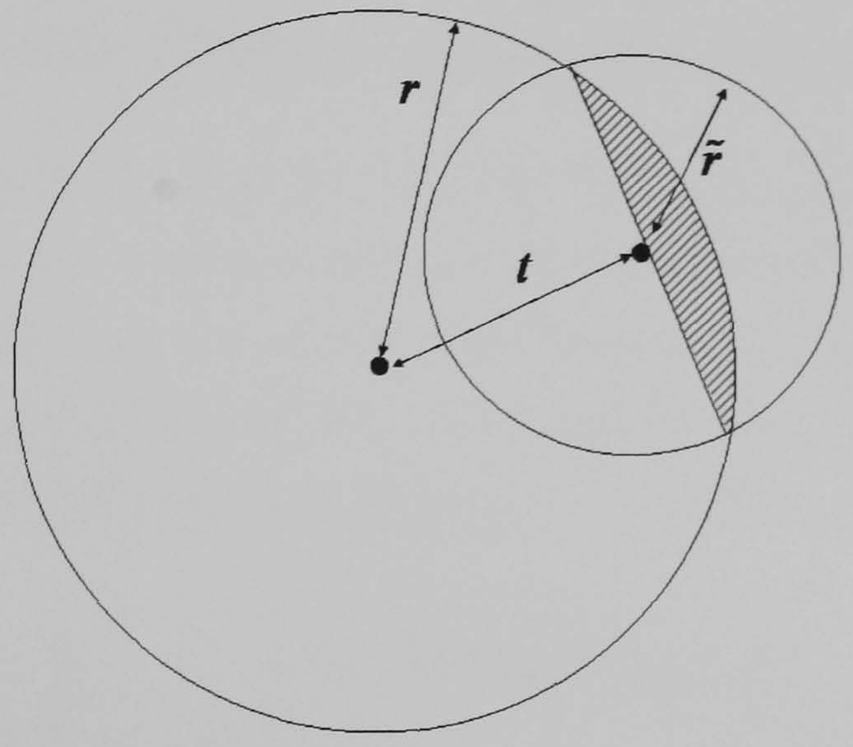


Figure 3.2: The geometrical interpretation of $r^2A_1(t, r, \tilde{r}) + A_2(t, r, \tilde{r})$

the algebraic expression for the $Q(r, \tilde{r})$-function of a homogeneous Poisson in $\mathbb{R}^2$ is somewhat complicated and requires numerical evaluation, it is (computationally) much more efficient to evaluate (3.5) compared to using simulation to obtain $Q(r, \tilde{r})$.

The $Q(r, \tilde{r})$-function of a homogeneous Poisson process takes a simpler form in the three-dimensional case (the reason for this being that the algebraic expression for the volume of a spherical cap is less complicated than that of the area of a circular segment), and is provided in the next proposition.

Proposition 2 *The $Q(r, \tilde{r})$-function of a homogeneous Poisson process in $\mathbb{R}^3$ is given by*

$$Q(r, \tilde{r}) = \begin{cases} \pi^2\tilde{r}^3\left(-\frac{1}{2}r^2\tilde{r} + \frac{1}{36}\tilde{r}^3 + \frac{8}{9}r^3\right) & \tilde{r} < 2r \\ \frac{8}{9}\pi^2 r^6 & \tilde{r} \geq 2r \end{cases} \tag{3.9}$$

Proof. As before, we have for $\tilde{r} < 2r$

$$\begin{aligned} Q(r, \tilde{r}) &= \frac{1}{2}\int_{b(o,r)} \nu(b(o, r) \cap b(\eta, \tilde{r}))d\eta \\ &= \int_0^{2\pi}\int_0^{\frac{\pi}{2}}\int_{r_{max}-r_{min}}^{r} \frac{t^2\pi(r + \tilde{r} - t)^2(t^2 + 2t\tilde{r} - 3\tilde{r}^2 + 2tr + 6\tilde{r}r - 3r^2)}{12t}\sin(\varphi)dtd\varphi d\theta \end{aligned}$$

$$+ \int_0^{2\pi} \int_0^{\frac{\pi}{2}} \int_0^{r_{max}-r_{min}} t^2 \frac{4}{3}\pi r_{min}^3 \sin(\varphi) dt d\varphi d\theta$$

evaluating the integrals gives the result, and the case $\tilde{r} \geq 2r$ follows as before. ■

3.2.2 The Gauss-Poisson process Q-function

Now consider a Gauss-Poisson process (see Chapter 2) having 0, 1, or 2 daughters with probability q_0, q_1, and q_2 respectively and separated by a distance s. Using the important result that the Palm distribution of a Poisson cluster process with intensity λ is the convolution of the Palm distribution c_o, of the cluster X_0, and the distribution P of the process (see Stoyan et al (1995), section 5.3.2), we have

$$2\lambda^2 Q(r,\tilde{r}) = \int \sum_{\eta,\zeta \in x \cap b^d(o,r)\setminus\{o\}} d_{\tilde{r}}(\eta,\zeta) P(dx) + 2\int\int \sum_{\substack{\eta \in x \cap b^d(o,r)\setminus\{o\} \\ \zeta \in y \cap b^d(o,r)\setminus\{o\}}} d_{\tilde{r}}(\eta,\zeta) P(dx) c_o(dy)$$
$$+ \int \sum_{\eta,\zeta \in y \cap b^d(o,r)\setminus\{o\}} d_{\tilde{r}}(\eta,\zeta) c_o(dy) \qquad (3.10)$$

The first term in (3.10) represents the number of candidate triples from the resulting cluster process. The second integral in (3.10) is the triples in which the typical, and one other event are members of the same cluster, and the third event is from a different cluster. The third term in (3.10) represents the number of candidate triples from the same cluster. Using similar arguments to Schladitz & Baddeley (2000) the integrals on the right hand side of (3.10) can be evaluated explicitly for a Gauss-Poisson process (see Appendix D for further detail) to obtain

$$Q(r,\tilde{r}) = Q_{\text{poi}}(r,\tilde{r}) + \frac{3q_2}{\lambda(q_1+2q_2)} I_{(r \geq s)} \nu(b^d(o,r) \cap b^d(\partial S_{b^d(o,s)}, \tilde{r}))$$

where $\partial S_{b^d(o,s)}$ is a point on the boundary of $b^d(o,s)$. An explicit expression for $\nu(b^d(o,r) \cap b^d(\partial S_{b^d(o,s)}, \tilde{r}))$ in the planar case when $r_{max} - r_{min} < s < \tilde{r} + r$ can be found in Stoyan & Stoyan (1994, Appendix K). For the $d = 3$ case we have

$$\nu(b^d(o,r) \cap b^d(\partial S_{b^d(o,s)}, \tilde{r})) = \pi(r + \tilde{r} - s)^2(s^2 + 2\tilde{r}s - 3\tilde{r}^2 + 2sr + 6r\tilde{r} - 3r^2)/12s.$$

For $s \leq r_{max} - r_{min}$ and $s \geq r + \tilde{r}$ the value of $\nu(b^d(o,r) \cap b^d(\partial S_{b^d(o,s)}, \tilde{r}))$ is trivial.

3.2.3 The Q-function of a p-thinned process

Now consider the $Q(r, \tilde{r})$-function for regular point patterns obtained by thinning. For point process X, the *thinned* point process $X_{th} \subset X$ is obtained by deleting the points of X according to some definite rule. In $p(\zeta)$-*thinning*, a point $\zeta \in X$ is deleted with probability 1-$p(\zeta)$, where $p(\zeta)$ is a deterministic function on $\mathbb{R}^d$, with $0 \leq p(\zeta) \leq 1$. The thinning process can be generalised to choosing p at random (see for example Stoyan et al (1995) , section 5.1). This type of thinning, where there is no interaction between points is known as independent thinning. The class of dependent thinnings is obtained by allowing dependence on the configuration of X.

In general, if a characteristic of a Poisson process X_{poi} with intensity measure Λ is known, then it is straightforward to obtain the characteristic for the process obtained by independent thinning. A useful result is that the process obtained by the $p(\zeta)$-thinning of X_{poi} is also a Poisson process with intensity measure Λ_{th} given by

$$\Lambda_{th}(B) = \int_B p(\zeta)\Lambda(d\zeta)$$

In particular, for the special case of p-thinning, where each point of X has probability $1 - p$ of being deleted, we have the following result

Lemma 3 *If X is a stationary Poisson process with intensity λ then the p-thinned process of X has $Q(r, \tilde{r})$-function, $Q_{th}(r, \tilde{r})$, given by*

$$Q_{th}(r, \tilde{r}) = Q_{poi}(r, \tilde{r})$$

Proof. The result follows since the p-thinning of X results in a stationary Poisson process with intensity $p\lambda$. ■

Note that the $Q(r, \tilde{r})$-function of point patterns obtained by dependent thinning, such as Matérn hardcore processes, is non-trivial to obtain.

3.3 Estimating Q

We consider estimation of the $Q(r, \tilde{r})$ function in a compact window $W \subset \mathbb{R}^d$ $(d = 2, 3)$. Like other popular summary statistics, we are faced with the problem of edge effects arising from unknown points outside of W. The easiest and one of the most widely used edge corrections is the border method (see for example Moller &

Waagepetersen (2004), Reed & Howard (1997)). The border edge-correction estimate for the $Q(r,\tilde{r})$ function, $\hat{Q}_{\text{border}}(r,\tilde{r})$ using (3.3) is

$$\hat{Q}_{\text{border}}(\tilde{r},r) = \frac{\nu(W)^2}{2X(W_{\ominus r})X(W)(X(W)-1)} \sum_{\zeta \in X \cap W_{\ominus r}} \sum_{\eta,\xi \in X} d_r(\eta,\zeta) d_{\tilde{r}}(\eta,\xi) d_r(\xi,\zeta)$$

and is defined whenever $X(W_{\ominus r}) > 0$ and $X(W) > 1$. Estimating the $Q(\tilde{r},r)$ function is in general harder than estimating the $K(r)$-function. The computation of the border estimate for the $Q(\tilde{r},r)$-function (for a particular r and $\tilde{r}$) is $O(n^3)$. where $n = X(W)$.

We can estimate $\lambda^3 Q(r,\tilde{r})$ using a statistic of the form

$$(2\nu(W))^{-1} \sum_{\eta,\zeta,\xi \in X \cap W} d_r(\eta,\zeta) d_r(\eta,\xi) d_{\tilde{r}}(\xi,\zeta) \omega(\eta,\zeta,\xi)$$

which makes use of the weight function $\omega(.)$ (see also Chapter 4). Choosing

$$\omega(\eta,\zeta,\xi) = \frac{\nu(W)}{\gamma_W(\eta,\zeta,\xi)}$$

where

$$\gamma_W(\eta,\zeta,\xi) = \nu(W_\eta \cap W_\zeta \cap W_\xi)$$

and W_ξ denotes the translation of W by ξ, as before, we can arrive at the translation correction estimator (see Schladitz & Baddeley (2000), section 4.2)

$$\hat{Q}_{\text{trans}}(r,\tilde{r}) = \frac{\nu(W)^3}{X(W)(X(W)-1)(X(W)-2)} \sum_{\eta,\zeta,\xi \in X \cap W} \frac{d_r(\eta,\zeta) d_r(\eta,\xi) d_{\tilde{r}}(\xi,\zeta) I_{(\gamma_W(\eta,\zeta,\xi) \neq 0)}}{\gamma_W(\eta,\zeta,\xi)}$$

and this is defined whenever $X(W) > 2$.

3.4 Practical applications

In this final section we demonstrate the use of the $Q(r,\tilde{r})$-function in practical situations. We estimate the $Q(r,\tilde{r})$-functions of some real data and compare the $Q(r,\tilde{r})$-functions of simulated data. We also use simulation studies to show that the $Q(r,\tilde{r})$-function is a powerful tool for certain hypothesis tests.

As a demonstration, Figure 3.3 shows surface plots of the theoretical $Q(r,\tilde{r})$-function

for (respectively) the homogeneous Poisson process and Gauss-Poisson process with $q_1 = 0.4, q_2 = 0.6, s = 0.4$ and $\lambda = 250$ (in $\mathbb{R}^2$). These parameters have been chosen rather arbitrarily but in such a way that we expect a high number of events to be generated in simulation studies (this is desirable as a higher number of events results in a smoother surface when estimating the $Q(r, \tilde{r})$-function using methods such as those outlined in section 3.3).

The difference in the two surfaces in Figure 3.3 is made more obvious by plotting the

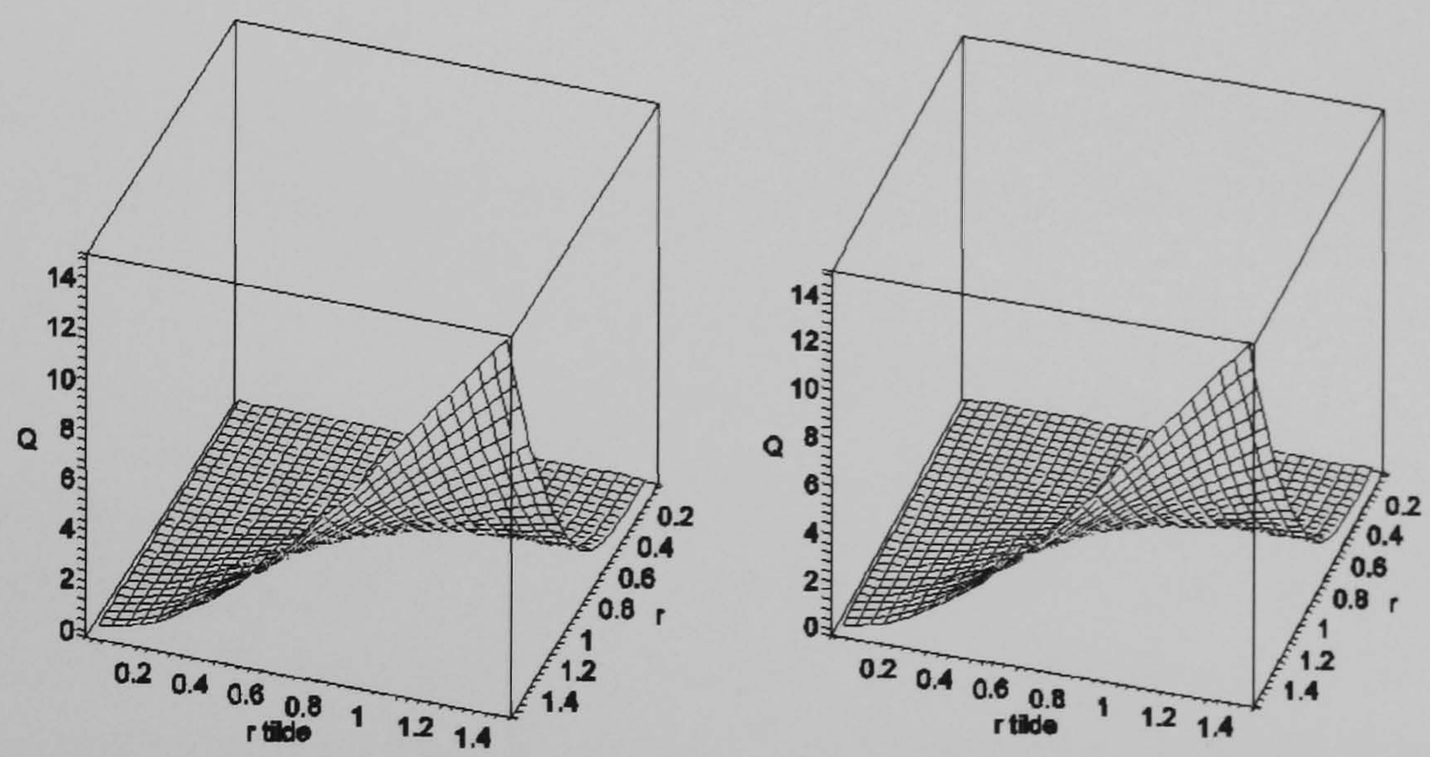


Figure 3.3: Theoretical Q-function of a homogeneous Poisson process (a) and Gauss-Poisson process (b)

difference between them and displaying them together on the same axis, as shown in Figure 3.4. Graph (b) in Figure 3.4 shows that (as suggested by the theoretical forms) the $Q(r, \tilde{r})$-function for the Gauss-Poisson process is the same as $Q_{\text{poi}}(r, \tilde{r})$ for $r < 0.4$, but becomes increasingly greater than $Q_{\text{poi}}(r, \tilde{r})$ as $\tilde{r}$ increases, for increasing values of $r \geq 0.4$. Note that numerical problems may be obtained when evaluating (3.5), particularly for the case when $r = \tilde{r}$. In the special case where $r = \tilde{r}$ it is advisable to use the solution

$$Q(r, \tilde{r}) = \frac{\pi(\pi - \frac{3\sqrt{3}}{4})r^4}{2}$$

which is the solution obtained by Schladitz & Baddeley (2000) for the $T(r)$-function in the planar case.

Figure 3.5 shows scatter plots (for the spatial case) of (respectively) a simulated homogeneous Poisson process, Gauss-Poisson process with the same parameters as

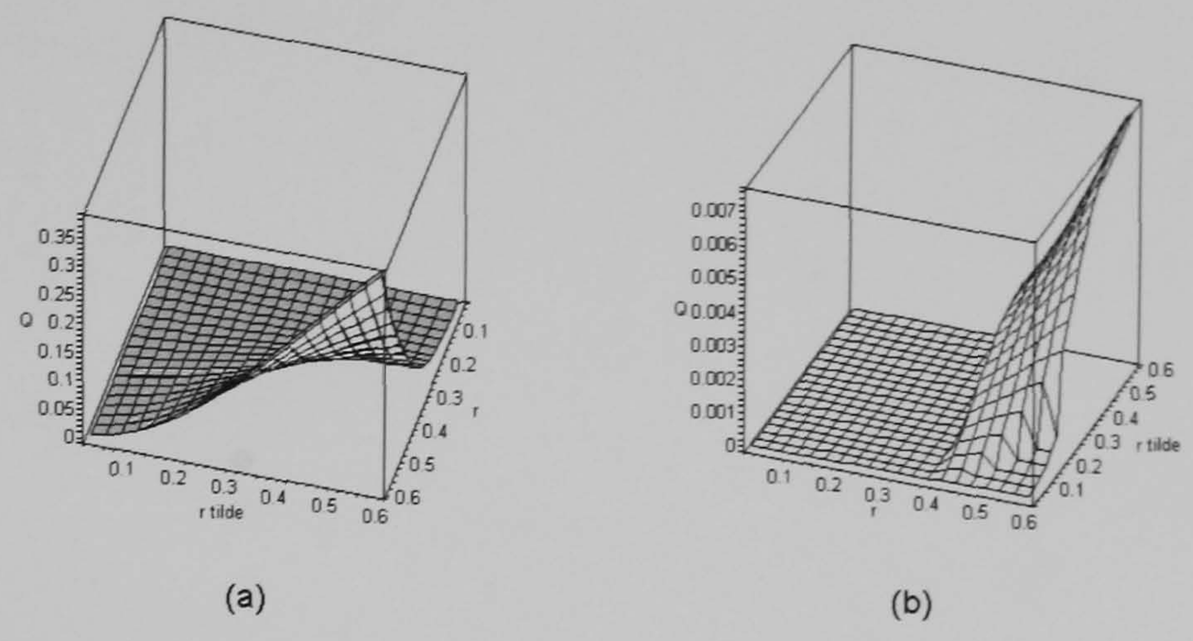
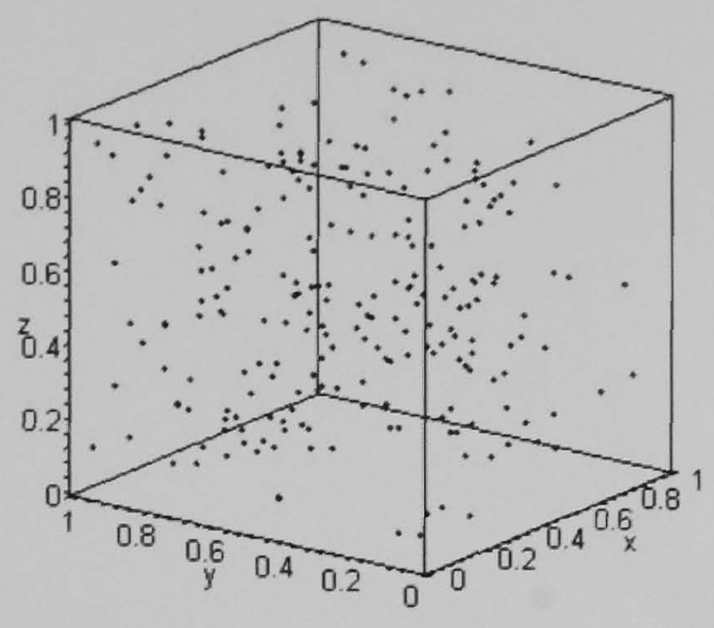


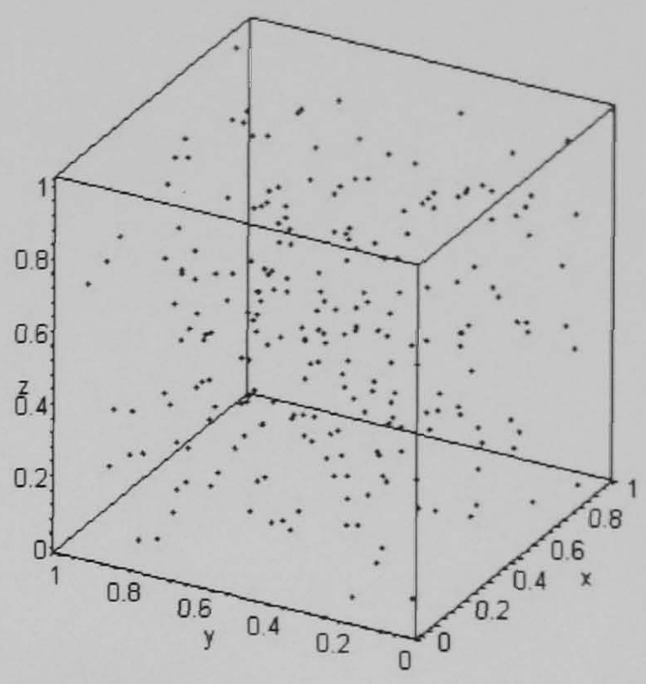
Figure 3.4: (a) Theoretical Q-function of a homogeneous Poisson process (white surface) and Gauss-Poisson process (Grey surface) (b) Q-function of a Gauss-Poisson process-$Q_{\text{poi}}(r, \tilde{r})$

used in the planar case above, and a p-thinned homogeneous Poisson process with $p = 0.4$ and intensity 250. Figure 3.5 demonstrates the need for descriptive point process functions. It is difficult to judge 'by eye' which spatial point process generated the spatial point patterns. Figure 3.6 shows a plot of the border estimates of the $Q(r, \tilde{r})$-function of the simulated homogeneous Poisson process and the simulated Gauss-Poisson process.

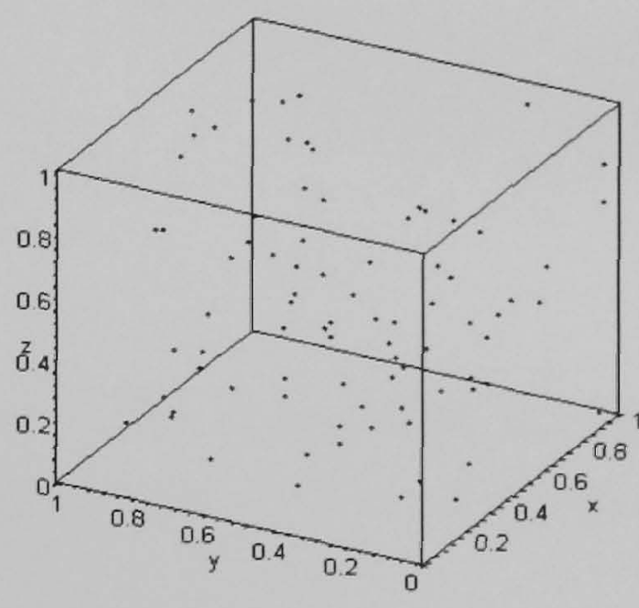
Figure 3.6 demonstrates that in this particular case, a border estimate is sufficient for highlighting the key features of the point patterns. As expected, the $Q(r, \tilde{r})$-functions are approximately the same for small r, then as r approaches 0.4, the $Q(r, \tilde{r})$-function of the Gauss-Poisson process becomes greater than the homogeneous Poisson process. Figure 3.7 shows a plot of the border estimates of the $Q(r, \tilde{r})$-function of the simulated homogeneous Poisson process and the simulated thinned Poisson process. For small r we approximately confirm the theoretical result that the $Q(r, \tilde{r})$-function of a thinned homogeneous Poisson process is the same as a the $Q(r, \tilde{r})$-function of a homogeneous Poisson process. We also demonstrate in Figure 3.7 that a border estimate coupled with a low intensity will result in increasingly unreliable results (in the sense of bias) as r increases. While these appear different, we also need to capture sampling variability. If the spatial point process that the spatial point pattern is generated by is known, then we could simply simulate many times from that process. We can then analyse the variance of estimated Q-functions for the simulated data. Often in practical situations, only a single realisation of a



(a)



(b)



(c)

Figure 3.5: Simulated homogeneous Poisson process (a), Gauss-Poisson process (b) and p-thinned Poisson process (c)

spatial point process is observed. In such cases, bootstrapping is a popular approach for assessing estimation uncertainty (see for example Künsch (1989), Cowling et al (1996), and Politis & Sherman (2001)).

The (respective) Japanese black pine, Redwood seedlings and cell centres data sets $Q(r, \tilde{r})$-function border estimates are plotted with the theoretical $Q(r, \tilde{r})$-function of a homogeneous Poisson process in Figure 3.8. The surface plots in Figure 3.8 provide some evidence for complete spatial randomness (CSR) for the Japanese black pine data, clustering for the Redwood seedlings data, and regularity for the cell data. The quantity

$$\sigma = \int_{0.02}^{0.2} \int_{0.02}^{0.2} \left(\frac{\hat{Q}(r, \tilde{r})}{Q(r, \tilde{r})} - 1 \right)^2 dr d\tilde{r} \tag{3.16}$$

which is an adjusted Cramér-von Mises test statistic (see Cressie, 1991, p667), was calculated for the data as a measure of the level of departure from CSR. Specifically, larger values of σ indicated a lack of fit for a CSR model. Monte Carlo

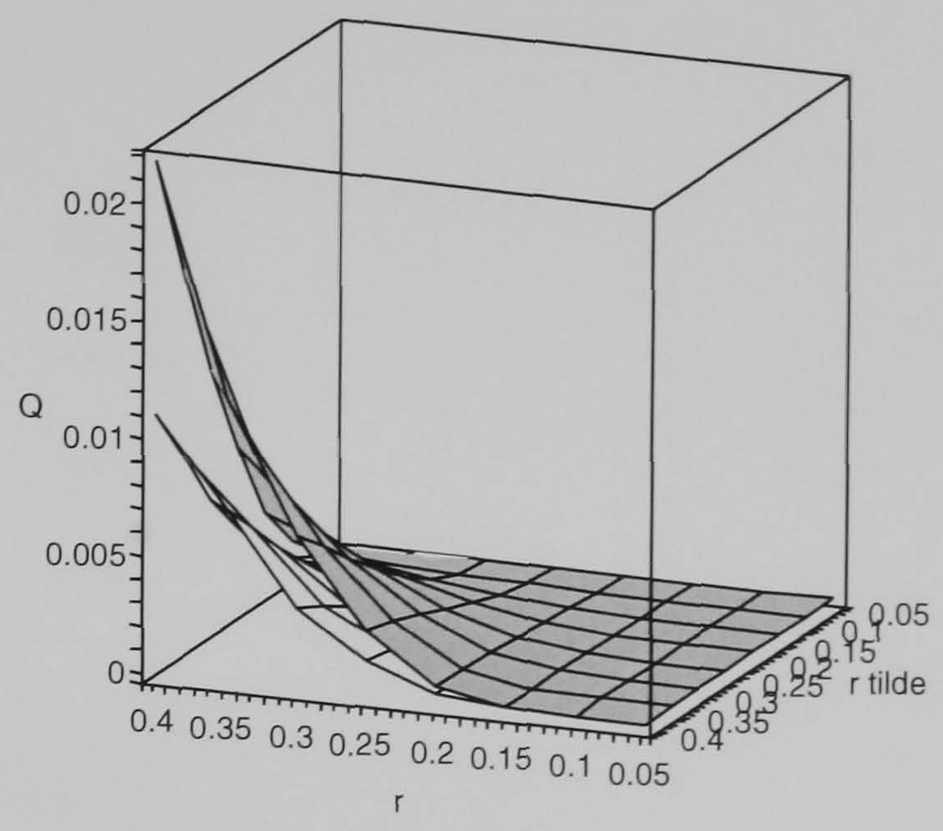


Figure 3.6: Border estimate of the Q-function of a simulated homogeneous Poisson process (white surface) and simulated Gauss-Poisson process (grey surface)

simulations were used to form a benchmark value for σ. We used small values of r and $\tilde{r}$, coupled with confidence intervals for the estimate, in order to maintain the usability of a border estimate. As we do not have the theoretical variance for the Q-function border estimate of any spatial point pattern, or knowledge of the underlying spatial dependence mechanism and marginal distribution that generated the data; the confidence intervals were formed by bootstrapping (Sherman & Carlstein, 1994). The re-sampling method used in the bootstrap procedure is a modification of the tiling method described by Loh & Stein (2004), with toroidal wrapping, as suggested by Politis & Romano (1992) for edge-correction purposes. More precisely, we re-sampled the point patterns M times to form M new estimates of $Q(r, \tilde{r})$. We then ordered the estimates. The Pth largest and Pth smallest were then used to form the confidence intervals at the required level of significance. The re-sampling method used entailed mapping the square W onto a torus by identifying opposite edges then applying a random toroidal shift. Two random shifts were used; the second shift was in the direction orthogonal to the first.

Integral (3.16) was approximated by the discrete summation:

$$\sum_{r=1}^{u}\sum_{\tilde{r}=1}^{u}\left(\frac{\widehat{Q}(\Delta r, \Delta\tilde{r})}{Q(\Delta r, \Delta\tilde{r})} - 1\right)^2 \tag{3.16}$$

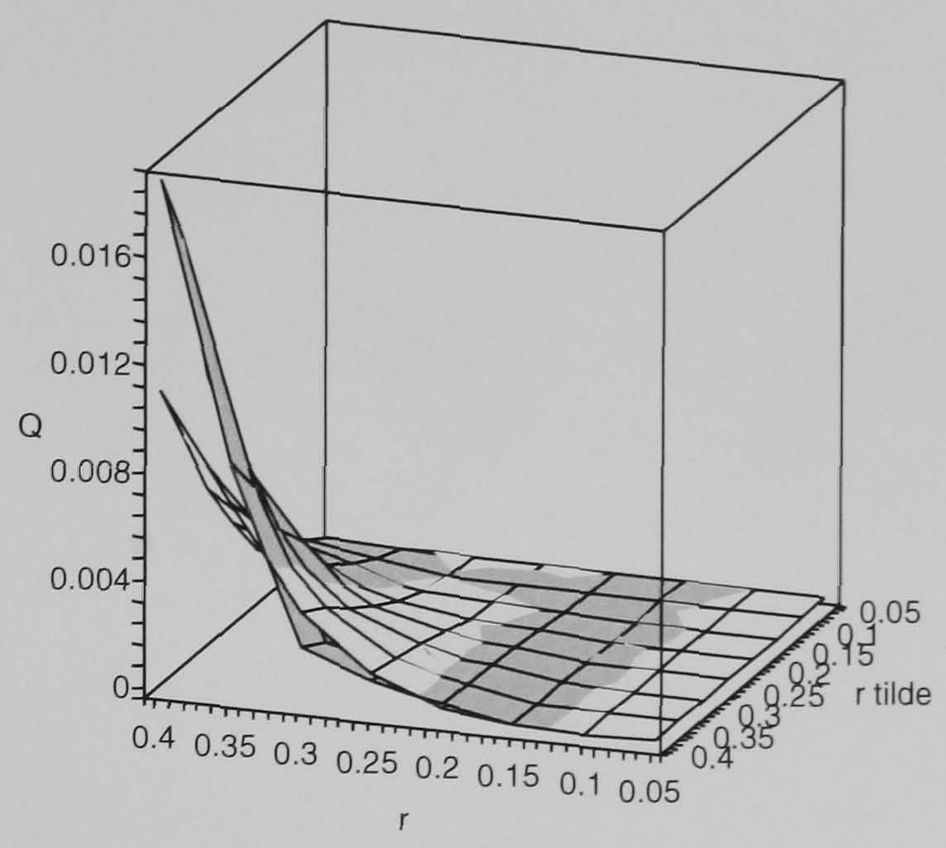


Figure 3.7: Border estimate of the Q-function of a simulated homogeneous Poisson process (white surface) and simulated p-thinned Poisson process (grey surface)

with $\Delta = 0.02$ and $u = 10$. We selected a value for Δ small enough to give good accuracy, while not requiring evaluation of many terms where $Q(\Delta r, \Delta \tilde{r})$ was zero. The maximum, σ_{MAX}, and minimum, σ_{MIN}, values of σ using border estimates of $Q(r, \tilde{r})$ of 100 simulations of homogeneous Poisson processes with intensities 65, 62 and 42 are given in the lookup table in the Appendix C. Note that these are estimates of the intensity of the Japanese black pine, Redwood seedlings and cell centres data respectively. The values in the lookup table are intended to act as a guideline. The values of σ for the border estimates of $Q(r, \tilde{r})$ for the Japanese black pine, Redwood seedlings and cell centres data are given in Table 3.1. We use $\sigma_{+99\%\mathrm{C.I}}$ and $\sigma_{-99\%\mathrm{C.I}}$ to denote the values of σ obtained by using (respectively) the upper and lower envelopes of a 99% confidence interval for the border estimates of $Q(r, \tilde{r})$. The confidence intervals were formed using the bootstrap method mentioned earlier with $M = 100$ and $P = 1$.

Table 3.1: Values of σ for Japanese black pine, Redwood seedlings and Cell centres

Point Pattern	σ	$\sigma_{+99\%\mathrm{C.I}}$	$\sigma_{-99\%\mathrm{C.I}}$
Japanese black pine	32	1210	83
Redwood seedlings	2010	14589	1956
Cell centres	94	99	86

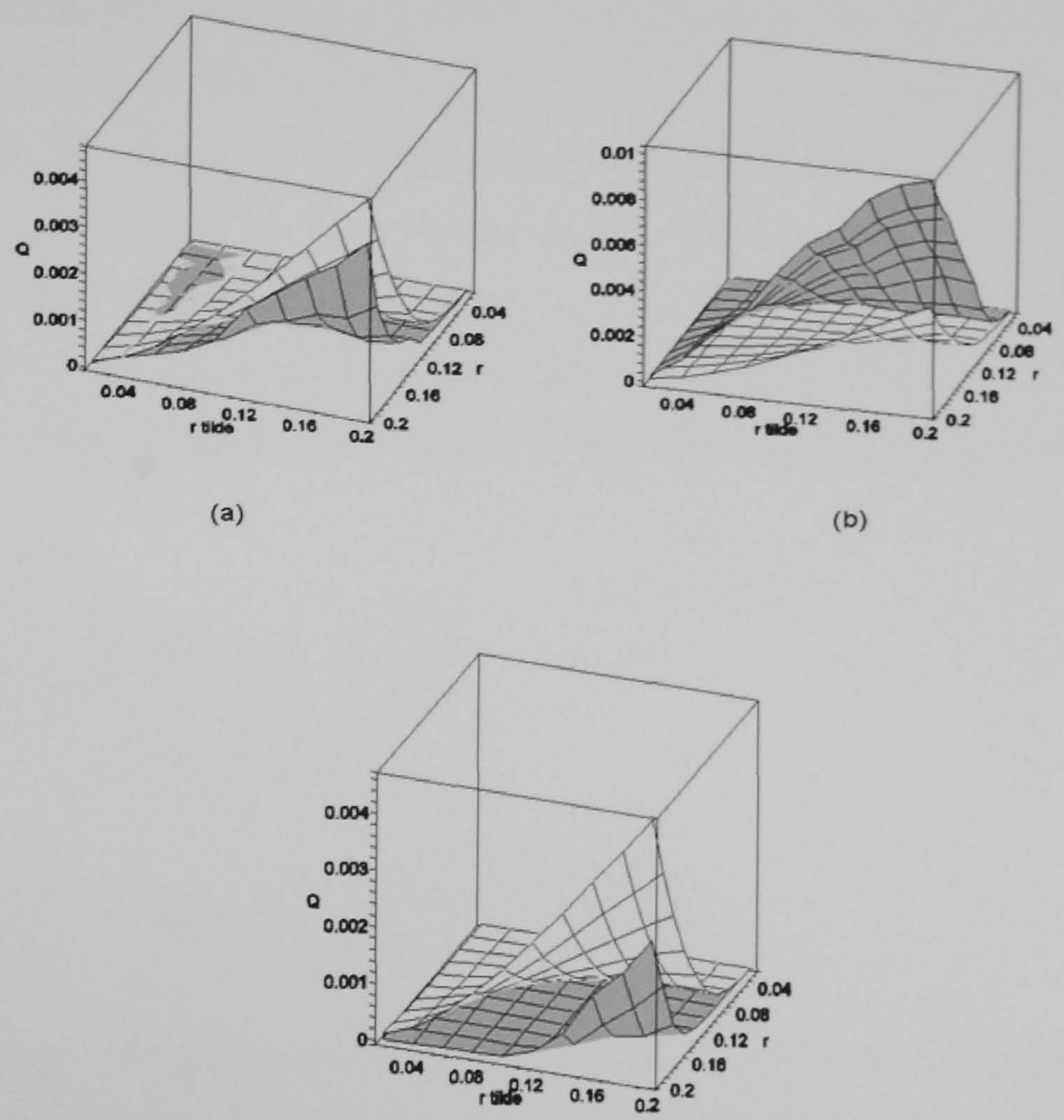


Figure 3.8: Q-function of the theoretical homogeneous Poisson process (white surfaces) with estimates for the Japanese black pine (a), Redwood seedlings (b) and cell centres data (b) (grey surfaces)

The values of σ for the real data are consistent with Figure 3.8, the clearest result being that the Japanese black pine data gives the strongest evidence for CSR. By assessing individual values of the $Q(r, \tilde{r})$-function of the cell centres data, a disadvantage of the $Q(r, \tilde{r})$-function and other characteristics such as the $K(r)$-function is highlighted, namely, that with few points and a less sophisticated form of edge-correction (for example the border method), less descriptive estimates may be obtained for small r and $\tilde{r}$.

3.4.1 Comparing the Q-function to other characteristics

Grabarnik & Chiu (2002) used simulations from a mixed Matérn cluster process with a conditional (on n events) Strauss process to compare the power of hypothesis

tests based on the Q^2 statistic with those based on four other summary statistics, including the $K(r)$-function. Using a similar experimental set-up to Grabarnik & Chiu (2002) we assessed how well tests for CSR on a mixture of regular and clustered data based on using the $Q(r, \tilde{r})$-function compared with the same test using the $L(r)$-function (which is $\sqrt{\frac{K(r)}{\pi}}$ for the two-dimensional case) in place of the $Q(r, \tilde{r})$-function. We simulated 100 events uniformly inside the unit square 99 times and used these simulations to form 99 separate estimates of σ, $\{\tilde{\sigma_1}, \ldots \tilde{\sigma_{99}}\}$ by using (3.16) with $u = 25$. From the same simulated data, we also produced 99 estimates of the quantity

$$\sigma_L = \int_0^\tau \left(\frac{\widehat{L}(r)}{L(r)} - 1 \right)^2 dr \tag{3.17}$$

using the estimate

$$\hat{\sigma_L} = \sum_{i=1}^{u_k} \left(\frac{\widehat{L}(\Delta r)}{L(\Delta r)} - 1 \right)^2, \tag{3.18}$$

$\{\tilde{\sigma_{L1}}, \ldots \tilde{\sigma_{L99}}\}$ with $u_k = 25$ and (as is also the case for σ in this case) with $\Delta = 0.02$. These quantities were produced using un-edge corrected estimates of the summary statistics and formed the significance levels for our simulation studies. Similarly to as done by Grabarnik & Chiu (2002), we simulated a mixed inhibition and cluster process formed by superimposing a conditional Matérn cluster process with n_p parent points, fixed n_d number of daughters per parent, and interaction radius R, with a Strauss process with interaction parameter γ and a conditional number of events n_s. The Matérn cluster process was simulated using the approach outlined by Stoyan et al (1995), while the Strauss process was simulated using the MCMC algorithm outlined by Diggle (2003, p76) with 40 'sweeps' of n_s depletion-replacement steps. We repeated this entire experiment 1000 times and calculated the percentage of times that the estimates of the quantities (again based on un-edge corrected summary statistics for a fair comparison) σ and σ_L for the mixed process, ($\tilde{\sigma_0}$ and $\tilde{\sigma_{L0}}$) where ranked (in terms of magnitude) in the top 5% of $\{\tilde{\sigma_0}, \tilde{\sigma_1}, \ldots \tilde{\sigma_{99}}\}$ and $\{\tilde{\sigma_{L0}}, \tilde{\sigma_{L1}}, \ldots \tilde{\sigma_{L99}}\}$ respectively. The results for different values of γ are shown in Figure 3.9. For values of $\gamma < 0.6$ the quantity σ_L does not appear to do the K-function as much justice compared with using the maximum statistic adopted by Grabarnik and Chiu (2002). That is, tests based on the K-function and the quantity σ_L yield a lower power for the hypothesis test that the mixed process is CSR, compared to those that are based on using the K-function with the maximum statistic. However, we obtain similar results to Grabarnik and Chiu (2002) for tests based on

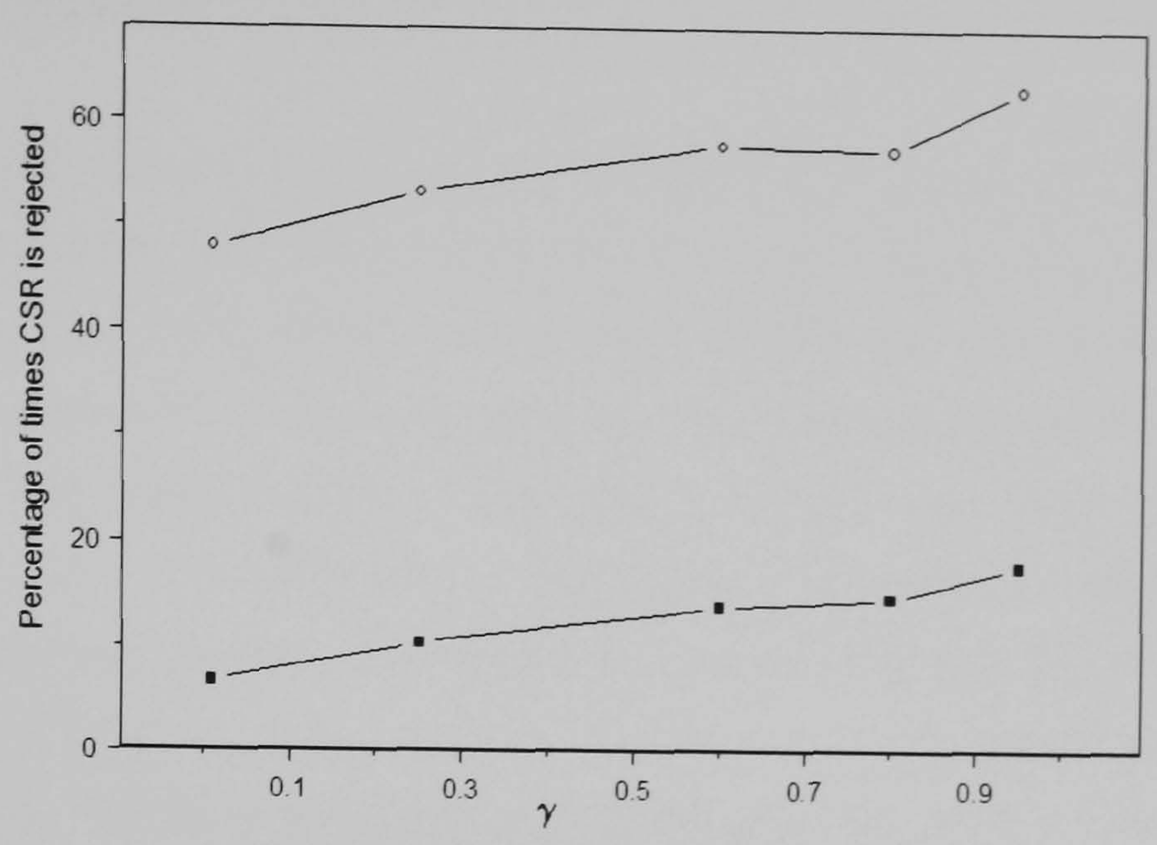


Figure 3.9: Graph showing the percentage of times that CSR is rejected at the 5% significance level for a simulated mixed Matérn cluster and Strauss process with $n_p = 17$, $n_d = 3$, $n_s = 50$ and $R = 0.06$. The points represented by a circle are the results for the $Q(r, \tilde{r})$-function and those represented by a square are for the $L(r)$-function.

the $K(r)$-function when $\gamma \geq 0.6$. Also, it is clear from Figure 3.9 that the tests based on the $Q(r, \tilde{r})$-function always strongly outperform those based on the $K(r)$-function. Furthermore, for all considered values of γ, the $Q(r, \tilde{r})$-function obtains similar or (in particular for $0.01 < \gamma \leq 0.8$) better results than tests based on the Q^2 statistic.

As a final simulation study we consider a process which we refer to as an *Nth generation spatial point process* (NGSPP). For this spatial point process, we begin with a realisation of a homogeneous Poisson process with intensity λ, and we call these events parents. These parent events then give birth to m_1 events, called daughters. Call these m_1 events the first generation. These daughters give birth to m_2 events and become parents of the second generation and so on. The spatial distribution of the daughters is such that they are scattered uniformly within the ball centred at the parent. Let $m^{(N)} = (m_1, \ldots, m_N)$ and $r^{(N)} = (r_1, \ldots, r_N)$ be vectors providing (respectively) the information for the number of daughter events, m_i, in generation i; r_i is the radius of the ball in which the daughters are uniformly placed. This type of model was used in an experiment to see how well the $K(r)$-, $T(r)$- and $Q(r, \tilde{r})$-

functions distinguish between point patterns from the same process but with slightly different parameters.

The experiment involved simulating an NGSPP with parameters $r^{(3)} = (0.1, 0.1, 0.1)$, call this the test data, and simulating 500 iid realisations from an NGSPP with parameters $r^{(3)} = (0.1, 0.05, 0.05)$, called the null model. For both the test data and null model, we use $m^{(3)} = (5, 5, 5)$ and $\lambda = 5$. The parameters were chosen in such a way that the expected number of events is similar to a high number of PNIL (for one cell nuclei), and it is difficult to distinguish between a realisation from the null compared to the test model. The steps for simulating the NGSPP process with parameters $m^{(N)} = (m_1, \ldots, m_N)$ and $r^{(N)} = (r_1, \ldots, r_N)$, inside the bounded window W are as follows. We first simulate a Poisson process with intensity λ, inside W, to produce n parent events, $\{\xi_1, \ldots \xi_n\}$. For each of the n parent event ξ_is, we simulate m_j events $\text{Gen}_{ij} = \{\xi_{ij1}, \ldots \xi_{ijm_1}\}$ uniformly inside $\{b^d(\xi, r_j) \cup W\}$ for the first generation ($j = 1$). The superposition $\cup_{i=1}^{n} \text{Gen}_{i1}$ forms the next set of parents. The process is repeated for $j = 2 \ldots N$. The set $\cup_{j=1}^{N}(\cup_{i=1}^{m_j} \text{Gen}_{ij})$ forms the realisation of the NGSPP process.

Simulations from the null model (typical realisations from this model are shown in Figure 3.11) were used to obtain significance levels for the test data (the first four of which are shown in Figure 3.10), and test the null hypothesis that the test data comes from the null model. Significance was determined by ranking estimates of the $K(r)$-, $T(r)$-, and $Q(r, \tilde{r})$-function estimates of the realisations from the null model and the test data. A curve (or surface in the case of the $Q(r, \tilde{r})$-function) of the test data was said to be ranked greater than that produced by a realisation of the null model if any point on the curve or surface was greater than a corresponding point on the curve or surface produced by the null data. This experiment was repeated 100 times. As in the previous experiment, we can safely ignore edge effects in this type of experimental setup and hence adopt un-edge corrected estimators. This is done to reduce the computational burden. The results of two-tailed tests are shown in Figures 3.12 and 3.13. As one can see from the bar chart (Figure 3.12), $Q(r, \tilde{r})$ and $T(r)$ reject at the 0.01 per cent levels more times than $K(r)$. Furthermore, tests based on $Q(r, \tilde{r})$ performs best overall. The null hypothesis that the test data is a realisation of the null model is correctly rejected more times (see Figure 3.13) using tests based on Q-function compared to those based on the K- and T-functions. Here we have not considered the frequency of a Type I error, and we reserve this as possible future work. However, we conjecture that using the $Q(r, \tilde{r})$-

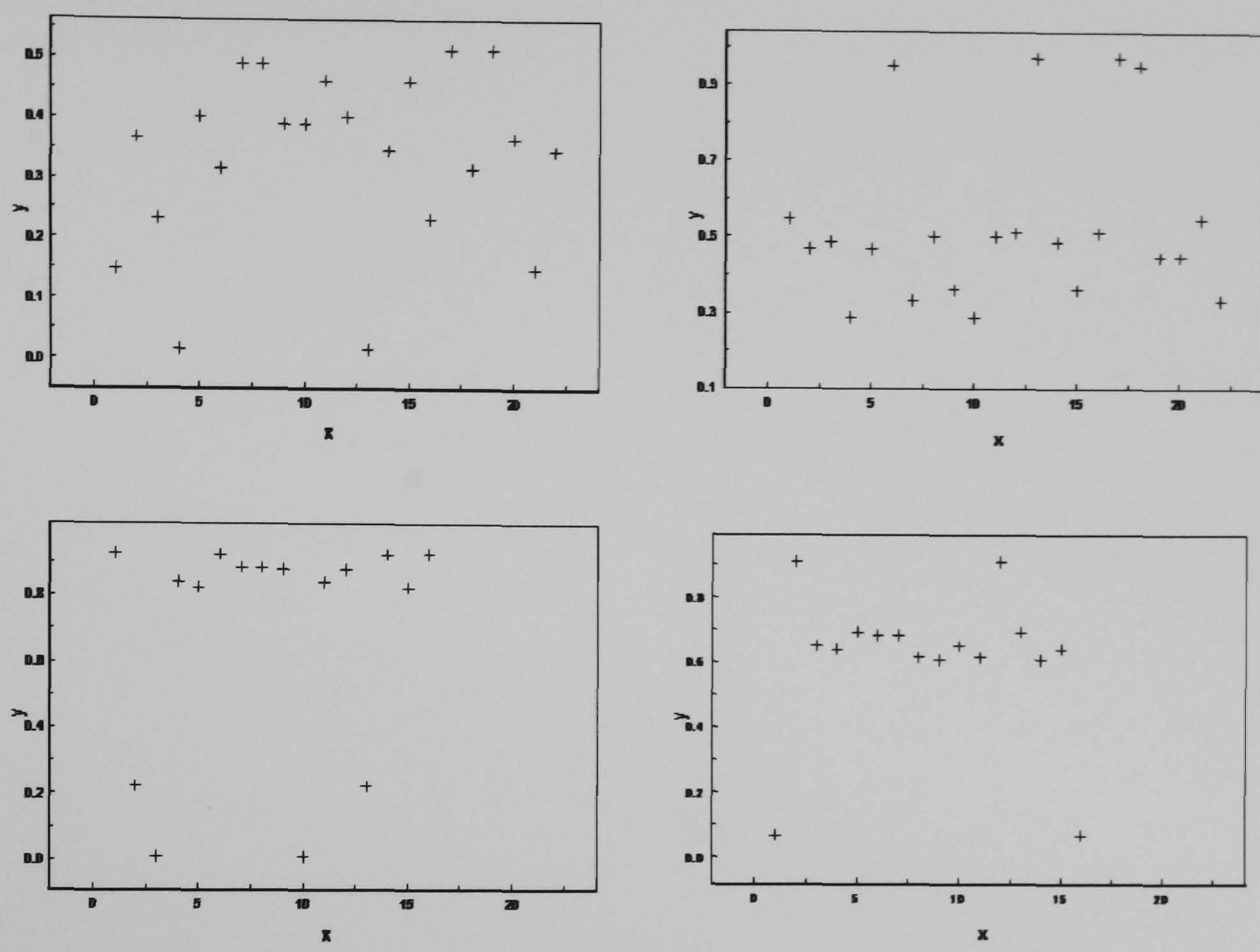


Figure 3.10: A sample of the NGSPP Model test data

as opposed to $K(r)$- and $T(r)$-functions is more likely to result in a more frequent occurrence of a Type I error in the above type of experiment. The reason being that the $Q(r, \tilde{r})$-function would detect smaller discrepancies in the data.

3.4.2 Q-function based CSR tests on PPDS1

We have shown that hypothesis tests based on the $Q(r, \tilde{r})$-function can be more powerful than those based on some other characteristics. We end this chapter with a repeat of the CSR tests on PPDS1 from Chapter 1, but this time using the $Q(r, \tilde{r})$-function. The outcome of these tests are provided in Figure 3.14. Similarly to the CSR tests in Chapter 1, cell nucleus 5 shows the greatest departure from CSR for the spatial distribution of PML NBs. All of the cell nuclei show signs of departure from CSR for small values of r and $\tilde{r}$. In particular, it is interesting that in these CSR tests that utilise the $Q(r, \tilde{r})$-function, the level of departure for the distribution of PML NBs in cell nucleus 1 appears (visually) much greater than that shown in Figure (1.5) from Chapter 1.

In this chapter we have defined a new third-order summary statistic and by utilising

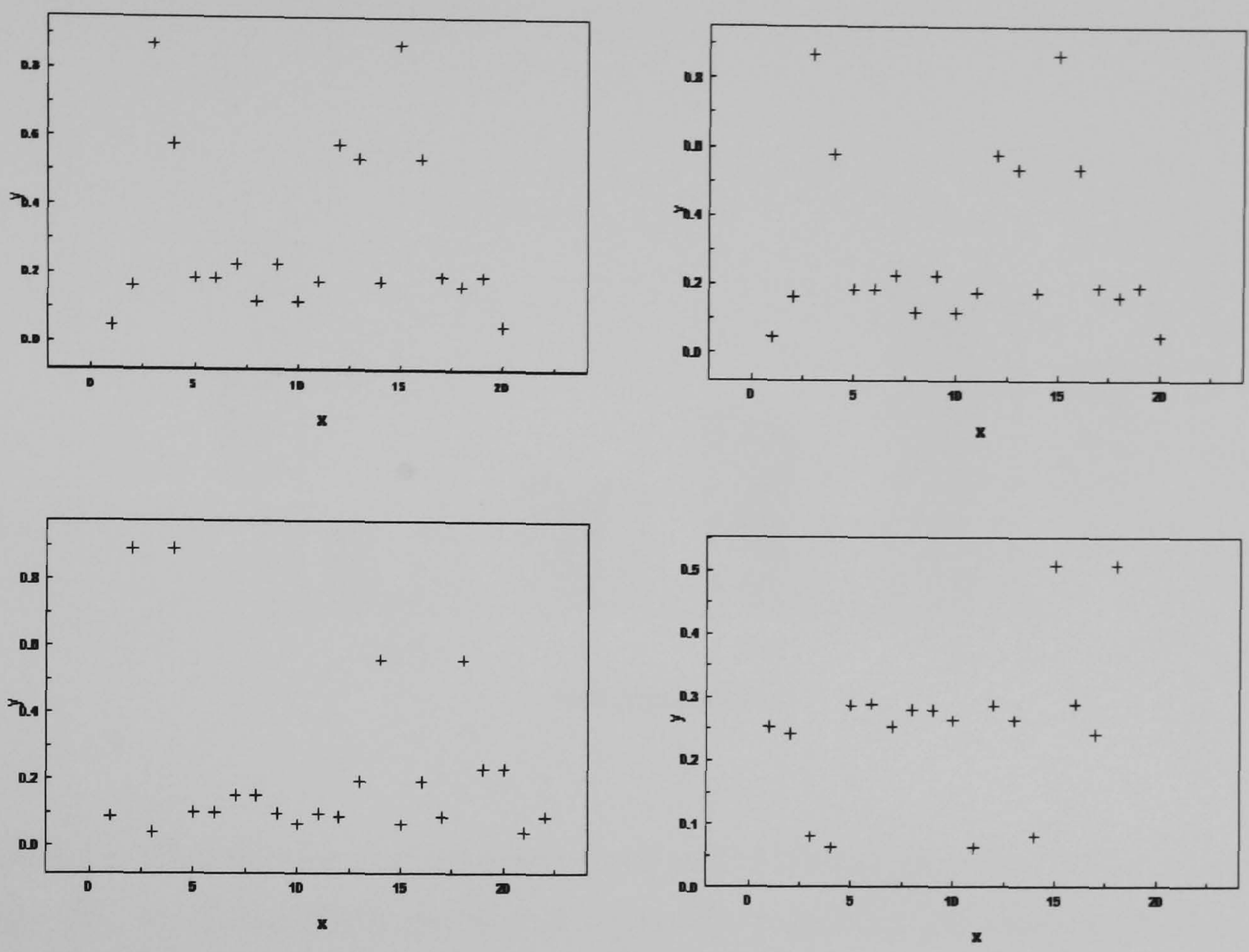


Figure 3.11: Typical realisations of the NGSPP null Model

some of the more complicated theory from Chapter 2, have derived its theoretical form for some important spatial point processes. We have shown through simulation studies that the $Q(r, \tilde{r})$-function can be a powerful tool in hypothesis tests compared to other characteristics such as the $K(r)$-function. With this in mind, we have revisited CSR tests that are similar to those performed at the end of Chapter 1, but this time basing the tests on the $Q(r, \tilde{r})$-function. From the results of these tests we are more confident that there is evidence for some departure from CSR, for the spatial distribution of PML NBs. Furthermore, the level of departure from CSR varies amongst cell nuclei. The next stage (and that naturally follows from the outcome of CSR tests) in this analysis, endeavours to bring us closer to understanding the organisation of PML NBs within the nucleus, through formulating and fitting a spatial point process model to the PML NB data. This is tackled in Chapter 5. However, before embarking upon this exposition, in the proceeding chapter we will present detailed discussions on edge effects and edge-corrections, as these play an important role in Chapter 5 and are an important issue in general spatial point process application.

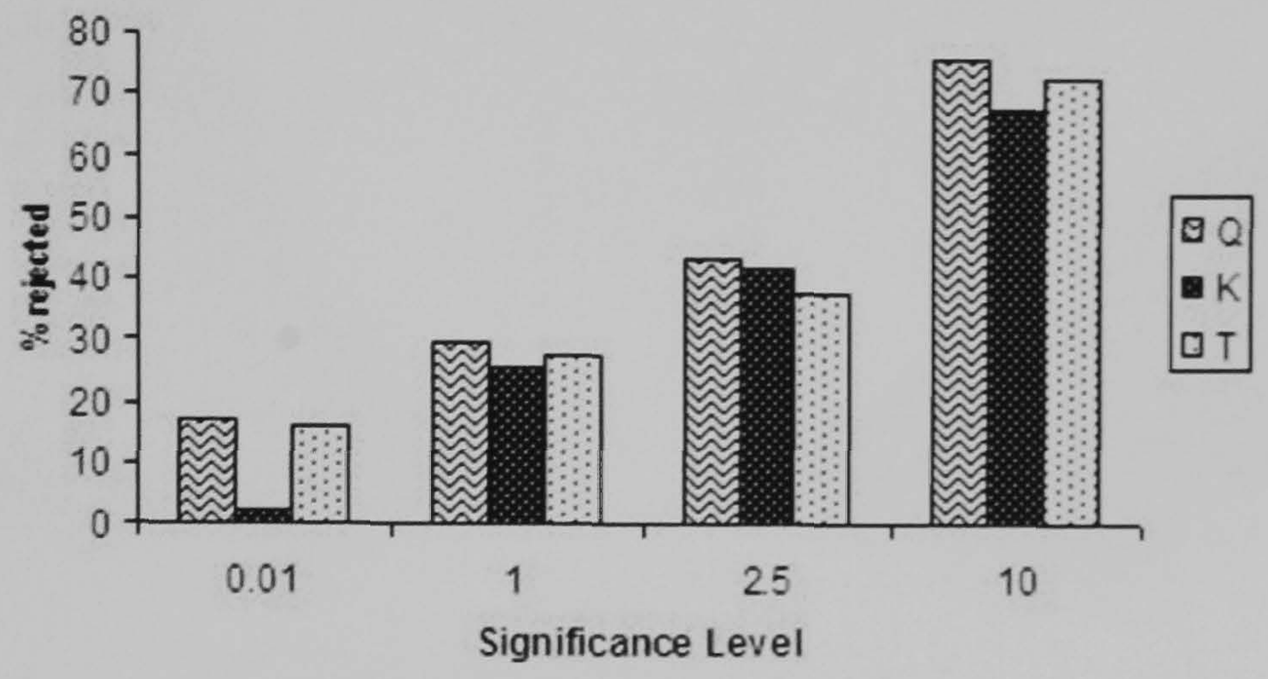


Figure 3.12: Results for the percentage of times that a hypothesis test based on using the Q-, K-, or T-function results in correctly rejecting the null model at particular significance levels

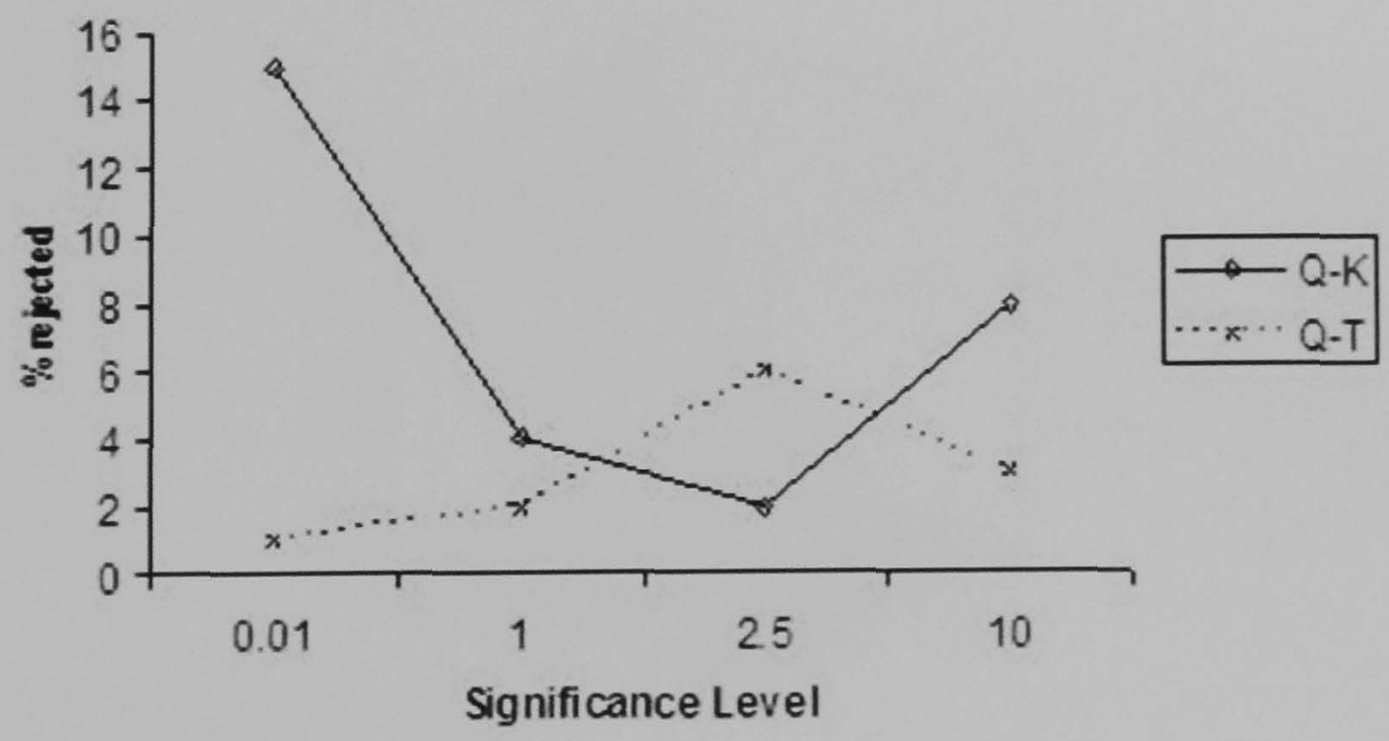


Figure 3.13: Results for the percentage of times that a hypothesis test based on using the Q-function results in correctly rejecting the null model at a particular level of significance compared to the K- and T-functions

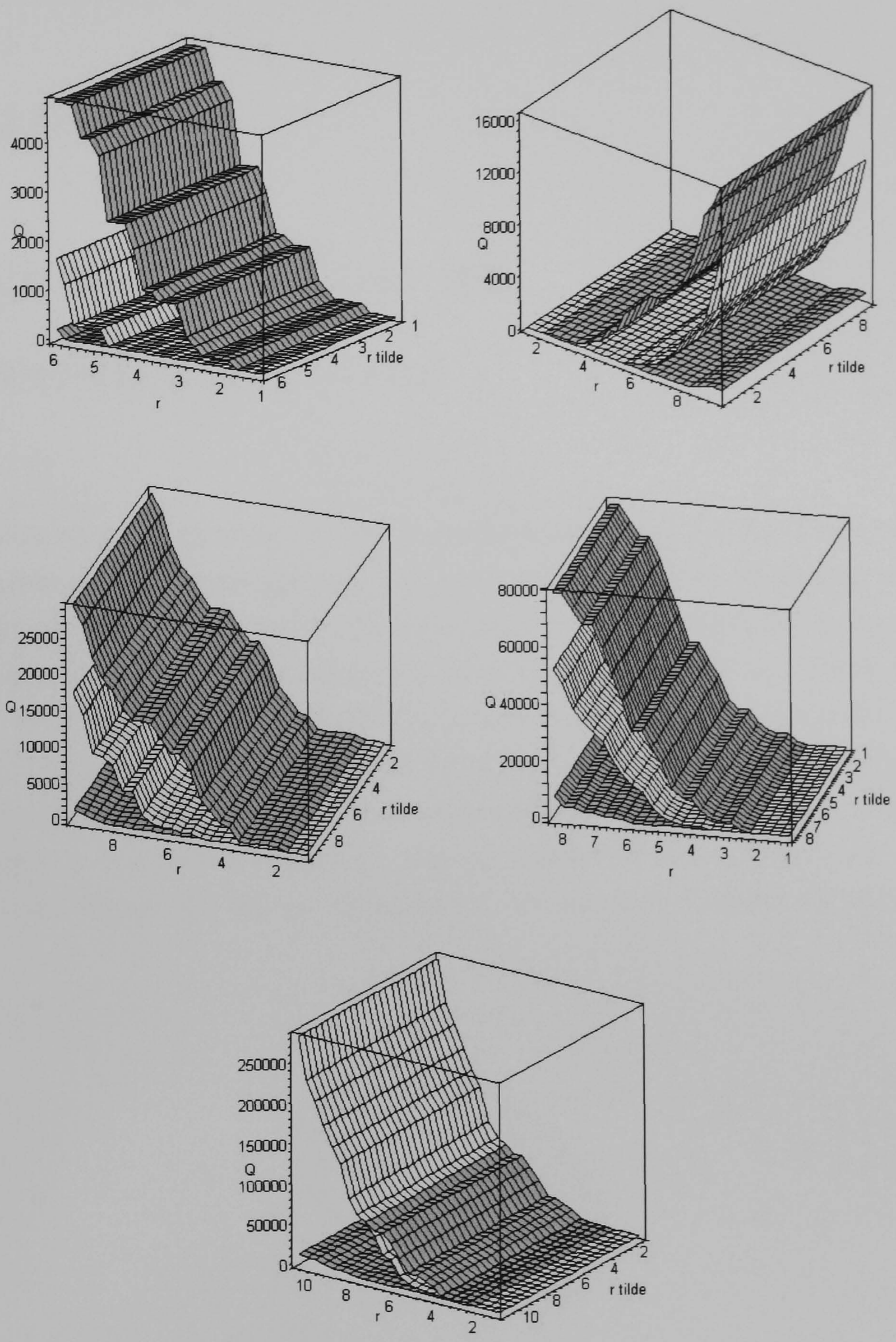


Figure 3.14: $Q(r, \tilde{r})$-functions of cells 1-5 (white surfaces) of PPDS1 with simulation envelopes (grey surfaces). Results for cell 1 are on the top left.

Chapter 4

Edge-correcting in three-dimensions

Estimation of point process summary statistics such as the $K(r)$-function is usually complicated by edge effects. The term 'edge effect' is often used to describe the estimation bias that arises when we attempt to calculate, for example, one of the distance-based summary statistics from Chapter 2 in some bounded compact window $W \subset \mathbb{R}^d$, that is a subset of a larger region on which the point process is defined. In the context of the PML NB data, the estimation of a distance-based summary statistic may require us to count PML NBs within some distance r of some point x, but this may include regions outside of the nucleus, that are not of interest (see Figure 4.1 for an illustration). In the next chapter we shall at times

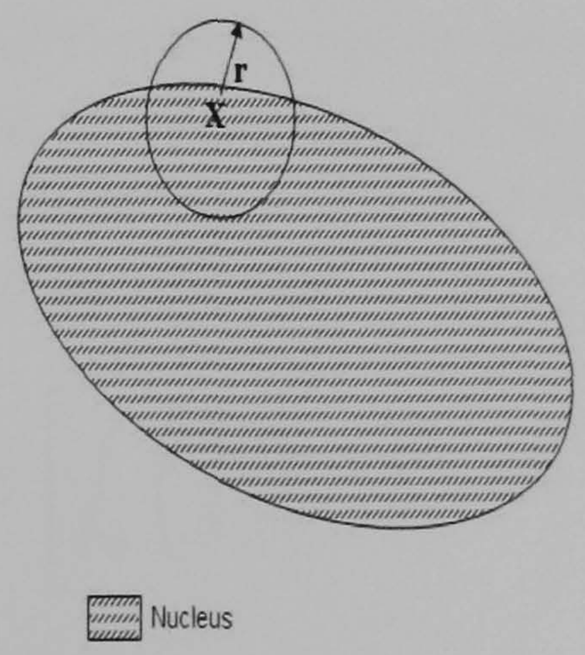


Figure 4.1: A demonstration of edge effects when estimating a distance-based summary statistic in the nuclear interior

find it important to carry out an edge-correction, for example when comparing a

theoretical value of a summary statistic with that estimated from simulated data.

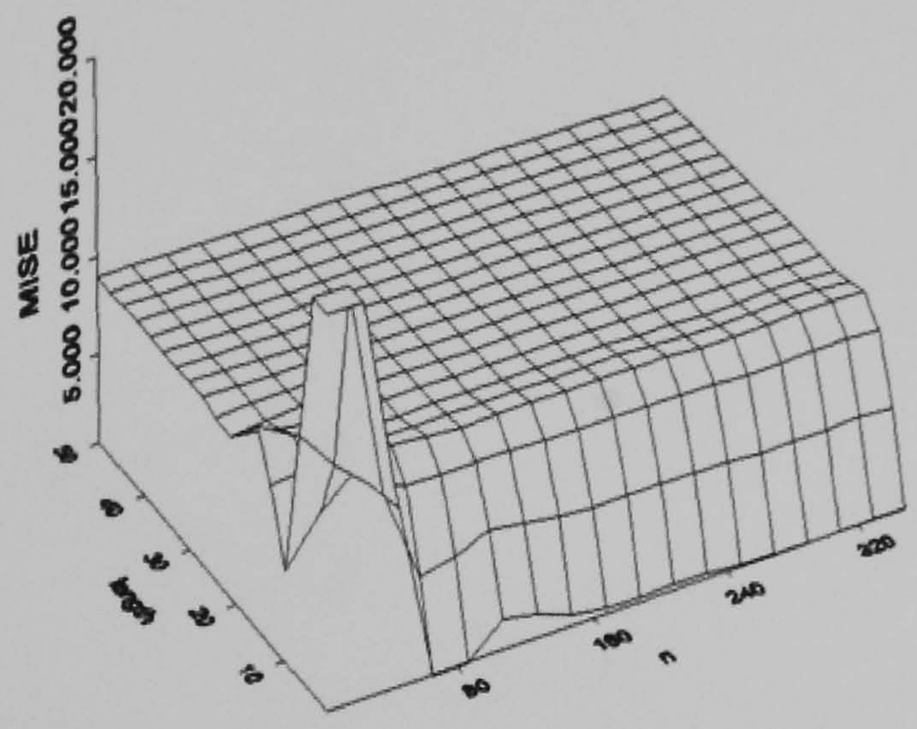
We have come across few, if any articles on point pattern analysis that discusses estimation of distance-based summary statistics but does not mention edge effects or edge-correction. The issue of edge effects is well studied for the two-dimensional case (see for example Ripley, 1979, Haase, 1995, Gignoux et al, 1999, Diggle, 2003 and more recently, Perry et al, 2006) but not so much so for the three-dimensional case. Discussion on the three-dimensional case is not as common place. This is mainly due to the complexity of edge-correction, increased computational expense, and perhaps the dearth of three-dimensional point patterns. For example, Gignoux et al (1999) report that three-dimensional ecological data is rare and costly to obtain. Notable contributions in the area of three-dimensional edge corrections include Baddeley et al (1987, 1993), König et al (1991), and Diggle et al (1991).

4.1 Initial illustrations

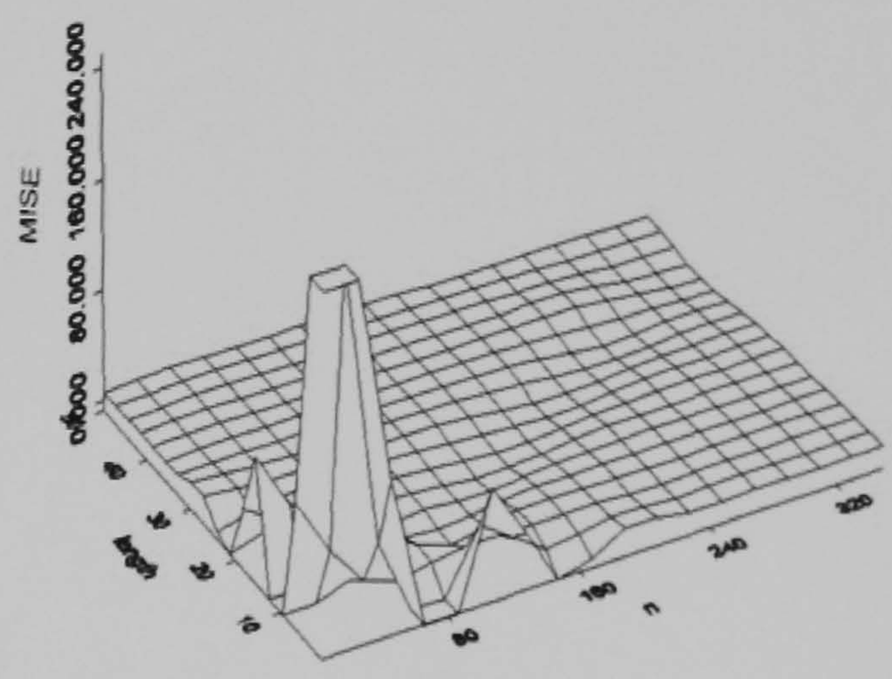
The need for edge-correction is more evident in three-dimensions compared to two-dimensions as edge effects become rapidly more severe, as the dimension of space increases, or as the distance r of interest increases (Gill, 1994). Baddeley et al (1993) have also identified informally that the increased dimensionality results in increased bias and variance. As an initial illustration of edge effects for the three- and two-dimensional cases, Figures 4.2-4.6 shows how an *adjusted Mean Integrated Squared Error* ($MISE$) varies with the size of window and number of events, for some important point processes. The results presented in Figure 4.2 were generated as follows. Given the number of events, n, and a cube window of a particular length, we simulate a conditional (on n) homogeneous Poisson process 100 times inside the cube window to form realisations $\{\tilde{x}_1, \ldots \tilde{x}_{100}\}$. For each set of 100 realisations, we compute the *adjusted* $MISE$ as given by

$$E\left[\int_{\Theta}\left(\frac{\widehat{S}(\theta)}{S(\theta)}-1\right)^2 d\theta\right]. \tag{4.1}$$

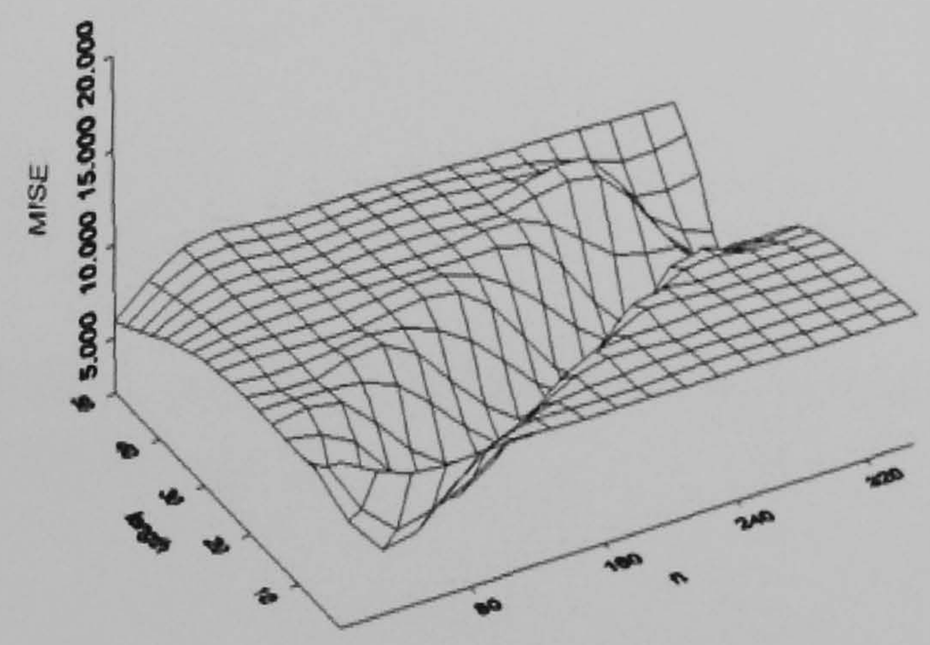
where $S(\theta)$ is the theoretical value of some summary statistics for a homogeneous Poisson process, and $\widehat{S}(\theta)$ is an unedge-corrected estimate of the summary statistic. Furthermore, Θ is the parameter space for $S(\theta)$. We refer to (4.1) as an 'adjusted' $MISE$ as it is an adjusted version of the $MISE$ given for example by Cressie (1991, section 8.5.1). By replacing the integral in (4.1) with a discrete summation,



(a)



(b)



(c)

Figure 4.2: $K(r)$-function adjusted $MISE$ for the homogeneous Poisson process in three-dimensions (for $t = \frac{1}{4}\sqrt{2}$ (a), $\frac{1}{2}\sqrt{2}$ (b), 1 (c))

we estimate (4.1) using

$$\frac{1}{100}\sum_{i=1}^{100}\sum_{\theta}\left(\frac{\widehat{S}(\theta;x_i)}{S(\theta)}-1\right)^2. \tag{4.2}$$

Figure 4.2 shows that when estimating the $K(r)$-function with t (the maximum value of r, that is, the upper limit of integration in (4.1)) small compared to the length of the cube window, the adjusted $MISE$ decreases as the intensity of the homogeneous Poisson process increases. The same experiment based on the $Q(r,\tilde{r})$-

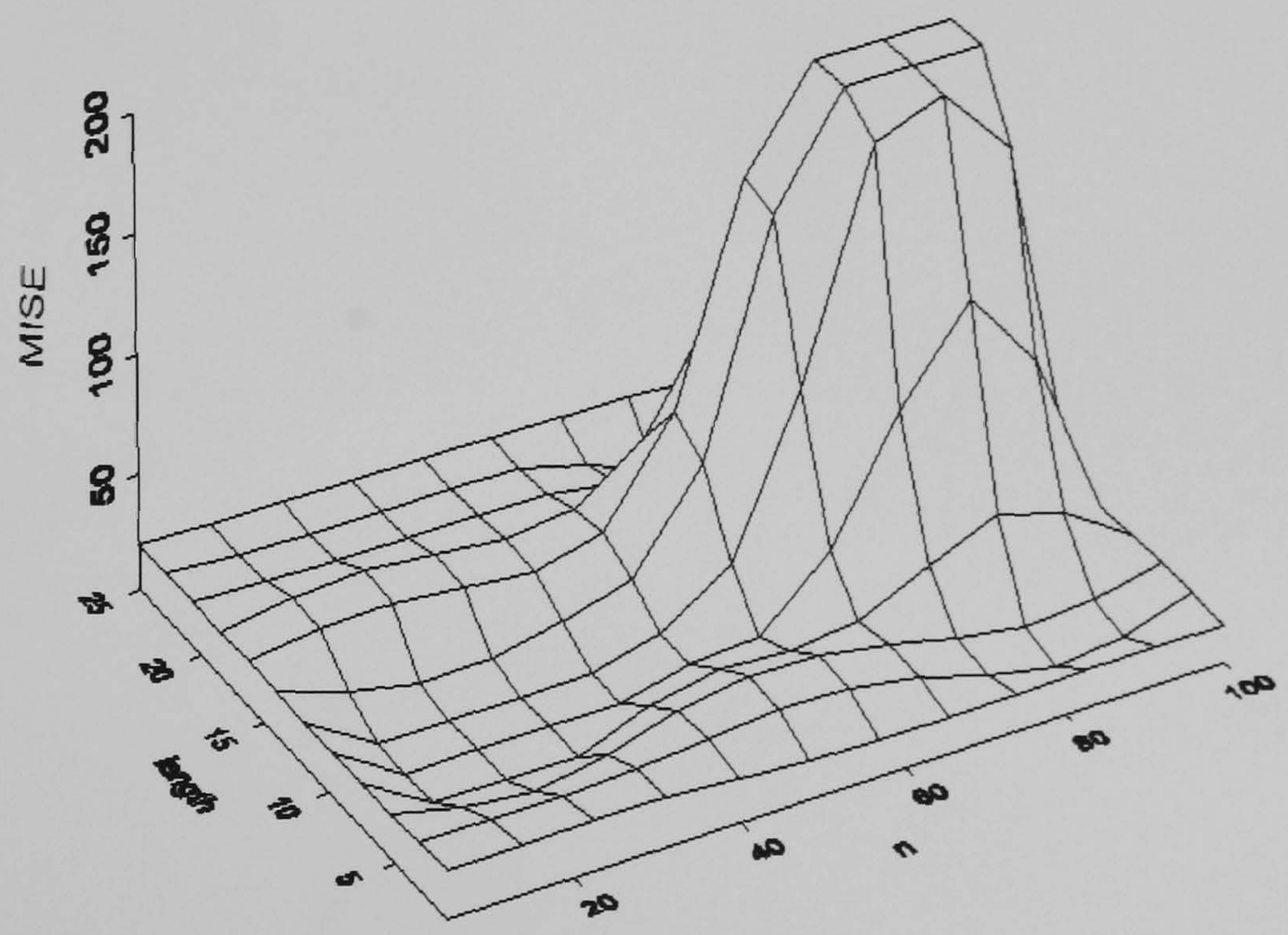
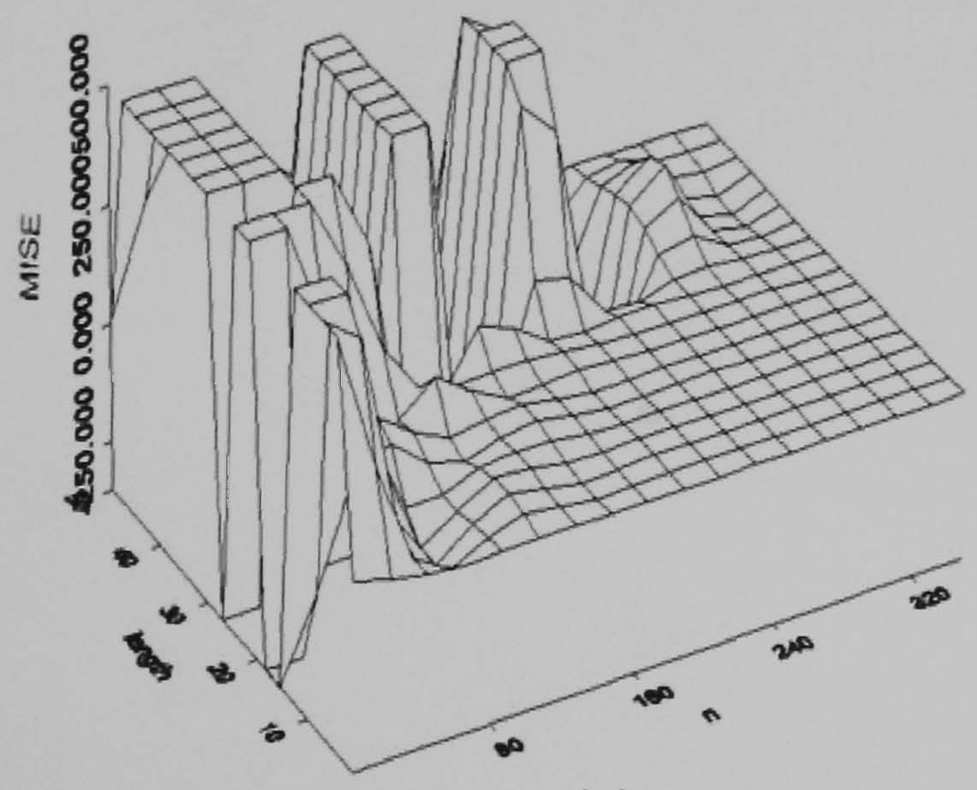


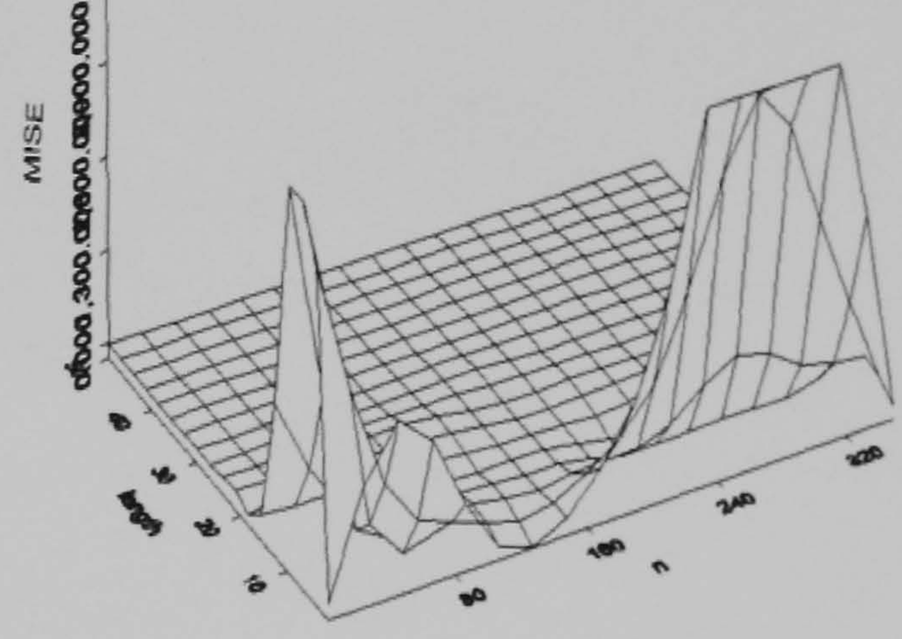
Figure 4.3: $Q(r,\tilde{r})$-function adjusted $MISE$ for the homogeneous Poisson process in three-dimensions (with $t = \frac{1}{2}\sqrt{2}$)

function (see Figure 4.3) showed that the adjusted $MISE$ is generally worse for the unedge-corrected $Q(r, \tilde{r})$-function compared to the unedge-corrected $K(r)$-function.

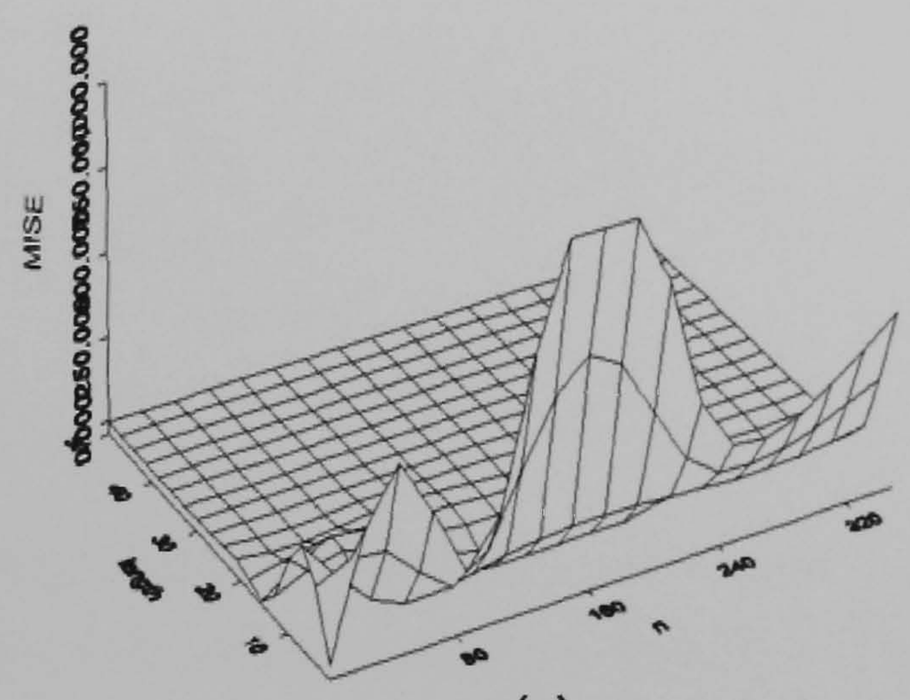
The same steps as those taken for producing Figure 4.2 were used to obtain the results in Figure 4.4 but for the two-dimensional case. The graphs in Figure 4.4 show that (as expected) as the parameter t is increased, the $MISE$ is reduced by increasing the size of the sampling window. Similar experiments for the adjusted $MISE$ for the Matérn cluster process and simple inhibition process (see for example Diggle (2003), page 72) for the case where we use the $K(r)$-function as S in the three-dimensional case yield interesting results (see Figures 4.5-4.6). In particular, for a simple inhibition process, the sampling window length, in comparison to the inhibition distance δ has notable bearing on the adjusted $MISE$. Also, for the Matérn cluster, when the radius of interaction, R is small, increasing the parent intensity does not have much of an impact on the adjusted $MISE$. A similar ex-



(a)



(b)

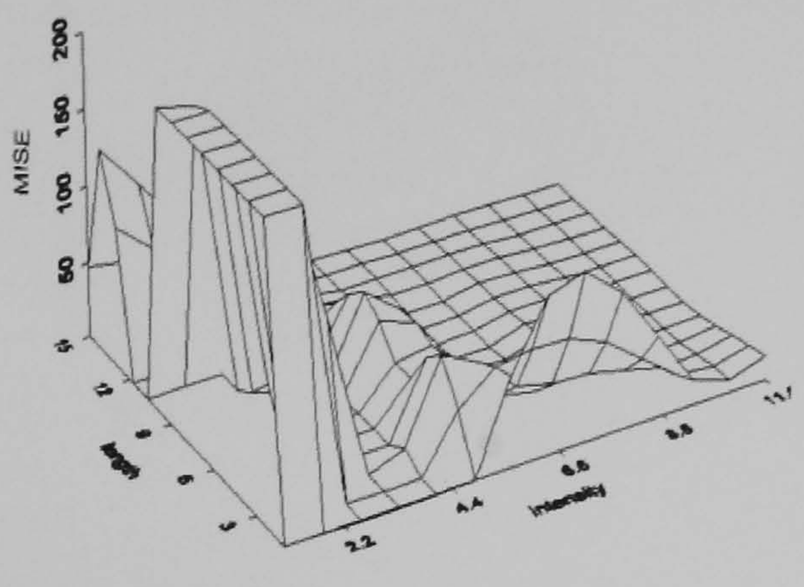


(c)

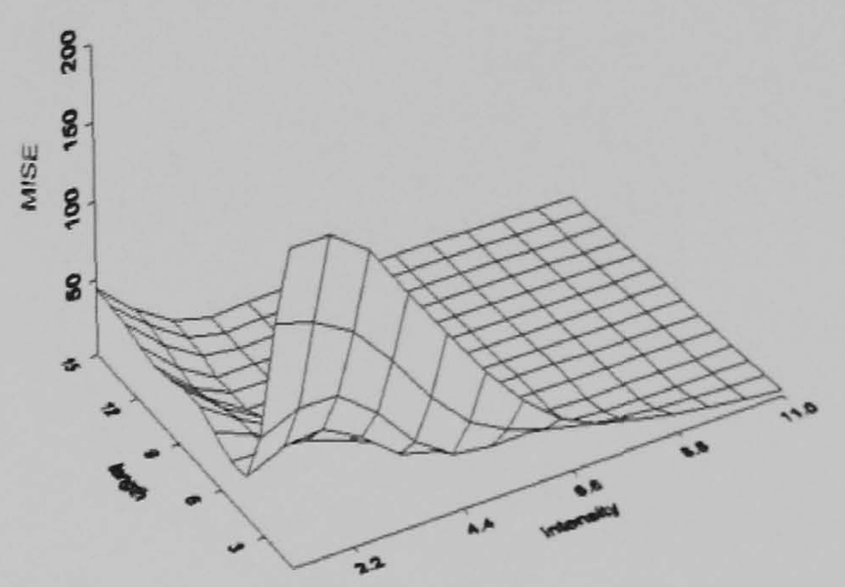
Figure 4.4: $K(r)$-function adjusted $MISE$ for the homogeneous Poisson process in two-dimensions for $t = \frac{1}{4}\sqrt{2}$ (a), $\frac{1}{2}\sqrt{2}$ (b), 1 (c)

periment was not carried out for the $Q(r, \tilde{r})$-function in place of the $K(r)$-function as we are unable to derive the $Q(r, \tilde{r})$-function for the Matérn cluster process and simple inhibition process.

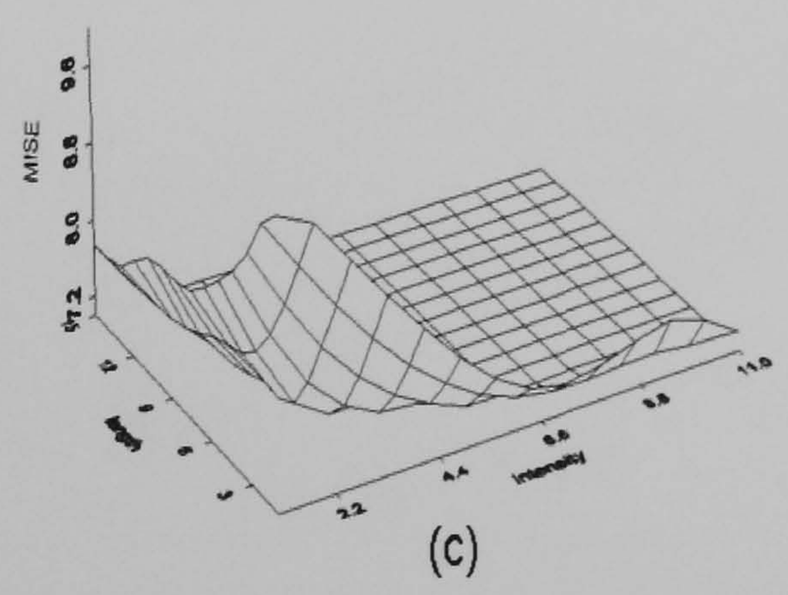
We are yet to state the $K(r)$-functions for the Matérn cluster process and simple inhibition process. A derivation now follows. Consider first a simple inhibition process. For this process the events come from a homogeneous Poisson process with intensity λ (which we refer to as the 'underlying intensity'), and events within dis-



(a)



(b)



(c)

Figure 4.5: $K(r)$-function adjusted $MISE$ for the simple inhibition process in two-dimensions with various underlying intensities and $\delta = 0.01$ (a), 0.1 (b), 0.25 (c), with $t = \frac{1}{4}\sqrt{2}$

tance δ of each other are thinned. The second product density of this process is given by

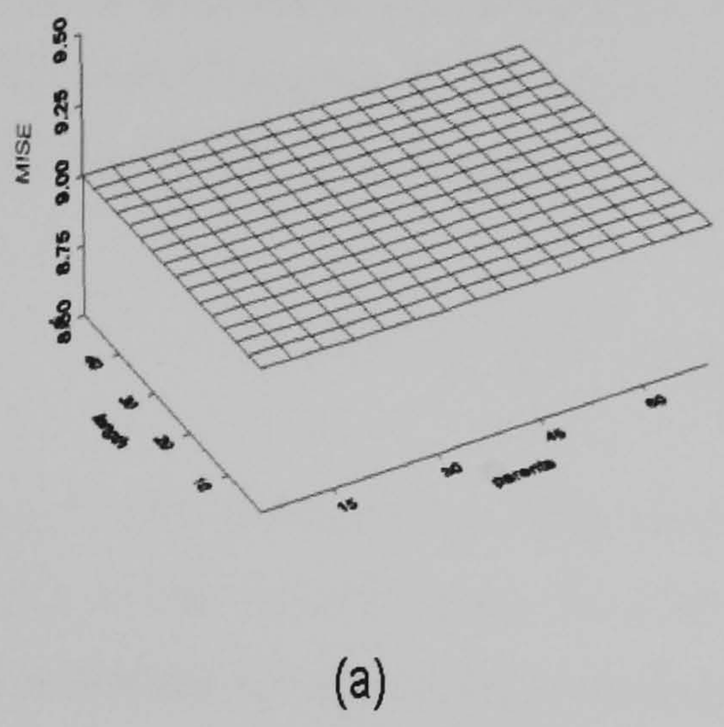
$$\varrho^{(2)} = \lambda^2 h(\|\eta - \xi\|) \equiv \lambda^2 h(r)$$

where

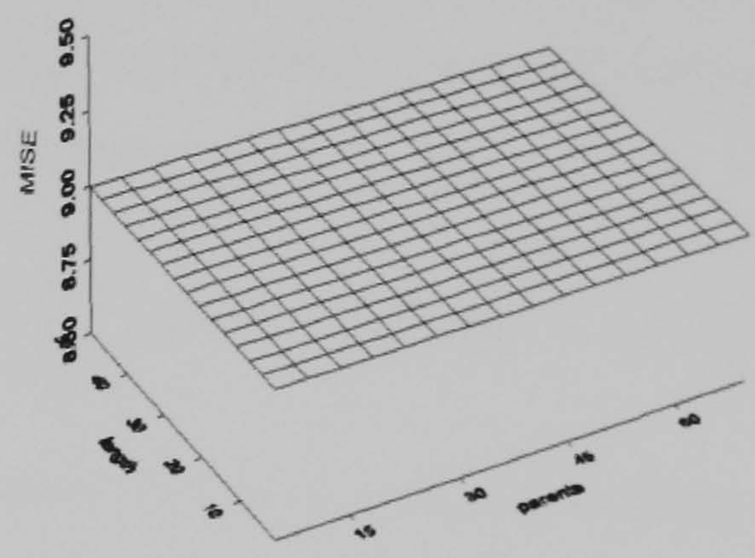
$$h(r) = \begin{cases} \exp\left(-\lambda\nu\left(b^d(o,\delta) \cap b^d(o + 1^{(d)}r, \delta)\right)\right) & r \geq \delta \\ 0 & r < \delta \end{cases}$$

and $1^{(d)}$ is the d-dimensional vector $(1, \ldots, 1)^T$. From (2.11) we have

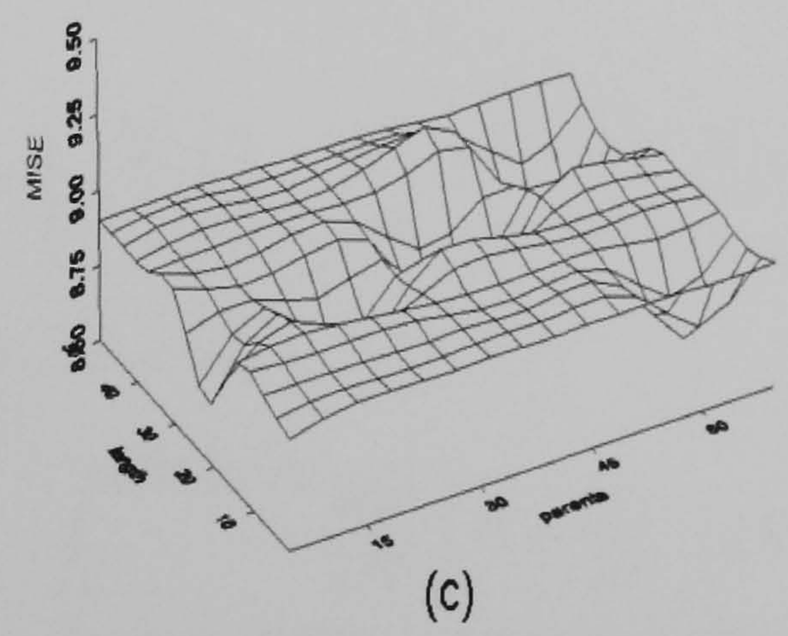
$$g(r) = h(r) \tag{4.3}$$



(a)



(b)



(c)

Figure 4.6: $K(r)$-function adjusted $MISE$ for the Matérn cluster process in three-dimensions for R = 0.01 (a), 0.1 (b), 0.25 (c) and $t = \frac{1}{4}\sqrt{2}$

Using (4.3) and the relationship between $K(r)$ and $g(r)$ (see Chapter 2, equation (2.13)) we have for the case $d = 3$ and $r \geq \delta$

$$\begin{aligned} K(r) &= 4\pi \int_{\delta}^{r} u^2 g(u) du \\ &= 4\pi \int_{\delta}^{r} u^2 \exp\left(-\frac{\lambda\pi}{12}(4\delta + u)(2\delta - u)^2 \right) du. \end{aligned} \tag{4.4}$$

Hence the $K(r)$-function for a simple inhibition process with inhibition radius δ (for $d = 3$) is

$$K(r) = \begin{cases} 4\pi \int_{\delta}^{r} u^2 \exp\left(-\frac{\lambda\pi}{12}(4\delta + u)(2\delta - u)^2\right) du & r \geq \delta \\ 0 & r < \delta. \end{cases}$$

As done throughout this document, the integral (4.4) can be evaluated using standard numerical procedures such as the trapezium rule.

We can derive the three-dimensional $K(r)$-function of a Matérn cluster process with parent intensity λ_p, daughter intensity λ_d and daughters uniformly placed in $b^{(3)}(o,R)$ (see Chapter 2) by using Stoyan et al's (1995) general formula (5.3.4) for the $K(r)$-function of a Neyman-Scott process:

$$K(r) = \frac{4}{3}\pi r^3 + \frac{1}{\lambda_p \lambda_d} \sum_{n=2}^{\infty} p_n n(n-1) D(r) \tag{4.5}$$

where p_n is the probability that there are n events of a representative cluster in $\mathbb{R}^d$ and $D(r)$ is the distribution function of the distance between two random events, whose locations are given by a uniform distribution on $b^{(3)}(o,R)$. By Santaló (1976), we have the density function f for the distribution function D (for the case $d=3$) is

$$f(r) = \begin{cases} \frac{3u^2}{2R^6}\left(R - \frac{u}{2}\right)^2\left(2R + \frac{u}{2}\right) & 0 < r < 2R \\ 0 \quad \text{otherwise.} \end{cases} \tag{4.6}$$

By (4.6) we have for $0 < r < 2R$

$$\begin{aligned} D(r) &= \int_0^{\min(r,2R)} \frac{3u^2}{2R^6}\left(R - \frac{u}{2}\right)^2\left(2R + \frac{u}{2}\right) du \\ &= \left(\frac{\min(r,2R)^6}{32R^6} - \frac{9\min(r,2R)^4}{16R^4} + \frac{\min(r,2R)^3}{R^3}\right) \end{aligned} \tag{4.7}$$

Combining (4.5) and (4.7) we obtain

$$K(r) = \frac{4}{3}\pi r^3 + \frac{\left(\frac{\min(r,2R)^6}{32R^6} - \frac{9\min(r,2R)^4}{16R^4} + \frac{\min(r,2R)^3}{R^3}\right)}{\lambda_p \lambda_d} \sum_{n=2}^{\infty} p_n n(n-1). \tag{4.8}$$

Further, for a Matérn cluster process $p_n \sim \text{Poisson}(\lambda_d)$ and hence

$$\sum_{n=2}^{\infty} p_n n(n-1) = \lambda_d^2. \tag{4.9}$$

Finally, by combining (4.8) and (4.9) we obtain the expression

$$K(r) = \frac{4}{3}\pi r^3 + \frac{\lambda_d}{\lambda_p}\left(\frac{\min(r,2R)^6}{32R^6} - \frac{9\min(r,2R)^4}{16R^4} + \frac{\min(r,2R)^3}{R^3}\right).$$

We have provided an initial illustration of edge effects by plotting the $MISE$ that is obtained when estimating the $K(r)$- and $Q(r,\tilde{r})$-functions without applying an

edge-correction. The structure of the rest of this chapter is as follows. We initially provide further demonstration, using both simulation and theoretical results, that edge effects are more severe in three-dimensions than two, and hence, given the three-dimensional nature of the data that is analysed throughout this document, there is a greater need for understanding methods for dealing with edge effects. We simultaneously highlight the bias obtained in failing to correct for edge effects. We then move on to discuss the various forms of edge-correction methods and make some suggestions for improving some of the currently adopted techniques. In Chapters 1 and 2, we often ignored edge effects without justification. In this chapter, we end with a discussion on when edge effects can be safely ignored and situations where it may be advantageous to not apply an edge-correction.

We shall refer to the functions $K(r)$, $L(r)$, $F(r)$, $G(r)$, and $g(r)$ (see Chapter 2) and use the notation $S(r)$ to mean any one of them, or statistics based on them. We will often use the quantity

$$\sigma_S = \int_0^\tau \left(\frac{\widehat{S}(r)}{S(r)} - 1 \right)^2 dr \quad \tau \in \mathbb{R}^+ \tag{4.10}$$

which will be estimated by the discrete summation

$$\widehat{\sigma_S} = \sum_{i=1}^{u_k} \left(\frac{\widehat{S}(\Delta i)}{S(\Delta i)} - 1 \right)^2 \tag{4.11}$$

as a means for quantifying the level of departure of estimated values, $\widehat{S}(r)$, from theoretical values $S(r)$. The quantity (4.10) will also be adjusted for application to the $Q(r, \tilde{r})$-function by simply extending to a double integral for the additional parameter $\tilde{r}$.

We will consider different types of observation region, $W \subset \mathbb{R}^d$, in which the point pattern is observed. The observation windows we are interested in include a brick (cuboid), sphere, and ellipsoid. Brick windows are common in practical situations (for example Baddeley et al (1993) and Fleischer et al (2006)). Our interest in spherical and ellipsoid windows follows from the shape of the mammalian cell nucleus (see Chapter 1). Throughout this chapter we limit our attention to stationary, and also in some cases, isotropic point patterns.

4.2 Comparing the severity of edge effects in three-dimensions with two-dimensions

Quantifying the severity of edge effects in two-dimensions compared to three is not trivial. One approach may be to compare simulation envelopes based on estimating a summary statistic of a spatial point process with a theoretical curve, as demonstrated in Figures 4.7 and 4.8. In both the two-dimensional and three-dimensional

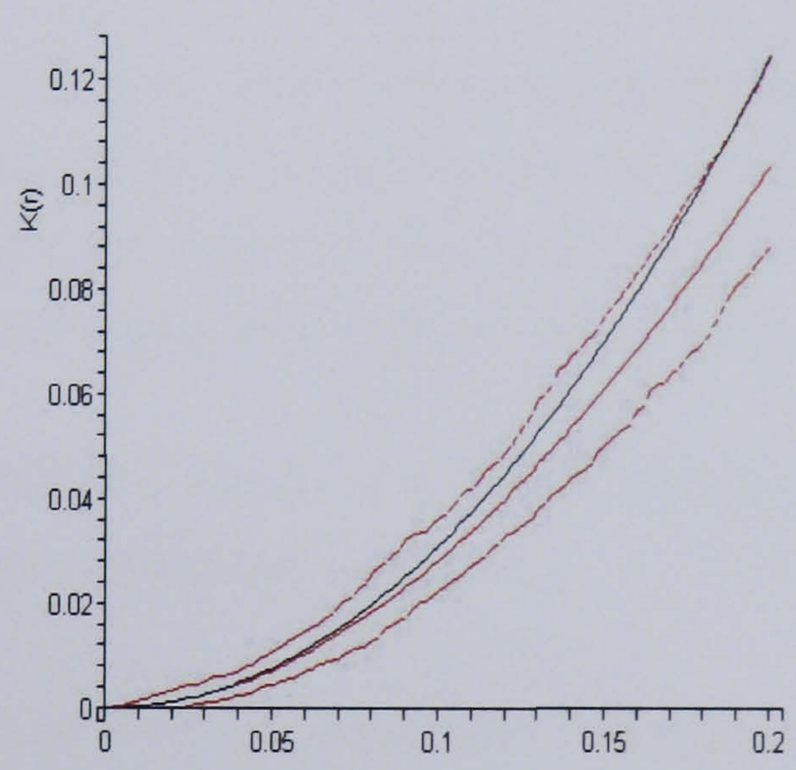


Figure 4.7: $E(\widehat{K}(r))$ $\times 10^2$ (red line) obtained by simulating 100 events uniformly 500 times inside the unit square. The envelopes for $\widehat{K}(r)$ $\times 10^2$ produced from these simulations (red dashed line) and the theoretical curve (black line) for a homogeneous Poisson process with intensity 100 are also displayed.

cases the un-edge corrected estimate of $K(r)$ is strongly biased downwards (noting the scaling of factor 100). However, it is difficult to draw any conclusions on the severity of edge effects in three-dimensions compared to 2 by simply comparing Figures 4.7 and 4.8. An approach that we adopt involves simulating n events m times in some window W and then estimating one of the summary statistics without correcting for edge effects and comparing with an estimate that incorporates edge-correction. For illustration, we use $n = 100$ uniformly distributed events inside a cube and square window W with lengths $l_0 = 1$ and set $m = 500$. We will at times refer to this as the 'standard scenario'. We carry out analysis in $W_{\ominus l}$ $(l < l_0)$. This is done so that to artificially create the scenario where we wish to have edge effects, by censoring the events lying outside of $W_{\ominus l}$. That is, if x is the point pattern observed in W, then $x \cap W_{\ominus l}$ is the resulting point pattern when events outside of the eroded window $W_{\ominus l}$ are censored (removed from the analysis). For the case where we do

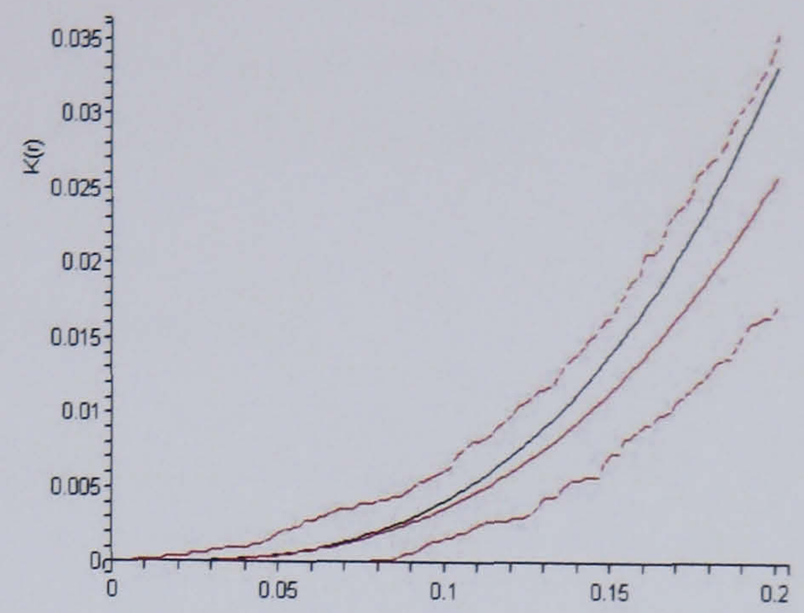


Figure 4.8: $E(\widehat{K}(r))$ $\times 10^2$ (red line) obtained by simulating 100 events uniformly 500 times inside the unit cube. The envelopes for $\widehat{K}(r)$ $\times 10^2$ produced from these simulations (red dashed line) and the theoretical curve (black line) for a homogeneous Poisson process with intensity 100 are also displayed.

not wish to have edge effects, we do not censor events. This is similar to carrying out a plus sampling edge-correction (see later).

To compare the expected level of departure between an estimate of a summary statistic $S(r)$ with the theoretical value, we consider a bias function, $BIAS(r)$, given by

$$BIAS(r) = \begin{cases} E(\widehat{S}(r_{2D}) - S(r_{2D})) & d = 2 \\ E(\widehat{S}(r) - S(r)) & d = 3 \end{cases}$$

where the adjustment $r_{2D} = \sqrt{\frac{4}{3}r^3}$. This adjustment is made for the two-dimensional case so that a better comparison can be made between three- and two-dimensions when using the standard scenario. That is, if $K(r)_{\text{poi}2D}$ and $K(r)_{\text{poi}3D}$ are the $K(r)$-functions for a homogeneous Poisson process in two- and three-dimensions respectively then, we account for a greater proportion of the total volume of a ball lying closer to the boundary in thee-dimensions than two by making distances commensurate by matching $K(r)$-functions. Specifically, we have $K(r_{2D})_{\text{poi}2D} = K(r)_{\text{poi}3D}$ implies that $r_{2D} = \sqrt{\frac{4}{3}r^3}$.

A candidate function for S in the $BIAS(r)$ function is the $L(r)$-function, that was briefly discussed in Chapter 2. This particular function is popular in spatial point process analysis because it has a simple closed form for the homogeneous Poisson process. The $L(r)$-function is also commonly used in order to stabilise variances (Stoyan, 2006). The un-edge corrected estimate $\widehat{L}_1(r)$ of $L(r)$ in compact window

W is given by

$$\widehat{L_1}(r) = \sqrt[d]{\nu(b^d(o,1))^{-1}(n/X(W))^{-2} \sum_{\eta,\zeta \in X \cap W} I_{(0<\|\eta-\zeta\| \leq r)}}.$$

The $BIAS(r)$ for $\widehat{L_1}(r)$ is therefore

$$BIAS(r) = \begin{cases} E(\widehat{L_1}(r_{2D}) - r_{2D}) & d = 2 \\ E(\widehat{L_1}(r) - r) & d = 3. \end{cases}$$

The curves of $BIAS(r)$ using the standard experimental setup inside a square and cube window for the L_1-function are presented in Figure 4.9

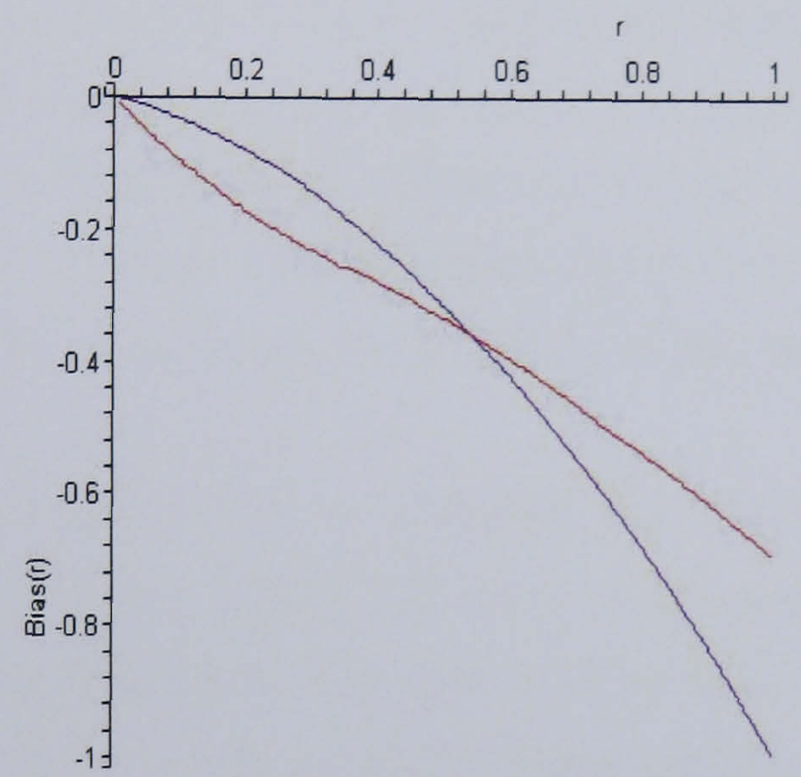


Figure 4.9: $BIAS(r)$ for the unedge-corrected L_1-function for the two-dimensional (blue) and three-dimensional case (red)

As intuition would dictate, the unedge-corrected estimate of the $L(r)$-function can be expected to be biased downwards. Further results for the bias obtained when estimating a summary statistic of point patterns, observed in cube and square windows, are displayed in Figures E.1-E.7 of Appendix E. Figure E.1 suggests that the absolute bias obtained in estimating the $K(r)$-function of a homogeneous Poisson process without edge-correcting, is greater for a unit square window, compared to a unit cube when $0.1 \leq r \leq 0.65$. This phenomena occurs for homogeneous Poisson processes of both high and low intensities. For small values of r ($r \leq 0.1$), the bias appears to be the same, regardless of whether the sampling window is a square or cube. When the length of the sampling windows are increased, less easily-described results are obtained (see Figure E.2). In this case, for a homogeneous Poisson process with low intensity ($n \leq 30$), there is not a clear pattern with regards to which

of the cube or square sampling window, results in the highest bias for $K(r)$-function estimation, for values of $r \leq 0.7$. For the case where we have a high intensity ($n = 100$), the absolute bias is consistently higher for the cube window. As the sampling window becomes very large compared to r (see Figure E.3), the absolute value of the bias is generally greater for the cube window compared to the square.

The bias is of course dependent on the spatial point process that the point pattern has been generated by. Figures E.4-E.6 suggest that the non edge-corrected estimate of the $K(r)$-function for a Matérn cluster process, with small interaction radius compared to the size of a cube sampling window, becomes strongly biased (compared to a homogeneous Poisson process) as the length of the sampling window increases. For a simple inhibition process inside a unit cube, the bias obtained in estimating the $K(r)$-function without edge-correction, is generally increased as the underlying intensity is increased. In particular, for $r \approx 0.7$, increasing the underlying intensity by a factor of 20, increased the absolute value of the bias by a factor of 100.

In Appendix E (Figures E.8-E.10) we present the $BIAS$ surfaces for the $Q(r, \tilde{r})$-function. These plots (Figures E.8-E.10) pose a demonstration that, as expected, the bias obtained when estimating the $Q(r, \tilde{r})$-function without edge-correcting, is higher than that obtained for a lower-order characteristic such as the $K(r)$-function. These observations hold as the sampling window size is increased. A somewhat interesting observation is that the estimation bias for the non edge-corrected $Q(r, \tilde{r})$-function remains relatively the same as the sampling window size is increased.

Combining the $BIAS$ function plots (E.1)-(E.7) with a plot of the mean square error ($MSE(r)$) function (see Figure 4.10) defined here for the L_1-function by

$$MSE(r) = \begin{cases} E((\widehat{L_1}(r_{2D}) - r_{2D})^2) & d = 2 \\ E((\widehat{L_1}(r) - r)^2) & d = 3 \end{cases}$$

we see that by using this approach as a means of comparing edge effect severity in two-dimensions with three, we obtain a counterintuitive outcome. Namely that, in some cases, for larger values of r the edge effects appear worse in two-dimensions. For example, edge effects seem worse in the two-dimensional case, after the point at which the MSE curves in Figure 4.10 cross (between $r \approx 0.7$ and 0.8, which is not too different to $\frac{\sqrt{2}}{2}$ and $\frac{\sqrt{3}}{2}$, which are the distances of the central point to the corners in the square and cube respectively) As a final comparison between edge effects in two- and three-dimensions, we consider a relative bias function ($RELBIAS(r)$)

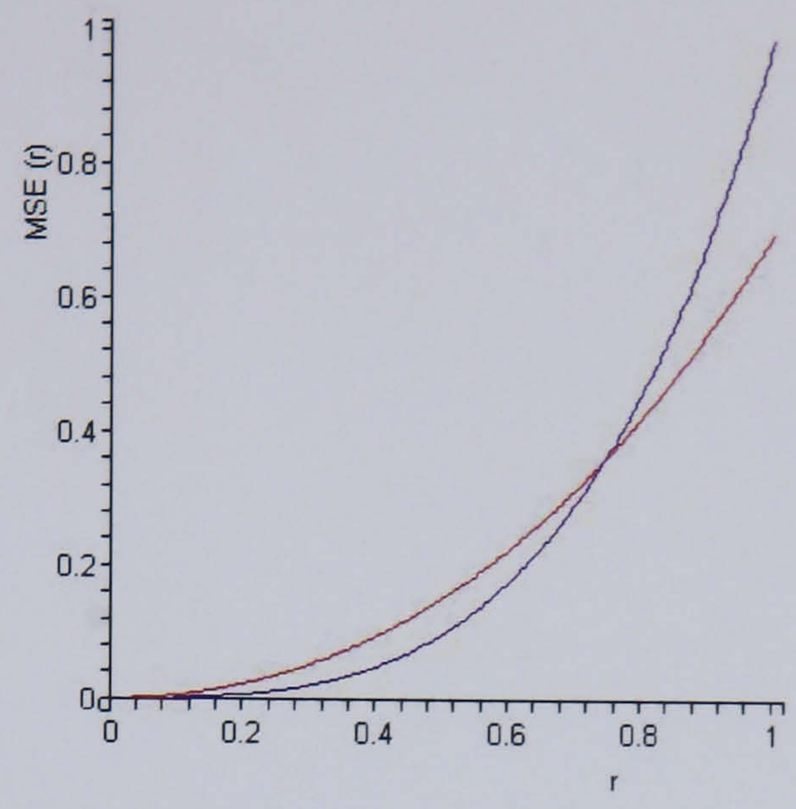


Figure 4.10: $MSE(r)$ for the two-dimensional (blue) and three-dimensional case (red)

defined as

$$RELBIAS(r) = \frac{BIAS(r)}{BIAS_0(r)}$$

where $BIAS_0(r)$ is the $BIAS(r)$ function for the point pattern when there are no censored events. As an example, for the $L(r)$-function this quantity is

$$\begin{aligned} BIAS_0(r) &= E(\widehat{L_0}(r) - r) \\ &= E\left(\sqrt[d]{\nu(b^d(o,1))^{-1}(n/X(W))^{-2} \sum_{\eta \in W_{\ominus l}, \zeta \in X \cap W} I_{(0<\|\eta-\zeta\| \leq r)}} - r \right). \end{aligned}$$

A different outcome is obtained for comparing edge effects in two- and three-dimensions using the $RELBIAS$ function (see Figure 4.11 and Figures E.11-E.12 of Appendix E). The effect on the summary statistic estimate as a result of having censored events is much greater in three-dimensions compared with two. Specifically, the curves in Figures E.11-E.12 demonstrate that the $RELBIAS$ is much greater in the three-dimensional case compared to 2, when we do not adopt an edge-correction for estimating a summary statistic.

The $RELBIAS$ is the preferred quantification of edge effects as it is independent of the level of departure from the theoretical value of the summary statistic. Furthermore, we are not required to make an adjustment to r for a better comparison. So far we have provided some experimental results for $BIAS(r)$ and $RELBIAS(r)$ obtained via simulation studies. We now consider the theoretical values of the

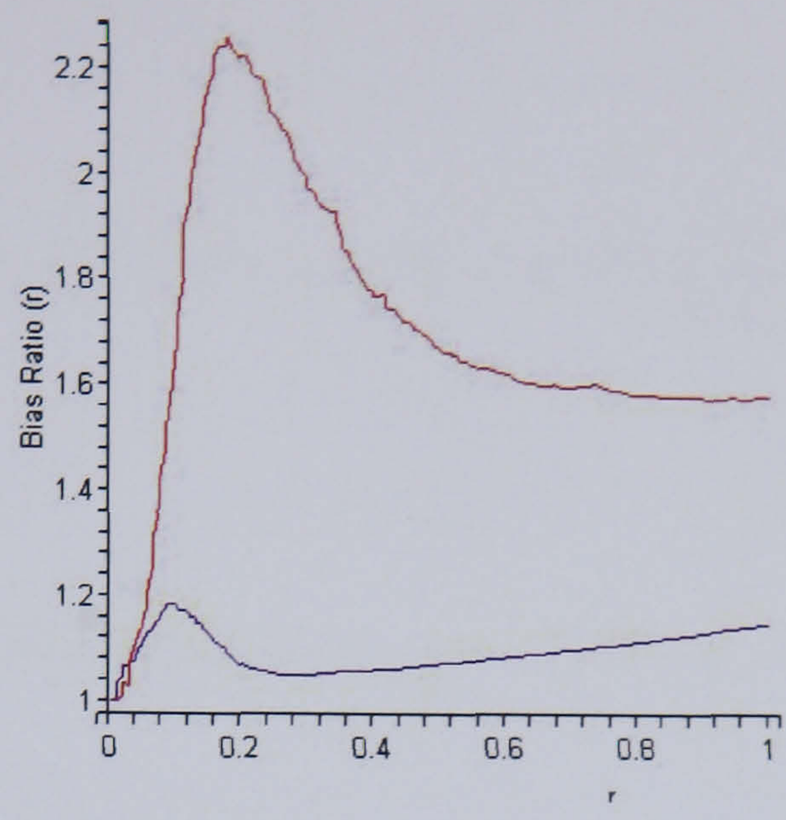


Figure 4.11: $RELBIAS(r)$ (the bias ratio) for the two-dimensional (blue) and three-dimensional (red) case produced through simulation studies (the standard scenario).

$BIAS(r)$ and $RELBIAS(r)$ functions for the two- and three-dimensional cases for the commonly used $K(r)$-function.

Suppose we have n_0 events of a homogeneous Poisson process, X_{poi} inside a square region W of length l_0. We wish to calculate

$$BIAS(r) = E(\widehat{K}(r) - K(r)) \tag{4.12}$$

where

$$\lambda^2 \widehat{K}(r) = \frac{1}{n} \sum_{\eta \in W_\ominus} \sum_{\zeta \in W_\ominus} I_{(0<\|\eta-\zeta\|\leq r)} \tag{4.13}$$

is an un-edge corrected estimate for the $K(r)$-function inside the eroded window, $W_\ominus$, of length l and n is the number of events in the eroded window (so that we mimic carrying out analysis in a sampling window that is a subset of a larger region on which the point pattern exists upon). Then for $BIAS(r)$ defined by (4.12) and (4.13) we have

Proposition 4

$$\begin{aligned} BIAS(r) \;=\; & \frac{1}{\lambda^2} \sum_{n=1}^{n_0} \sum_{k=0}^{n(n-1)} \frac{k}{n} \binom{n_0}{n} \binom{n(n-1)}{k} \times \\ & p_0^n (1-p_0)^{n_0-n} p^k (1-p)^{n(n-1)-k} - \pi r^2 \end{aligned} \tag{4.14}$$

where

$$p_0 = \frac{\nu(W_\ominus)}{\nu(W)} \tag{4.15}$$

$$p = \frac{1}{l^2}\left(\frac{20}{3}r^3 l - r^4(2+\pi^2) + \pi r^2(l-4r)^2\right)\left(\frac{\pi r^2}{l_0^2} - \frac{8r^3}{3l_0^3} + \frac{r^4}{2l_0^4}\right) \quad 4r \leq l \tag{4.16}$$

Proof.

$$E(\widehat{K}(r) - K(r)) = E(\widehat{K}(r)) - \pi r^2 \tag{4.17}$$

Let

$$Y_1 = \frac{1}{N}\sum_{\eta\in W_\ominus}\sum_{\zeta\in W_\ominus} I_{(0<\|\eta-\zeta\|\leq r)}$$

$$Y_2 = \sum_{\eta\in W_\ominus}\sum_{\zeta\in W_\ominus} I_{(0<\|\eta-\zeta\|\leq r)} \tag{4.18}$$

Where

$$N = \sum_{\xi\in X_{\text{poi}}} I_{(\xi\in W_\ominus)}$$

Then

$$P(Y_1 = y_1) = \sum_{n=0}^{n_0} P(N=n)P(Y_1 = y_1|N=n) \quad y_1 \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n(n-1)}{n}\right\}$$

$$= \sum_{n=0}^{n_0} P(N=n)P(Y_2 = y_2|N=n) \quad y_2 \in \{0,1,2,\ldots,n(n-1)\}$$

Now,

$$E(Y_1) = \sum_{y_1} y_1 P(Y_1 = y_1)$$

$$= \sum_{y_1} y_1 \left(\sum_{n=0}^{n_0} P(N=n)P(Y_2 = ny_1|N=n)\right)$$

$$= \sum_{n=1}^{n_0}\sum_{y_2=1}^{n(n-1)} \frac{y_2}{n} P(N=n)P(Y_2 = y_2|N=n) \tag{4.19}$$

We have that $N \sim \text{Binomial}(n_0, p_0)$, where

$$p_0 = \frac{\nu(W_\ominus)}{\nu(W)} \tag{4.20}$$

Also, $Y_2 = y_2|N = n \sim \text{Binomial}(n(n-1), p)$, where

$$\begin{aligned} p &= P\left(0 < \|\eta - \zeta\| \leq r \text{ and } \zeta \in W_\ominus | \eta, \zeta \in X_{\text{poi}}\right) \\ &= p_3 \int_0^r \frac{2\pi u}{{l_0}^2} - \frac{8u^2}{{l_0}^3} + \frac{2u^3}{{l_0}^4} du \end{aligned} \tag{4.21}$$

The integral in (4.21) is the probability that two events, uniformly and independently placed in a square of length l_0 are within distance r of each other. We obtained the integral (4.21) using the discussion by Diggle (2003, section 2.2), Taylor et al (2001) and Bartlett (1964). Further to this,

$$\begin{aligned} p_3 &= \frac{1}{\nu(W_\ominus)} \int_{W_\ominus} b^{(2)}(\zeta, r) \cap W_\ominus d\nu(\zeta) \\ &= \frac{4}{l^2}\left[\int_{\eta^{(2)}=0}^{\eta^{(2)}=r} \int_{\eta^{(1)}=\eta_0^{(1)}/2}^{\eta^{(1)}=l-\eta_0^{(1)}/2} A_1 d\eta^{(1)} d\eta^{(2)} + \int_{\eta^{(2)}=0}^{\eta^{(2)}=r} \int_{\eta^{(1)}=l-\eta_0^{(1)}/2}^{\eta^{(1)}=l+\eta_0^{(1)}} A_2 d\eta^{(1)} d\eta^{(2)} + \frac{\pi r^2 A_3}{4}\right] \\ &= \frac{4}{l^2}\left[\int_{\eta^{(2)}=0}^{\eta^{(2)}=r} d\eta^{(2)} (l - 2\sqrt{\eta^{(2)}(2r - \eta^{(2)})}) [r^2 \cos^{-1}\left(\frac{r-\eta^{(2)}}{r}\right)\right. \\ &\quad \left. -(r-\eta^{(2)})\sqrt{2r\eta^{(2)} - \eta^{(2)2}} + 2\sqrt{\eta^{(2)}(2r-\eta^{(2)})} r^2 \cos^{-1}\left(\frac{r-\eta^{(2)}}{r}\right) d\eta^{(2)} + \frac{(l-4r)^2}{4}\right] \\ &= \frac{1}{l^2}\left(\frac{20}{3} r^3 l - r^4(2+\pi^2) + \pi r^2 (l-4r)^2\right), \quad 4r \leq l \end{aligned} \tag{4.22}$$

Where $\eta_0 = 2r$. p_3 can be thought of as the probability that an event is uncensored, given that it is within distance r of an event in $W_\ominus$. A geometric description of the terms A_1, A_2 and A_3 are shown in Figure 4.12. Noting that $E\left(\frac{1}{\lambda^2} Y_1\right) = \frac{1}{\lambda^2} E(Y_1)$ and combining (4.17)-(4.22) we obtain the result. ■

A graph of (4.14) is shown in Figure 4.13.
Taking limits, we see that as $l_0 \to \infty$, $BIAS(r) \to -\pi r^2$ and as $r \to \infty$, $BIAS(r) \to -\infty$. To compute the theoretical $RELBIAS(r)$ given by

$$RELBIAS(r) = \frac{E(K(r) - K(r))}{E(\widehat{K}_+(r) - K(r))}$$

We use $p_3 = 1$ for the $K(r)$-function estimate for the events in $W_\ominus$ when there are no censored events (denoted $\widehat{K}_+(r)$). A theoretical $RELBIAS$ is plotted in Figure 4.14.

We can adjust (4.14) to apply to the three-dimensional case. What now follows is an outline of how we can obtain an approximation to the theoretical $BIAS(r)$ function (4.12) for the three-dimensional case. The argument is similar to that given

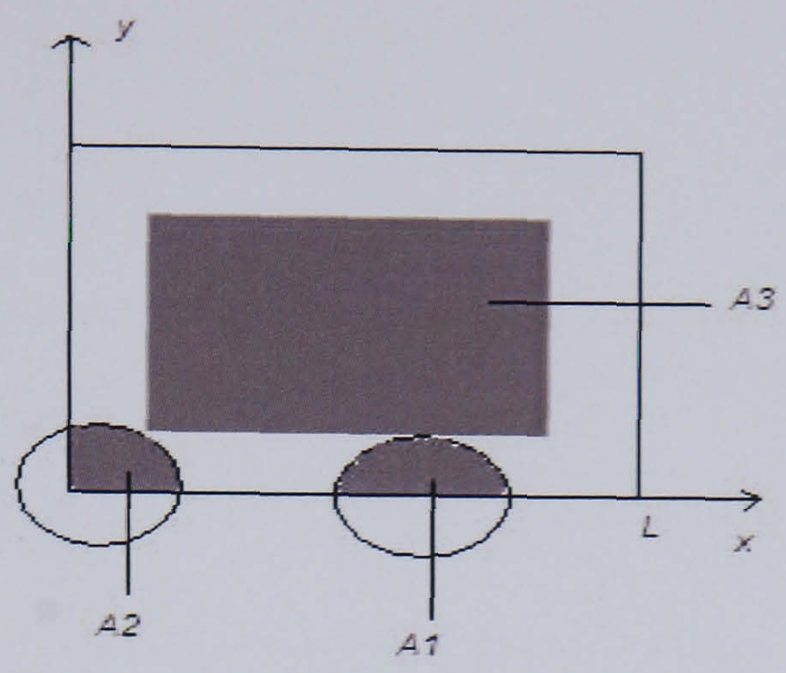


Figure 4.12: The geometric meanings of A_1, A_2, and A_3

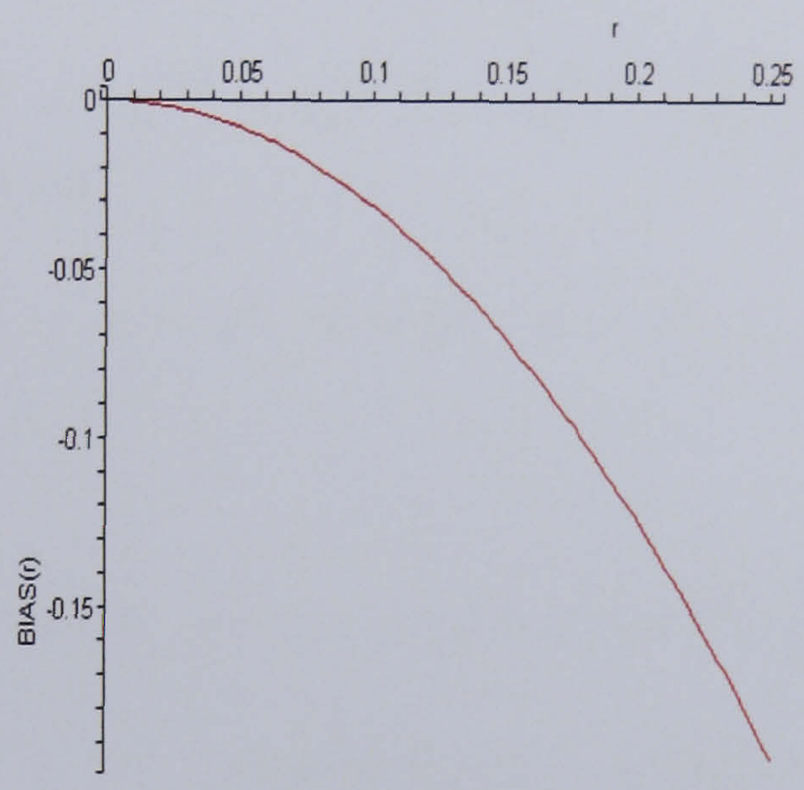


Figure 4.13: A plot of the theoretical $BIAS(r)$ for 100 events in a square window of length 3 and eroded window of length 1

in the derivation of the two-dimensional version (4.14). However, there are two steps that make the three-dimensional case more complicated. The first of these is the calculation of the probability that two events, η_1 and η_2, uniformly and independently placed in a cube of length L are within distance r of each other. To tackle this complicated step, we extend the approach adopted by Bartlett (1964) for the two-dimensional case. We have that the joint distribution of $|x|$, $|y|$, and $|z|$ is

$$f_{|X|,|Y|,|Z|}(|x|,|y|,|z|) = 8(L-|x|)(L-|y|)(L-|z|)$$

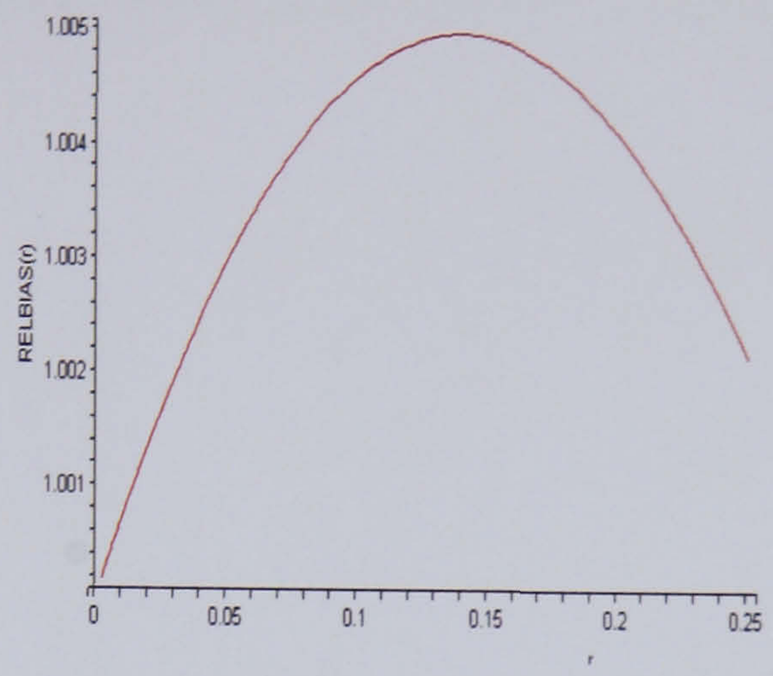


Figure 4.14: A plot of the theoretical $RELBIAS(r)$ for 100 events in a square window of length 3 and eroded window of length 1

where $r^2 = x^2 + y^2 + z^2 = (\eta_1^{(1)2} - \eta_2^{(1)2})^2 + (\eta_1^{(2)2} - \eta_2^{(2)2})^2 + (\eta_1^{(3)2} - \eta_2^{(3)2})^2$.
Making the transformation

$$u = x^2 \quad v = y^2 \quad w = z^2$$

gives

$$\begin{aligned} g_{U,V,W}(u,v,w) &= f_{|X|,|Y|,|Z|}(\sqrt{u},\sqrt{v},\sqrt{w}) \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| \\ &= \frac{1}{2\sqrt{u}\sqrt{v}\sqrt{w}}(L-\sqrt{u})(L-\sqrt{v})(L-\sqrt{w}) \end{aligned}$$

Transforming again ($k = r^2 = u + v + w$)and integrating to find the marginal distribution of r, we are left to evaluate

$$\frac{1}{4}\left[\int_0^t \int_0^{L^2} \int_0^{L^2} \frac{r}{\sqrt{r^2 - v - w}\sqrt{v}\sqrt{w}}(L-\sqrt{r^2-v-w})(L-\sqrt{v})(L-\sqrt{w})dvdwdr\right]$$

The integral appears difficult to evaluate. Numerical methods such as adaptive Gaussian quadrature (see for example Press et al (1992)) appear to fail. This is possibly due to continuity issues and machine accuracy. Specifically, during the numerical integration, some of the quantities under the square root may evaluate to negative terms. To approximate the integral, one can use a combination of Monte Carlo simulations and interpolation. To demonstrate this, we simulated two events uniformly in a cube 65,000 times and calculated the distance, d between them to form the dataset $D = \{d_1, \ldots, d_{65000}\}$. An empirical CDF, $\widehat{F}_G(r) \approx P$(two events

are within distance r) was then formed from the dataset D, using a linear interpolation to form a continuous curve. Examples of such curves are presented in Figure 4.15.

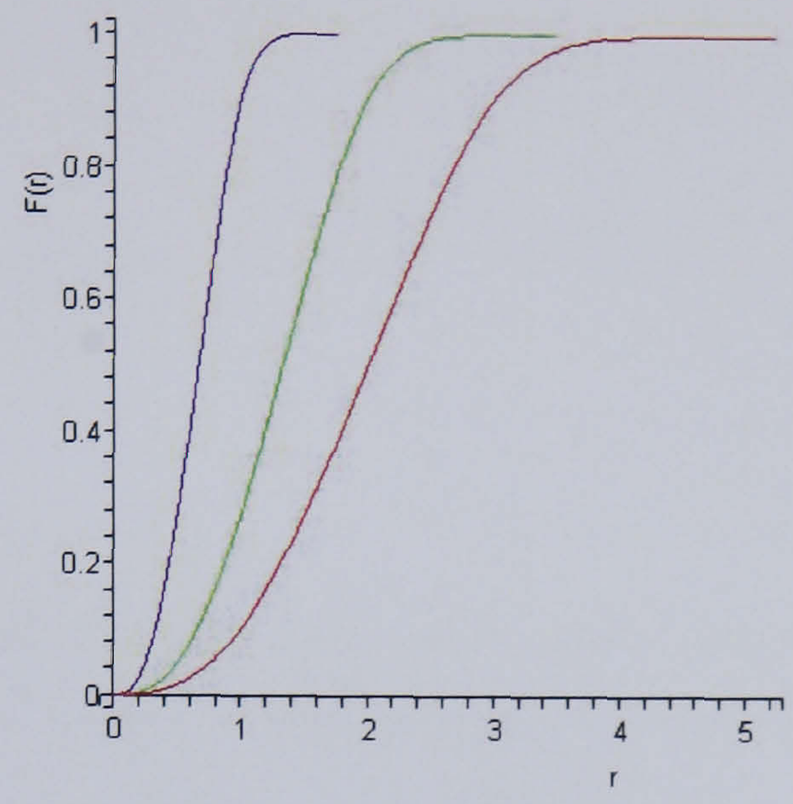


Figure 4.15: The empirical CDF, $\widehat{F}_G(r)$ of the distance between two events dropped in a cube of lengths 1 (blue), 2 (green), and 3 (red)

The second complication in obtaining an estimate for the $BIAS(r)$ function for the three-dimensional case is the need to calculate the quantity p_3, which, recall, heuristically, is the probability that an event is uncensored, given that it is within distance r of another event. We are able to evaluate p_3 for the three-dimensional case by integrating (the very complicated) expression given by Baddeley et al (1993, page 663, equation (27)) for the intersection volume of a brick and sphere. The dimensions of the cube, (and hence also r) need to be adjusted to meet the conditions (28) of the formula. To reduce numerical problems, we use modulus signs in the square roots in expression (34) on page 665 of Baddeley et al (1993). An alternative to using Baddeley et al's (1993) closed form expressions is to use Monte Carlo integration.

Following the above discussions of the two complicated steps required for computing the theoretical $BIAS$ for the three-dimensional case, we are in a position to calculate approximately the quantity (4.12) for the three-dimensional case. Figure 4.16 shows a plot of the theoretical $BIAS(r)$ for the two and three-dimensional case. The three-dimensional version uses Baddeley et al's (1993) closed form expression for the intersection volume of the cube and sphere. The theoretical $RELBIAS(r)$ functions are shown in Figure 4.17. Figure 4.17 and Figures E.11-E.12 in Appendix E show that using the $RELBIAS$ function for comparing the severity of edge effects

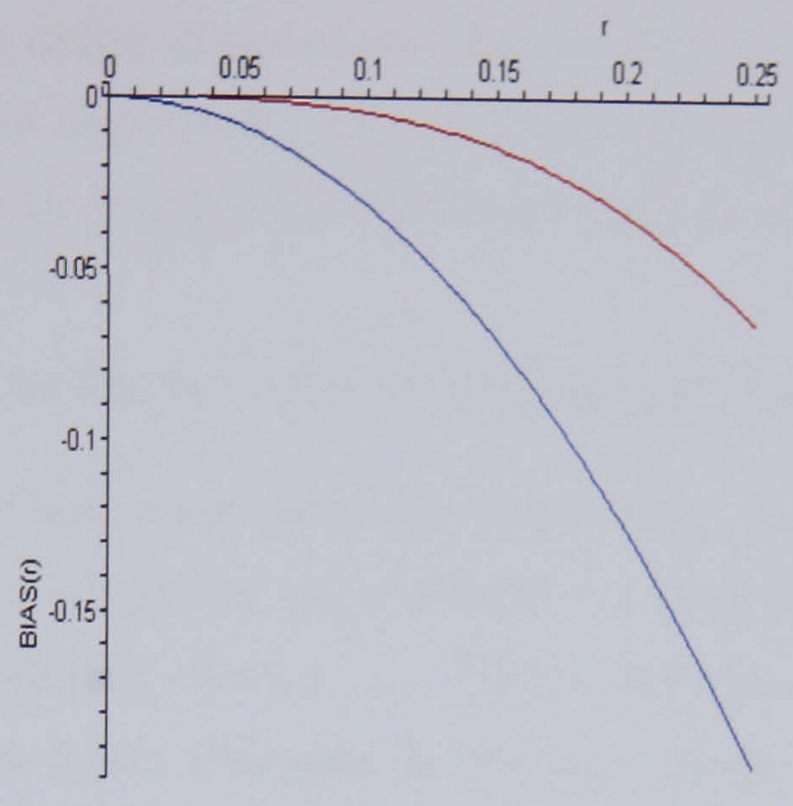


Figure 4.16: Theoretical $BIAS(r)$ for 100 events uniformly placed in the square (blue) and cube (red) of length 3 with eroded window of length 1

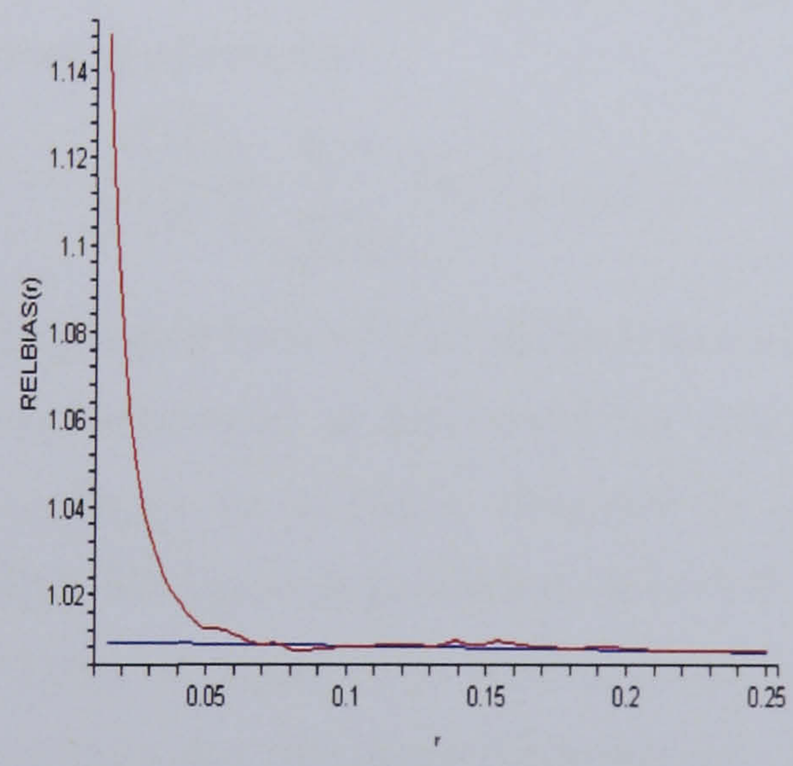


Figure 4.17: Theoretical $RELBIAS(r)$ for 100 events uniformly placed in the square (blue) and cube (red) of length 3 with eroded window of length 1

in two-dimensions with three-dimensions suggests that edge effects, as expected, appear greater in three-dimensions. More importantly, the plots provide some important and useful insight into how the difference in edge-effect severity varies with r, for a range of r, for two-dimensions compared with three-dimensions.

4.3 Techniques for edge-correcting

Having established the greater severity of edge effects in three-dimensions, we not turn to the various methods for edge-correcting. The most common approaches for

edge correcting include using estimators that explicitly account for edge effects, or using a naive buffer zone approach. The aim of edge-corrections is usually to obtain unbiased estimators, or at least ratio-unbiased estimators (Stoyan, 2006).

4.3.1 Edge-corrections via weighting factors

Perhaps the most common approach for correcting for edge effects involves estimators that explicitly account for edge effects by multiplying the unedge-corrected estimator by some weighting factor ω. These weights are sometimes referred to as Horvitz-Thompson weights (Stoyan & Stoyan, 2000). A good overview of these types of estimators in the three-dimensional case is provided by Stoyan & Stoyan (2000). In particular, they provide some experimental results regarding the use of these estimators when the sampling window is cubic.

One of the least complicated and widely used estimator of this kind, as shown here for the $K(r)$-function is given by

$$\widehat{K}(r) = \frac{\nu(W)}{X(W)^2} \sum_{\eta,\xi \in W} I_{(0<\|\eta-\xi\|\leq r)} \omega_1(\eta,\xi)^{-1} \tag{4.23}$$

with $\omega_1(\eta,\xi)$ equal to the proportion of the surface area or volume of $b^{(3)}(\eta, \|\eta-\xi\|)$, that lies within W. The estimator is unbiased for isotropic point processes. The analytical solution for $\omega_1(\eta,\xi)$ in a brick window is provided by Baddeley et al (1993), with the algorithm for its computation detailed by König et al (1991). For a spherical window the form of $\omega_1(\eta,\xi)$ is well known as it is simply the volume of intersection of two spheres (see for example Chapter 3). The computation of $\omega_1(\eta,\xi)$ is non-trivial for the ellipsoid window case. Estimator (4.23) has been adopted by practitioners such as König et al (1991) for point process applications.

The translation-corrected estimate due to Miles-Lantueoul-Stoyan-Hanisch is

$$\widehat{K}(r) = \frac{\nu(W)^2}{X(W)^2} \sum_{\eta,\zeta \in W} I_{(0<\|\eta-\zeta\|\leq r)} \omega_2(\eta,\zeta) \tag{4.24}$$

where

$$\omega_2(\eta,\zeta) = \frac{1}{\nu(\{\xi \in W : \xi + (\eta-\zeta) \in W\})}.$$

is valid for second-order stationary point processes. Stoyan (2006) states that the reduced second-order moment measure $\lambda\mathcal{K}(B)$ can be estimated using

$$\kappa(B) = \sum_{\eta,\xi \in X \cap W} \frac{I_{(\eta-\xi \in B)} I_{(\{\eta\} \cap \{\xi\} \equiv \emptyset)}}{\nu(W_\eta \cap W_\xi)}$$

for bounded Borel set B. The estimator is unbiased (see Stoyan et al (1995) for a proof). If the process is known to be isotropic then an improved alternative for the estimator for $0 \leq r < r^*$, where

$$r^* = \sup\{r : \nu(W^{(r)}) > 0\}$$

and

$$W^{(r)} = \{\xi \in W : \partial S(\eta, r) \cap W \neq \emptyset\}$$

is

$$\kappa(B) = \sum_{\eta,\xi \in X \cap W} \frac{I_{(0<\|\eta-\xi\|\leq r)}\gamma_0(\eta,\xi)}{\nu(W^{\|\eta-\xi\|})}$$

where

$$\gamma_0(\eta,\xi) = \frac{4\pi}{\alpha_{\eta,\xi}}$$

and $\alpha_{\eta,\xi}$ is the sum of all central angles that belongs to arcs of the ball centred at η, radius $\|\eta - \xi\|$. Moreover, in the isotropic case (4.23) can be improved for larger r (see Baddeley et al (1993)) to obtain the estimator

$$\widehat{K}(r) = \frac{\nu(W)}{X(W)^2} \sum_{\eta,\zeta} I_{(0<\|\eta-\xi\|\leq r)}\omega_1(\eta,\zeta)^{-1}\gamma_1(\eta,\zeta) \tag{4.25}$$

where

$$\gamma_1(\eta,\zeta) = \frac{\nu(W)}{\nu(\{\xi \in W : \partial S(\eta, r) \cap W \neq \emptyset\})}$$

and $\partial S(\eta, r)$ is the surface of $b^{(3)}(\eta, r)$. Estimate (4.25) is consistent and approximately unbiased (Baddeley et al, 1993).

Figure 4.18 is a plot of $\widehat{K}(r)$ using (4.23) and $\omega_1(\eta,\zeta) \equiv \omega(\eta,\zeta) = \nu(b^{(3)}(\eta, \|\eta-\xi\|) \cap W)$ for events uniformly placed inside a sphere of radius 0.62 (and hence the sphere has unit volume). Figure 4.18 is a visual demonstration that the estimator (4.23) is approximately unbiased for values less than between about 0.3 and 0.4. This is approximately half the radius of the sphere, and is hence consistent with some of Diggle's (2003) comments on the use of the estimator. Specifically, Diggle (2003) reports that the estimator (4.23) is unbiased for sufficiently small r. This restriction on r arises because the weighting factor $\omega(.)$ can become unbounded as r increases. Moreover, when the sampling region is the unit square, the theoretical upper limit

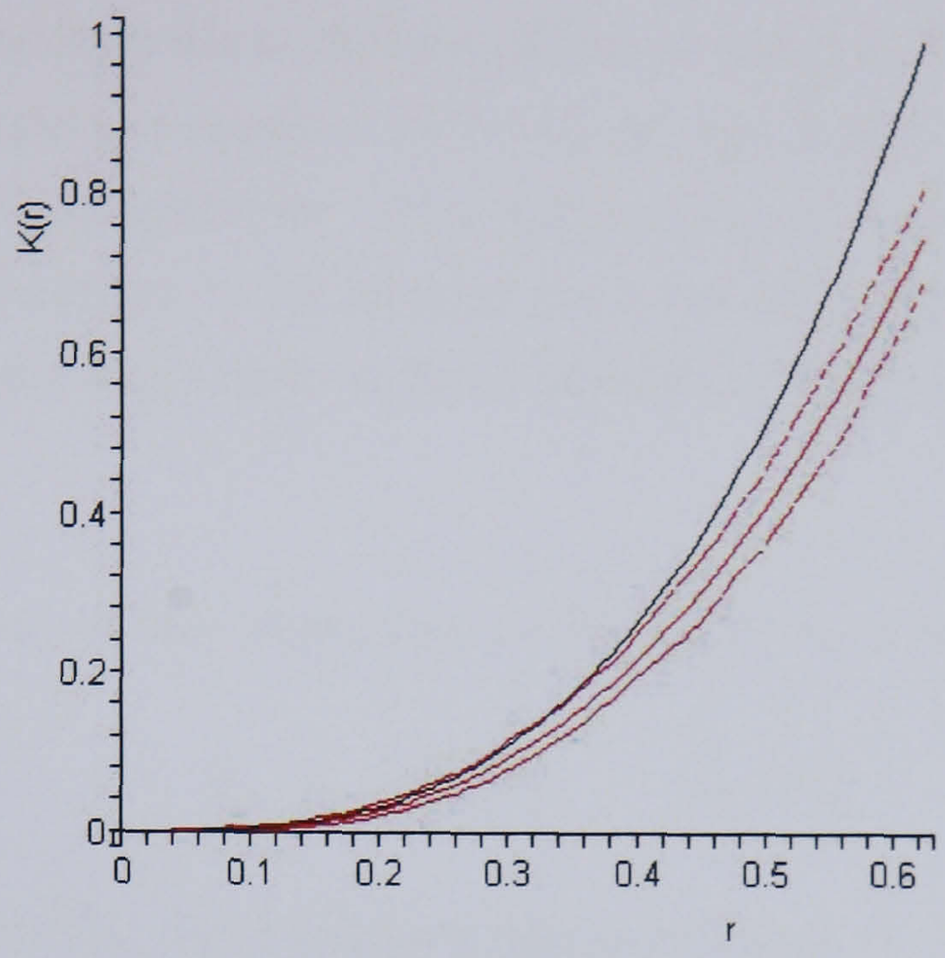


Figure 4.18: A plot of $\widehat{K}(r)$ (using estimator (4.23) (red)) with simulation envelopes based on 100 simulations (red dashed curves) of 100 events uniformly placed in a sphere of radius 0.62. The theoretical $K(r)$-function curve (black line) for a homogeneous Poisson process is also provided for comparison

of r is $\frac{\sqrt{2}}{2}$. (4.23) and (4.24) were used by Baddeley et al (1993) for analysing a point pattern defined in a brick window. In Baddeley et al's (1993) practical investigation, it was concluded that the estimators performed comparably.

Although analytical expressions for $\omega_1(.)$ in the case where W is an ellipsoid can be difficult to come by, we can adopt numerical techniques such as Monte Carlo integration in order to obtain approximations. The results in Table 4.1 are a demonstration of the estimation accuracy obtained when estimating the $K(r)$-function for different type of spatial point patterns, using (4.23) (with $\omega_1(\eta, \zeta) = \omega(\eta, \zeta)$, calculated by Monte Carlo integration) for W an ellipsoid window with semi-axis lengths 1.0, 0.5 and 0.25. In Table 4.1 we have used Poisson(7) to denote a Poisson process with

Table 4.1: Experimental results for applying a weighting factor edge-correction

Process	$\sigma_{T_{\text{MIN}}}$	$\sigma_{T_{\text{MAX}}}$	$\sigma_{\overline{T}}$
Poisson(7)	100	2×10^{11}	42
Matérn(250, 30, 0.05)	100	74	11
Hardcore(7,0.1)	157	4	64

intensity 7, Matérn(250,30,0.05) a Matérn cluster process with parent intensity 250,

daughter intensity 30, and radius of interaction 0.05. Finally, Hardcore(7,0.1) is a simple inhibition process with underlying Poisson process of intensity 7 and hardcore distance 0.1. The expected number of events for the Poisson(7) and Hardcore(7,0.1) processes is 4 and 5 respectively. The motivation behind considering such a low expected number of events is for insight into the performance of the estimator in the case where we have the lowest typical number of PML NBs in a cell nucleus (see Chapter 1).

A detailed description of the quantities in Table 4.1 is as follows. Let

$$T = \{\widehat{K}_1(r), \ldots, \widehat{K}_N(r)\}$$

denote N estimates of the $K(r)$-function obtained from $N = 500$ replicates of spatial point process X, then the quantities $\sigma_{T_{\mathrm{MIN}}}$, $\sigma_{T_{\mathrm{MAX}}}$, and $\sigma_{\overline{T}}$ in Table 4.1 denote (respectively) the quantity obtained by replacing $\widehat{S}(r)$ with $\sup\{T\}$, $\inf\{T\}$, and $N^{-1}\sum_{i=1}^{N}\widehat{K}_i(r)$ in (4.10). Note that undefined values of $\widehat{K}_i(r)$ were ignored. These arise when fewer than 1 events are generated in the simulation process. Furthermore, for the Hardcore process we use

$$\widehat{\sigma_S} = \sum_{i=1}^{u_k}\left(S(\Delta i) - \widehat{S}(\Delta i)\right)^2$$

in place of (4.11) in order to avoid a division by zero. The results in table 4.1 show that, in general, the various quantities used to assess the level of departure between an estimated value of the $K(r)$-function, with a theoretical one, agree that the level of departure is higher for low intensity spatial point processes, when using (4.23) as an edge-correction.

Simulating 10,000 realisations of a homogeneous Poisson process conditional on 30 events (this is the upper limit of the typical number of PML in a cell nucleus) inside the unit cube, followed by estimating the unedge-corrected $K(r)$-function took 20 seconds on a modern computer with a 2.8GHz Pentium (IV) processor. The computation time was increased by 45% when using the volume weighting factor edge-correction, with the weighting factor estimated using a Monte Carlo integration. This demonstrates how (computationally) costly applying an edge-correction can be. In this case, the additional computation time that is required to carry out the edge-correction can be eradicated when a closed form for the edge-correction term is adopted.

4.3.2 Kaplan-Meier estimation

The use of Kaplan-Meier type estimators is also an explicit form of edge-correction. These estimators are an innovative alternative to the use of weighting factors. The Kaplan-Meier approach to edge-correction was introduced by Baddeley & Gill (1994). Reed & Howard (1997) applied this type of estimation to the nearest neighbour distance distribution function

$$\widehat{G}(r) = 1 - \prod_{m=1}^{j} \frac{\sum_{\eta \in W} I_{(s_i = r_m, s_i \leq b_i)}}{I_{(s_i \geq r_m, b_i \geq r_m)}}$$

where s_i and b_i denote the distance from each event in W, to the nearest other event in W and to the nearest point on the boundary of W respectively. Also, $s_i = r_m$ is defined as $r_m \leq s_i < r_{m+1}$ and we discretise r, $r_j = j\Delta$ and Δ is the increment of the distance. Reed & Howard (1997) found good results for a higher number of points per unit volume, when using the estimator in simulation studies. Baddeley & Gill (1997) compared a Kaplan-Meier estimator of the empty space function, $F(r)$ to the *minus sampling* method through simulation studies. They found that for a Poisson process in a square window, the Kaplan-Meier estimator appeared to be uniformly more efficient and in general performed better. The more theoretical properties of the Kaplan-Meier estimator of the empty space function are discussed by Gill (1994).

4.3.3 Minus sampling (the "border" method)

The border method is a simple and most commonly used form of edge-correction (Floresroux & Stein, 1996). The method was first suggested by Ripley (1977), but is analogous to the reduced sample estimator, discussed by Kaplan and Meier (1958). The idea behind the border method is to restrict attention (when estimating $F(r)$, $G(r)$ or $K(r)$ at distance r) to events more than distance r from the boundary (Gill. 1994). More precisely, we only consider events in the eroded window

$$W_{\ominus r}. \tag{4.26}$$

For the case where W is a brick window, (4.26) is straightforward to calculate. and even more so for a sphere. Figure 4.19 shows the geometrical interpretation of (4.26) for W a brick window. The border method has been reported to having an inferior performance compared to other edge-corrections in certain analysis (see for example Schladitz & Baddeley (2000)). This is usually a result of the technique requiring important information from the point pattern to be discarded. In some exceptional

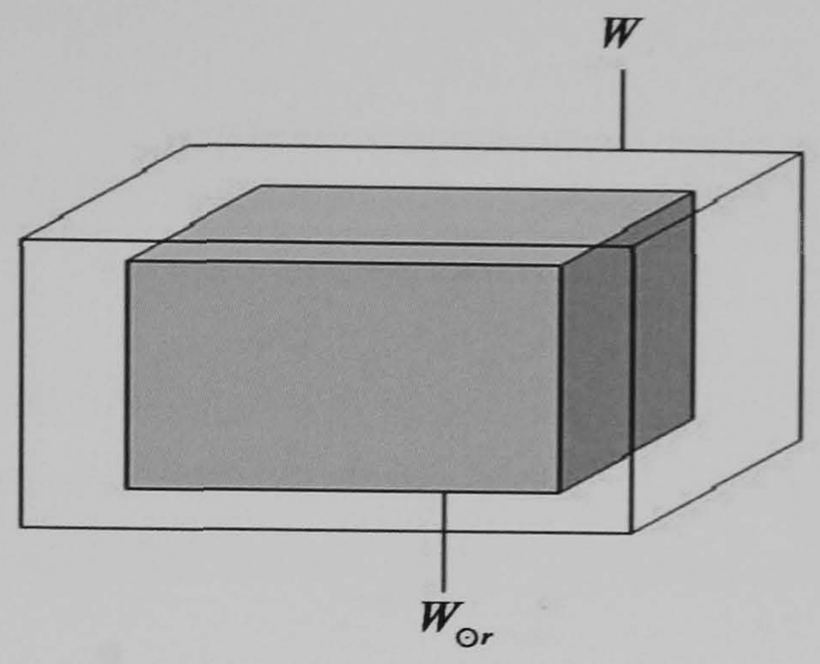


Figure 4.19: A brick window and eroded window (grey)

cases, the method is rendered useless if the majority of the events are located close to the boundary of W. However, at the other extreme (when most of the events are located far from the boundary) the border method may be a good method to adopt as it is can be computed quickly and is usually straight-forward to implement.

Figure E.16 compares the adjusted $MISE$ obtained when estimating the unedge-corrected $Q(r, \tilde{r})$-function and the border estimate of the $Q(r, \tilde{r})$-function (see Chapter 3). Clearly, in some cases a border estimate can significantly reduce the adjusted $MISE$ compared to not applying an edge-correction. Specifically, Figure E.16 demonstrates that applying even a simplistic edge-correction such as the border method, when estimating the $Q(r, \tilde{r})$-function, can have benefits such as reducing the $MISE$.

4.3.4 Plus sampling

Plus sampling is almost the opposite of minus sampling in the sense that, while the border method requires one to discard some events, a plus sampling estimate requires additional information regarding events that lie outside of the estimation window W. However, one could also loosely describe a plus sampling estimate as a border estimate with the entire sampling window being the eroded window and the dilated window $W_{\oplus r}$ as the sampling window. As an example, consider the plus sampling estimation of the $K(r)$-function; in this case, when counting the number of events within distance r of events inside W, we would also count any events which lie outside of W. That is, for spatial point process X, the plus sampling estimate is

given by

$$\widehat{K}(r) = \frac{\nu(W)}{X(W)^2} \sum_{\eta \in W, \zeta \in X} I_{(0<\|\eta-\zeta\|\leq r)}. \tag{4.27}$$

Plus sampling is unbiased (Stoyan et al, 1995) but of course, realistically, has limited application. Clearly the problem with adopting the plus sampling approach is that the method requires more information than might be available.

4.3.5 Probabilistic edge-correction

Reed & Howard (1997) describe a probabilistic edge-correction for the three-dimensional nearest-neighbour distance distribution function $G(r)$.

$$\widehat{G}(r) = \frac{\sum_{\eta \in W \cap X} I_{(s_\eta \leq r)} + \phi(E_\eta)}{X(W)}, \tag{4.28}$$

where $\phi(a) = a$ if a < 1 and $\phi(a) = 1$ if a ≥ 1. Furthermore, E_η is equal to 1 if $s_\eta \leq b_\eta$ and $\nu(b^d(\eta, \rho(\eta, X\backslash\{\eta\}))\backslash(W \cap b^d(\eta, \rho(\eta, X\backslash\{\eta\}))))$ (this is the shaded region shown in Figure 4.20 with r equal to the distance between the event η and its nearest other event) otherwise. They also describe a related estimator developed by Doguwa & Upton (1990) for two-dimensional point patterns. The outcome of simulation studies carried out by Reed & Howard (1997) inside a brick window was that the probabilistic edge-correction (4.28) gave similar results to the one suggested by Doguwa & Upton (1990). Furthermore, an increase in the number of replicates did not make the average value of $\widehat{G}(r)$ closer to $G(r)$. Further to this, a higher number of points per brick did not improve the estimator. The problem with the estimator proposed by Doguwa & Upton (1990) is that it assumes certain properties of the point process are the same as that of a homogeneous Poisson process (even when this is not necessarily the case). Floresroux & Stein (1996) suggested an improved version of the estimator of Doguwa & Uptons (1990). Their improved estimator does not require the properties of a homogeneous Poisson assumption.

We have seen so far that the principal aim of an edge-correction is to reduce the estimation bias. The implementation of the edge-correction in terms of computational time and ease of implementation is also important. With this in mind, we will introduce two edge-correction techniques of our own. The first edge-correction that we propose is a probabilistic edge-correction estimator for the $K(r)$-function.

The estimator is of the following form, $\lambda^2 \widehat{K}_{\text{Prob}}(r)$ for $\lambda^2 K(r)$:

$$\lambda^2 \widehat{K}_{\text{Prob}}(r) = \sum_{\eta \in X \cap W} \left(\sum_{\zeta \in X \cap W} I_{(0<\|\eta-\zeta\| \leq r)} \right) + \left(\nu(b^{(3)}(\eta, r) \backslash \{b^{(3)}(\eta, r) \cap W\}) \right) X(W) \nu(W)^{-1}, \quad (4.29)$$

For our new estimator (4.29), we are counting the number of events that we observe within distance r of an event η and we also consider the number of events that we expect to observe (given the number of events in W) in the shaded region shown in Figure 4.20.

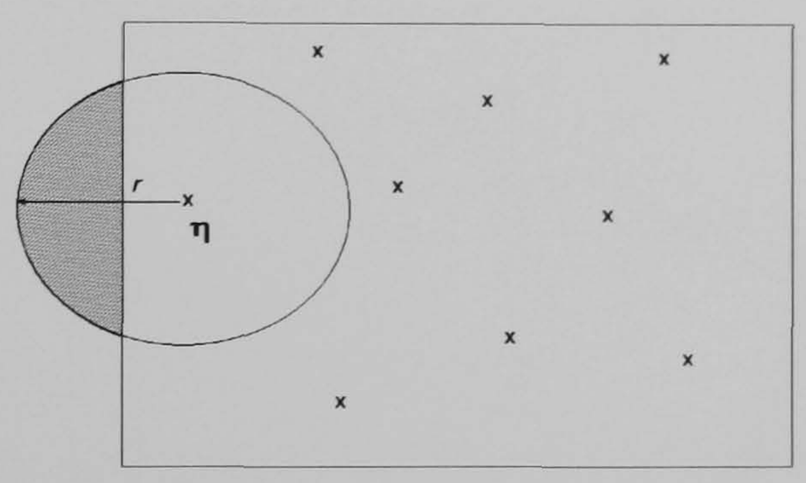


Figure 4.20: Geometrical illustration of estimator (4.29). The crosses represent events of the spatial point process

Proposition 5 *Estimator (4.29) is asymptotically unbiased for a homogeneous Poisson point process.*

Proof. Using the definition of the factorial kth moment measure (see Chapter 2), for Borel set B and the second reduced moment measure $\mathcal{K}$ we have

$$\begin{aligned} I_1 &= E\left(\sum_{\eta \in X \cap W} \sum_{\zeta \in X \cap W} I_{(0<\|\eta-\zeta\| \leq r)} \right) \\ &= \int \int I_{(\|\xi\| \leq r)} I_{(\varsigma \in W)} I_{(\varsigma+\xi \in W)} d\varsigma \mathcal{K}(d\xi) \\ &= \lambda^2 \mathcal{K}(b^{(3)}(\xi, r) \cap W) \end{aligned}$$

Noting that for an ergodic spatial point process, if $b^{(d)}(o, r) \subset W$, then as $r \to \infty$ we have

$$E\left(\frac{x(W)}{\nu(W)}\right) = \lambda$$

(see for example Stoyan & Stoyan, 1994, p194). We have that

$$\begin{aligned} I_2 &= E\left(\sum_{\eta\in X\cap W} \nu(b^{(3)}(\eta,r)\backslash\{b^{(3)}(\eta,r)\cap W\})X(W)\nu(W)^{-1}\right) \\ &= \lambda^2\mathcal{K}(b^{(3)}(\eta,r)\backslash\{b^{(3)}(\eta,r)\cap W\}) \end{aligned}$$

For the homogeneous Poisson process case we can replace $b^{(3)}(\eta,r)$ with $b^{(3)}(o,r)$ and obtain $\lambda^2 K(r)$ by summing I_1 and I_2. ■

Figure 4.21 shows simulation results (100 events uniformly placed in the sphere) for using (4.29) as an edge-correction with $\widehat{\lambda^2} = X(W)^2/\nu(W)^2$. Figure 4.21 suggests

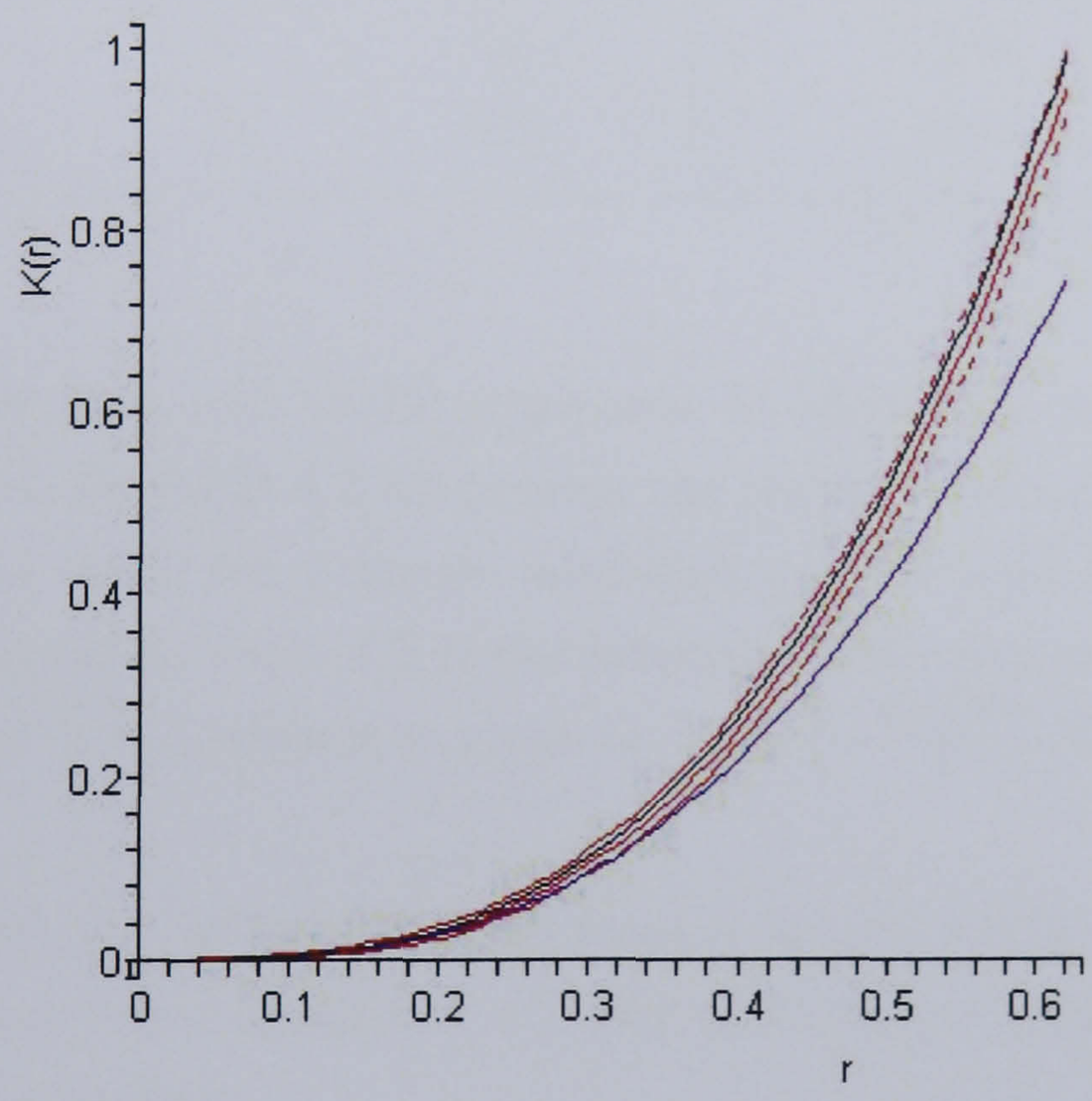


Figure 4.21: A plot of $E(\widehat{K}_{\text{Prob}}(r))$ (red) with simulation envelopes (red dashed curves) for $\widehat{K}_{\text{Prob}}(r)$ compared with the weighted edge-correction estimate (4.23) (blue) and the theoretical curve for a homogeneous Poisson process (black)

that in this case the new probabilistic estimator (4.29) outperforms the weighting factor edge- correction in terms of accuracy and variance. The difference is more apparent for larger r.

Tables 4.2 provides the results of a further comparison between estimators (4.23) and (4.29) for the homogeneous Poisson process case. The results in Table 4.2 give rise to further evidence that in certain cases, the new probabilistic edge-corrected

estimator for the $K(r)$-function, that we have proposed, can be a more desirable choice of estimator compared to (4.23). More precisely, for different assessments of departure between an estimated and theoretical value of the $K(r)$-function, estimates based on the new probabilistic edge-correction generally provide a smaller degree of departure.

Table 4.2: Results for 500 simulations from a homogeneous Poisson process using the weighted edge-correction (4.23, labelled 'A') and the probabilistic edge-correction (labelled 'B') for estimating the $K(r)$-function

Edge-correction	λ	$\sigma_{K\text{MIN}}$	$\sigma_{K\text{MAX}}$	$\sigma_{\overline{K}}$
A	50	18	7295	6
A	100	13	944	5
B	50	13	7290	3
B	100	9	607	3

We now consider how well (4.29) compares to (4.23) for non-stationary Poisson distributed events. In Table 4.3 we present the results of simulation studies for comparing (4.23) and (4.29) for a simple inhibition process with hardcore distance 0.1. The intensity λ given in Table 4.3 is the intensity of the underlying Poisson process, the true intensity of the process is given by $\lambda \exp\left(-\frac{4\lambda\pi\delta^3}{3}\right)$. Here σ_S^0 is given by

Table 4.3: Results for 500 simulations from a simple inhibition process using the weighted edge-correction (labelled 'A') and probabilistic edge-correction (labelled 'B') for the $K(r)$-function

Edge-correction	λ	$\sigma^0_{K\text{MIN}}$	$\sigma^0_{K\text{MAX}}$	$\sigma^0_{\overline{K}}$
A	50	81	113	94
A	100	105	134	118
A	250	107	131	1
B	50	148	173	158
B	100	179	203	188
B	250	180	200	1

$$\sigma_S^0 = \frac{1}{u}\sum_{1}^{u}(S(r) - S_{\text{Hardcore}}(r))^2$$

where $S_{\text{Hardcore}}(r)$ is the theoretical S function of the simple inhibition process.

We present simulation results for comparing (4.29) and (4.23) for the case of a

Table 4.4: Results for 500 simulations from a Matérn cluster process using the weighted edge-correction (labelled 'A') and the probabilistic edge-correction (labelled 'B') for estimating the $K(r)$-function

Edge-correction	λ_p	σ_{MIN}	σ_{MAX}	$\sigma_{\overline{K}}$
A	50	100	288	60
A	100	100	70	43
A	250	100	293	34
B	50	98	155	50
B	100	86	54	37
B	250	85	268	29

Matérn cluster process with fixed daughter intensity 30 and radius of interaction 0.05 in Table 4.4. In summary, the results provided in Tables 4.2-4.4 demonstrate that the new probabilistic estimator works well for non-hardcore models compared to (4.23). In particular, the results in Table 4.3 suggest that there is a larger discrepancy between the theoretical, and the estimated (using the new probabilistic method, (4.29)) $K(r)$-function of a simple inhibition process compared to edge-corrected estimation based on using (4.23). We would expect this result since, for inhibition models, there is a danger that the new probabilistic method over-estimates the number of events that one would expect to observe in the shaded region in Figure 4.20. Unfortunately, the new estimator (4.29) is also remarkably more computationally expensive to compute compared to the weighting factor edge-correction, when a Monte Carlo approximation for computing $\nu(W \cap b^{(3)}(o,r))$ is required. Computing the probabilistic $K(r)$-function estimate of a single realisation from a homogeneous Poisson process with 30 events took 90 minutes. Recall from earlier (Section 4.3.1) that the same computation based on the weighting edge-correction, took 29 seconds for 100 realisations, on the same machine.

Figure E.17 (see Appendix E) shows plots of $K(r) - \frac{4}{3}\pi r^3$ for the PML NB data in PPDS1, estimated using the (volume) weighting factor edge-correction and the new probabilistic edge-correction that we have introduced. The probabilistic $K(r)$-function edge-correction yields results that are consistent with the CSR tests carried

out in Chapter 1. More specifically, for cell 1, both estimates suggest that there are fewer PML NBs within distance $r \leq 2$ units of a randomly selected PML NB compared to CSR, indicating low scale inhibition. However, for cells 3-5 there is some strong conflicting information provided by the different estimators for large values of r. For example, for cell 5, the $K(r)$-function estimate based on the probabilistic estimator suggests that for $r \geq 1.5$ there are more PML NBs within distance r of a randomly chosen PML body compared to CSR. This is consistent with the signs of clustering seen in the CSR results for cell 5 in Chapter 1. However, the results based on using the weighting factor edge-correction (4.23) suggest that for cell 5, there are fewer PML within distance $r \geq 1.5$ units compared to CSR.

4.3.6 The probabilistic border edge-correction

As mentioned earlier, the border method is easy to implement and although it suffers from being strongly biased downwards, can perform competitively compared to its more complicated rivals such as the explicit forms of edge-correction. The second edge-correction we propose is an improved border method that incorporates the probabilistic edge-correction idea. More precisely, we increase the size of the window W (and hence implicitly assume our point pattern is a subset of a process defined on a large set $V \supset W$). We do this to include regions where there are unlikely to be unobserved events. The motivation behind this approach is to reduce the number of events in W that we must discard. We expect the new method to produce better results than the border method, especially for stationary hardcore processes. For this particular case, we expect fewer points to be discarded in the edge-correction process compared with the standard border method. The probabilistic border method that we propose, for the $K(r)$-function case is given by

$$\widehat{K}(r) = \frac{\nu(W_{\ominus r})}{X(W_{\ominus r})^2} \sum_{\eta \in W_{\ominus r}^{\oplus r^*}} \sum_{\zeta \in W_{\ominus r}^{\oplus r^*}} I_{(0<\|\eta-\zeta\|\leq r)} \tag{4.30}$$

where

$$W^{\oplus r^*} = W \cap \bigcup_{\eta \in \partial W} b(\eta, r^*) \tag{4.31}$$

and $r^* = \inf\{d_1, d_2\}$, where d_1 is the maximum of 0 and the expected distance between the typical point and its nearest neighbour, $E_{\rho(\eta,X)}$, minus the distance between the boundary ∂W and the event in W closest to the boundary. Also, d_2 is the maximum of 0 and the distance between the event ξ, in W closest to the boundary of W and the event in W closest to ξ, minus the distance between ξ and the

boundary. In other words, we extend the sampling window, W, in such a way that the event closest to the boundary, ξ, is at least distance $E_{\rho(\eta,X)}$ from the boundary of W, unless there is an event that is within distance $E_{\rho(\eta,X)}$ to ξ. r^* is generally unknown and is estimated from the point pattern by calculating the average of the observed nearest neighbour distances and shortest distance to the boundary of the sampling window. Clearly, in certain situations (4.30) will collapse to the usual border estimate.

Figure 4.22 demonstrates that, as anticipated, the probabilistic border method is less biased downwards compared to the standard border method. Further compar-

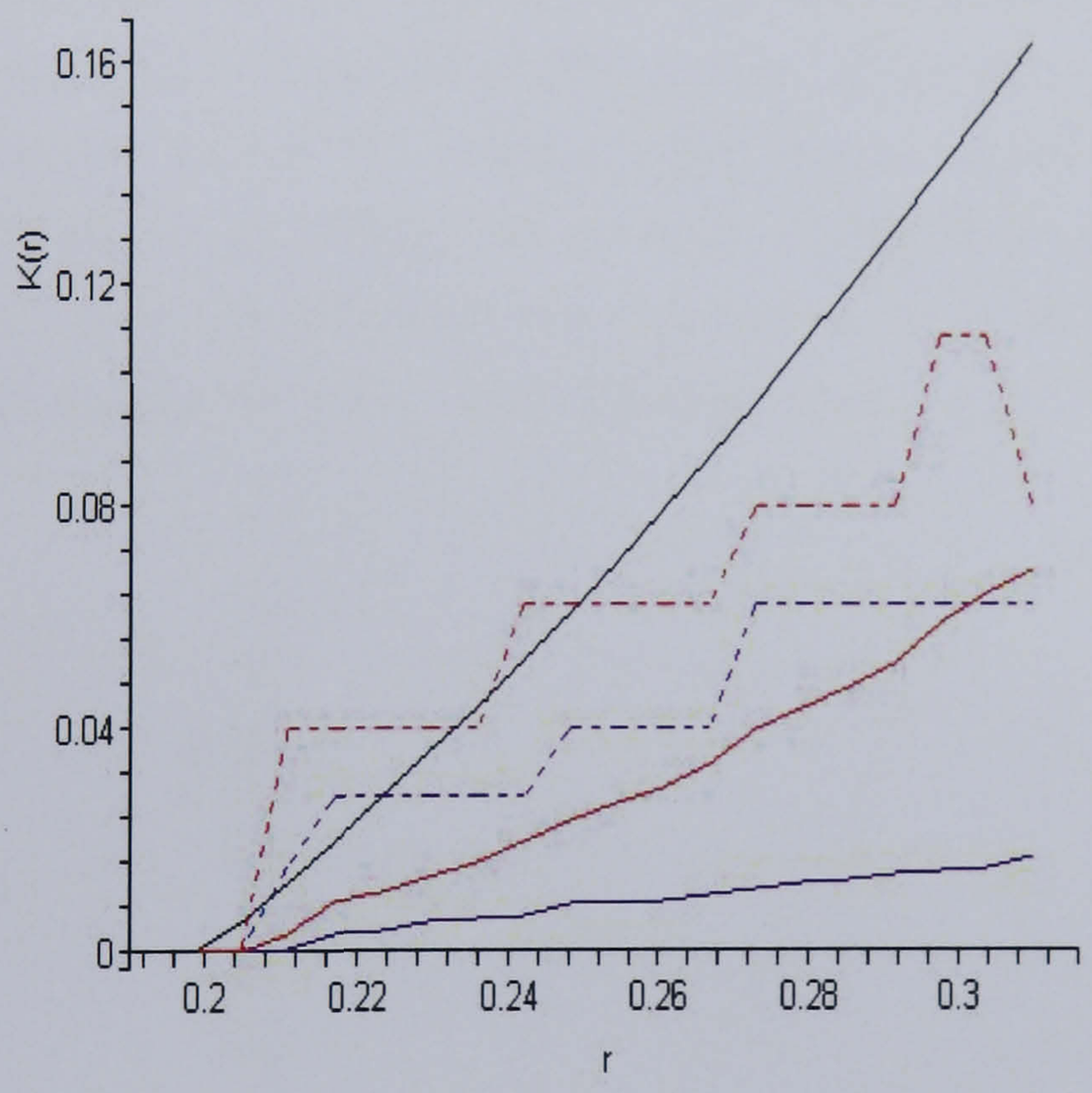


Figure 4.22: The mean probabilistic border estimate of the $K(r)$-function (red) compared to the mean of the usual border method (blue) and theoretical curve (black) using the standard scenario. The respective upper simulation envelopes for the border estimates are given by the red and blue dashed curves

isons between the border method and probabilistic border method are presented in Figures E.13-E.15 in Appendix E. The graphs in Figures E.13-E.15 show that in this case, the most notable improvement in terms of reduction in bias when using the probabilistic border method in place of the standard border method occurs when the spatial point process is a homogeneous Poisson. Although the graphs E.14-E.15 show comparable results between the border method and probabilistic border when

considering the simple inhibition process and Matérn cluster process, the data used to generate the curves indicated some slight reduction in bias, for both processes, when using the probabilistic border method instead of the standard border method.

4.4 Edge-correction for irregular windows

In many practical applications (such as that discussed in the next chapter) the sampling window W can not be well approximated by some regular body such as a cuboid or ellipsoid. For PPDS2 and PPDS3, the sampling window is an irregular one, with a boundary that is approximated by a set of discrete points. This discrete set of points form a convex hull. In order to asses the accuracy of the edge-correction methods discussed, for the case where the sampling window is approximated by a convex hull, we simulated N points on the surface of an ellipsoid. We subsequently calculated the convex hull of the points, using the C implementation of the qhull algorithm (see Barber et al, 1995). An estimate of the $K(r)$-function of 100 events uniformly placed inside the ellipsoid was calculated using the border method. The results are shown in Figure 4.23. The distance between an event and a facet was

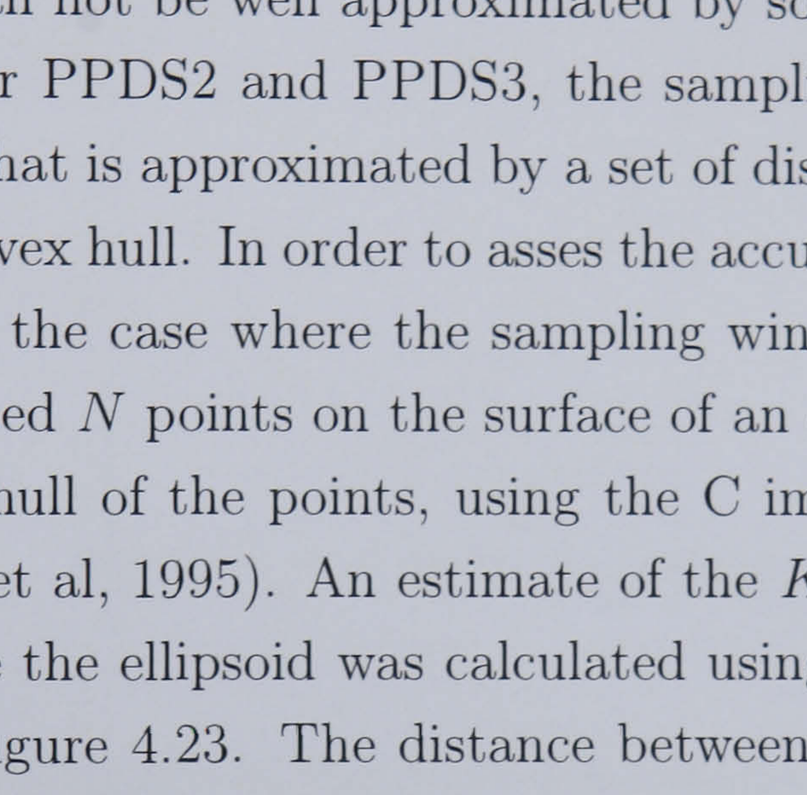


Figure 4.23: Plot of an estimate of the $K(r)$-function (estimated using the standard border method) for 100 events uniformly placed in an ellipsoid (approximated by a convex hull) with semi-axis lengths 5, 4 and 2 and theoretical curve (red)

calculated by taking the distance between the average of the points on the facet and the event. Although the curve is (as expected) strongly biased downwards, this method is more accurate than using the weighting factor edge-correction with a non-time consuming Monte Carlo approximation for the weighting factor. In particular,

we estimated the $K(r)$-function using the weighting factor edge-correction method and adopted Monte Carlo integration for calculating $\omega(.)$. This involved simulating uniformly inside the convex hull. The results using a Monte Carlo integration that did not require extensive computational times were rather disappointing (see Figure 4.24).

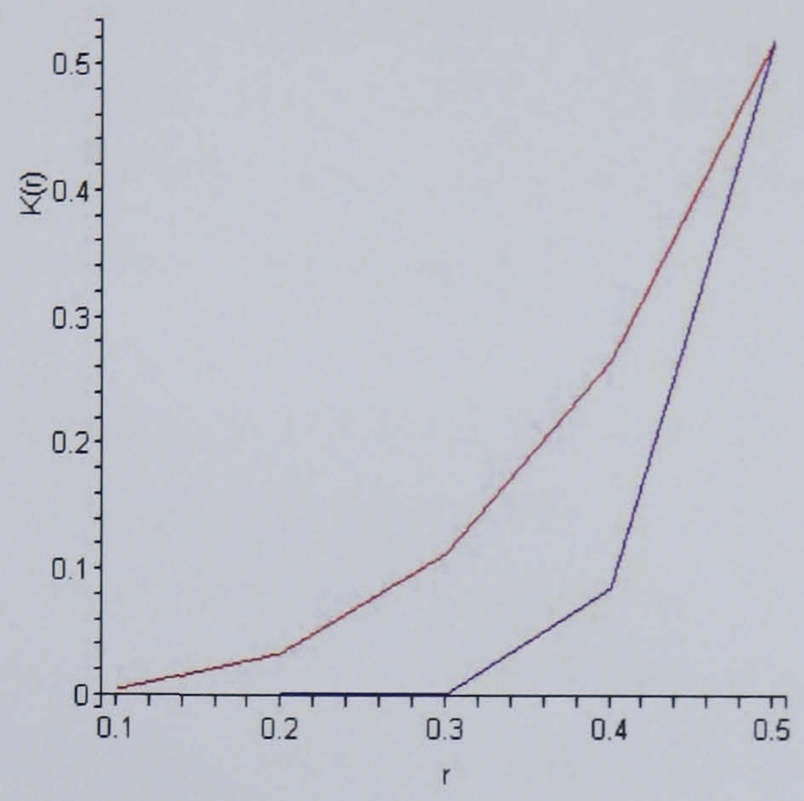


Figure 4.24: Plot of an estimate of the $K(r)$-function ($\times 324$), using equation (4.23) with $\omega(.)$ calculated using Monte Carlo integration (blue curve) for 100 events uniformly placed in an ellipsoid with semi-axis lengths 5, 4 and 2, and theoretical curve (red curve)

4.5 Occurrences where edge-correcting may not be advantageous

We have so far provided a discussion that indicates the great importance of edge-corrections. However, there are cases where edge effects can be safely ignored. Examples of such situations include hypothesis testing (Baddeley et al, 1993) and are further outlined by Diggle (2003). Indeed in the majority of the simulation studies that we describe throughout, where we compare a characteristic that has been estimated for the PML NB data, with those estimated from simulated data, we do not carry out an edge-correction, since we are comparing empirical estimates of the characteristics (that have been estimated in the same way for the PML NB data and simulated data) as apposed to comparing an empirical estimate with a theoretical form. A consequence of this is, any difference in bias, arising when estimating a characteristic for PML NB data compared to the same computation for simulated

data, can have a positive impact on the analysis (in the sense of improving our understanding of the differences between the PML NB data and simulated data). We are reluctant to carry-out an edge-correction when not required, as edge-correcting is generally computationally expensive. Further to this, there are situations where it may be advantageous to not carry out an edge-correction. Diggle (1979) stated that edge-correction such as applied in the $K(r)$-function estimate can substantially reduce the sampling fluctuations of a statistic and thus improve the power of a test based on it. However, Gignoux et al (1999) showed that for small samples (about 10-30 events), the power of CSR tests, such as those we have adopted throughout, is improved if an edge-correction is not applied.

We believe that we have obtained a convincing amount of evidence to suggest that there is significant level of departure from CSR, for the spatial distribution of PML NB centroid locations within some of the cell nuclei. The next chapter provides a detailed account of attempting to fit an alternative model to the PML NB data. Some of the analysis in the next chapter require the use of edge-corrections. This chapter has raised our awareness of various edge-correction techniques and has addressed the importance of edge correcting (when it is required), in particular for the three-dimensional case.

Chapter 5

Modelling replicated PML confocal microscopy data

In Chapter 1 we presented the results of CSR tests on PPDS1 and found that in all but one cell there is some evidence to reject the null hypothesis of CSR for the spatial distribution of PML NBs. The outcome of these CSR tests were backed by further CSR tests in Chapter 3, that utilised the $Q(r,\tilde{r})$-function, and plots presented in Chapter 4, that showed the differences between edge-corrected estimates of $K(r)$-functions and the theoretical $K(r)$-function of a homogeneous Poisson process. In this chapter we take the next natural step in point pattern analysis, namely to find an appropriate alternative model to CSR. We focus our attention on PPDS2 as it contains valuable information on the nuclear boundary, while PPDS1 only contains details of PML NB centroid locations within the nucleus (see Appendix B). Besides being important for edge-correction when estimating distance-based summary statistics (see Chapter 4), knowledge of the nuclear boundary can be useful for implementing models that support recent biological theory, such as some of those outlined by McManus et al (2006). More precisely, McManus et al (2006) have reported that chromosomes and regions of chromosomes segregate differently within the nucleus depending on whether or not they are rich in potentially transcribed genes. The individual interphase chromosome territories segregate their gene rich R-bands into the interior of the nucleoplasm, whereas their gene poor G-bands are gathered against the periphery of the nucleus and against the nucleolar surface (see for example Shopland et al, 2003). Euchromatin sequences are further organised such that they maintain a spatial relationship with the predominant nucleoplasmic nonchromatin structure, the splicing factor compartments (McManus et al, 2006). Smaller nonchromatin structures such as PML associate with specific regions of the

genome. In summary, this means that there is biological reason to suggest that the spatial location of PML NBs is related to the nuclear boundary.

Up until now nucleoli have been believed by some biologists to act as exclusion regions for certain other compartments (Shiels et al, 2007). In particular, PML NBs do not exist within regions of nucleoli (Shiels, 2005). However, this idea has been challenged following very recent unpublished work by Imperial College London centre for structural biology. We apply the idea of nucleoli as PML exclusion regions to help us construct a model for the spatial distribution of the PML NBs within the nucleus.

Throughout this chapter a dataset is generally assumed to be a collection of k iid replicates. This means that we have k cell nuclei, where it is assumed that the spatial distribution of the PML NBs in each of the nuclei is governed by the same spatial point process.

We begin this chapter with a test for assessing whether the assumption of iid replicates is valid for this carefully selected dataset and find that there is not strong evidence to reject this assumption. Following this, we go on to fit our first model, Model 1, a stationary Poisson process defined in the nucleus interior excluding the regions occupied by nucleoli. After some rigorous diagnostic tests that include an interpretation of $Q(r, \tilde{r})$-function plots, we find that although this first model is consistent with some of the biological ideas and some of the results of CSR simulation studies from Chapters 1 and 3, it is inappropriate for describing the spatial distribution of PML NBs in some of these cell nuclei. Model 1 is subsequently used as a 'straw man' comparison for more complicated models. We go on to develop this simplistic model into a more sophisticated one, Model 2, that incorporates PML-boundary interaction.

We initially define Model 2 through simple postulates, and go onto prove that these postulates characterise an inhomogeneous Poisson process. As a consequence, we are able to make use of the established theory on inhomogeneous Poisson processes. Specifically, as also is the case for Model 1, we are in a position to fit Model 2 using a maximum likelihood estimation (MLE) approach and are able to simulate from the model using the standard approaches. We move on to consider a final complication, whereby we incorporate interaction between the PML bodies. Through simulation studies, we explore the possibility of fitting a Markov model, Model 3, to PPDS2 by

simulating replicates via perfect simulation from Model 3 and studying how well we are able to estimate the parameters of Model 3 by utilising a Metropolis-Hastings algorithm with the ratio of partition functions in the Hastings ratio approximated by thermodynamic integration.

Models 1 - 3 are constructed in such a way that Model 2 is a special case of Model 3 and Model 1 is a special case of Model 2. Following some observations made whilst fitting Model 2, we attempt to fit a log-Gaussian Cox process to PPDS2. We end this chapter with a discussion on selecting between competing models. using techniques that make use of partial Bayes factors and pattern recognition schemes.

5.1 Testing for similarity

An initial step taken before attempting to fit a model to replicated PML data involves testing the assumption that the PML NBs in the cell nuclei that makes up PPDS2 come from the same spatial point process distribution. To do this, we use the method described by Diggle (2003) and Webster et al (2006) for group comparisons for testing within group spatial data similarity. The procedure makes use of a non-stationary version of the $K(r)$-function. Since we are working with three-dimensional data, we extend the definition of the (two-dimensional) inhomogeneous $K_{\text{inhom}}(r)$- function as given by Baddeley et al (2000(b)), to the three-dimensional case as follows.

Definition 3 *Let X be a second-order intensity-reweighted stationary point process. Define the inhomogeneous $K(r)$-function of X, $K_{inhom}(r)$ by*

$$K_{inhom}(r) = \frac{1}{\nu(B)} E\left(\sum_{\eta \in X \cap B} \sum_{\zeta \in X \setminus \{\eta\}} \frac{I_{(\|\eta-\zeta\| \leq r)}}{\lambda(\eta)\lambda(\zeta)} \right), \quad r \geq 0$$

for any $B \in \mathcal{B}_0$ with $\nu(B) > 0$ and $\mathcal{B}_0$ the class of bounded Borel sets in $\mathbb{R}^3$. We take $u/0 = 0 \; \forall \; u \geq 0$.

Furthermore, a point process X is said to be *second-order intensity-reweighted stationary* if

$$F(W, V) = F(W + \eta, V + \eta)$$

for all $\eta \in \mathbb{R}^3$, and $W, V \in \mathcal{B}_0$. Where

$$F(W, V) = E\left(\sum_{\eta \in X \cap W} \sum_{\zeta \in X \cap V} \frac{1}{\lambda(\eta)\lambda(\zeta)} \right).$$

We have a collection of k replicates. The test for similarity between the replicates can be outlined as follows. Let $\hat{K}_{\text{inhom}_i}(r)$ denote the $K_{\text{inhom}}(r)$-function estimate of the ith replicate, Γ_i in the set $\mathcal{D} = \{\Gamma_1, \ldots \Gamma_k\}$ of replicates. We estimate a weighted average $K_{\text{inhom}}(r)$ for the set

$$\bar{K}_{\text{inhom}}(t) = N^{-1} \sum_{i=1}^{k} n_i \widehat{K}_{\text{inhom}_{i(t)}},$$

where n_i is the total number of events in replicate Γ_i and

$$N = \sum_{i=1}^{k} n_i.$$

We use the statistic

$$D = \sum_{i=1}^{k} \int_0^{t_0} w(t) n_i [\widehat{K}_{\text{inhom}_i}(t) - \bar{K}_{\text{inhom}}(t)]^2 dt \tag{5.1}$$

to test for significant differences within the group. The function $w(t) = t^{-2}$ is used to stabilise the variance. Following Webster et al (2006), due to the intractability of the distribution of D, we use a bootstrap based significance test for the null hypothesis of no difference between the replicates. Specifically, this involves computing a set of residual $K_{\text{inhom}}(r)$-functions,

$$R_i(t) = \sqrt{n_i}(\widehat{K}_{\text{inhom}_i}(t) - \bar{K}_{\text{inhom}}(t)).$$

We then randomly resample, with replacement from the set $\{R_1(t) \ldots R_k(t)\}$ of residual $K_{\text{inhom}}(r)$-functions to obtain bootstrap estimates of residual $K_{\text{inhom}}(r)$, $R_i^*(t)$ and calculate

$$K^*_{\text{inhom}_i}(t) = \bar{K}_{\text{inhom}}(t) + \frac{1}{\sqrt{n_i}} R_i^*(t).$$

These bootstrapped residual $K_{\text{inhom}}(r)$-functions are used to form a realisation D^*. By ranking the estimated value of D against the bootstrapped, we attain a significance level.

Applying this procedure to 5 replicates (to increase the power of the test, we use the cell nuclei containing the greatest number of PML NBs), we attained a p-value of **0.18**. This suggests insignificant reason to reject the null hypothesis of similar $K_{\text{inhom}}(r)$-functions between replicates. Hence we assume that the spatial distribution of PML NBs within the cell nuclei that form PPDS2 is governed by the same

spatial point process. We note that these results are consistent with the cell nuclei being similar in a biological sense. Specifically, the cell nuclei that make up PPDS2 are from cells that are in a G_0, or "resting phase" of the cell cycle (see for example Naniche et al, 1999). The importance of analysing nuclei from the same stage of the cell cycle stems from the biological idea that nuclei undergo extensive changes in organisation as they progress through the cell cycle and development (Thakar & Csink, 2005).

To reduce computational burden we replace the integral in (5.1) with the discrete summation

$$\sum_{t=1}^{y} w(\delta t) n_i [\widehat{K}_{\text{inhom}_i}(\delta t) - \bar{K}_{\text{inhom}}(\delta t)]^2$$

where $\delta = t_0/y$. Further to this, we apply a border estimate (see Chapter 4) for $K_{\text{inhom}(r)}$ given by

$$\widehat{K}_{\text{inhom}}(r) = \frac{1}{\nu(W)} \sum_{\eta \in X \cap W \ominus r} \sum_{\zeta \in X \setminus \{\eta\}} \frac{I_{(\|\eta - \zeta\| \leq r)}}{\hat{\lambda}(\eta)\hat{\lambda}(\zeta)}$$

with an edge-corrected kernel estimator, $\hat{\lambda}(\eta)$ for $\lambda(\eta)$ (see Chapter 2, section 2.7), that utilises the multivariate Gaussian kernel

$$\kappa(\xi) = (2\pi)^{-3/2} \exp(-\xi^T \xi/2).$$

We adopt the quadrature scheme

$$\int_W h^{-3} \kappa\left(h^{-1}(\eta - \xi)\right) \nu(d\xi) \approx \sum_{i=1}^{U} h^{-3} \kappa\left(h^{-1}(\eta - \xi_i)\right) \Delta V \tag{5.2}$$

for calculating the edge-correction (2.8), where $\Delta V = \nu(W)/U$ and $\{\xi_1, \ldots \xi_U\}$ are point locations inside the convex hull (see Appendix B and Chapter 1 for details on constructing the datasets). In the procedure outlined above, we used $U = 10,000$. Integral approximations similar to that given in (5.2) are used throughout the rest of this document.

An appropriate bandwidth, h, in (5.2) can be obtained by minimising the mean integrated squared error (Cressie, 1991)

$$MISE(\hat{\lambda}) = E\left(\int_W (\widehat{\lambda}(\eta) - \lambda(\eta))^2 \nu(d\eta)\right) \tag{5.3}$$

A pitfall with such an approach is that $\lambda(\eta)$ in (5.3) is usually unknown. Guan (2007) recently proposed a more complicated data-driven approach for bandwidth selection that is based on a cross-validation idea from density estimation. The approach that we adopt is to obtain an estimate of an appropriate bandwidth by replacing $\lambda(\eta)$ in (5.3) by a constant intensity λ_n. We then minimise

$$\frac{1}{n}(\Delta V \sum_{j=1}^{U}(\widehat{\lambda}(\eta_j) - \lambda_n)^2) \tag{5.4}$$

by using a grid search for h.

Carrying out the minimisation of (5.4), for example, for cell 1 of PPDS2 yielded a bandwidth of 5.99. A rule-of-thumb method by Møller and Waagepetersen (2003) uses the estimate $h \approx c\lambda^{-\frac{1}{2}}$, with c some constant between 0.1 and 0.2. Such an approximation provides a bandwidth value between $32.67 - 65.33$. However, the approximation by Møller and Waagepetersen (2003) is based on simulations and practical experience for point patterns that are in two-dimensions and consist of 50-300 events. Both the bandwidth approximation by Møller and Waagepetersen (2003) and the bandwidth selection procedure that requires minimisation of (5.4) with respect to h, necessitate that the intensity does not vary greatly within the cell nuclei. This is true in particular for cells containing PML bodies that do not exhibit significant departure from CSR.

The aim in this chapter is to fit an appropriate spatial point process model to replicated PML data. We have begun with a test for whether the cell nuclei that make up the replicated data (PPDS2) contain PML bodies that have similar spatial distributions. The procedure that we have adopted is a bootstrapping-dependent one that makes use of an inhomogeneous version of the $K(r)$-function. As a result, we have formally extended the definition of the $K(r)$-function, as provided by Baddeley et al (2000(b)) to the three-dimensional case. The outcome of the test for similarity suggested that we can assume that the PML NBs have similar distributions within the cell nuclei that make up PPDS2. We will now move onto fit an initial model to PPDS2.

5.2 Model 1: Stationary Poisson process with exclusion zones

Consider a simple model for the spatial locations of PML NBs, in which the PML bodies are uniformly distributed in the region of the cell nucleus not occupied by nucleoli. That is, we define the model by the following postulates:

- **M1P1** The number of PML NBs n, is a realisation of a Poisson process with intensity measure $\lambda\nu(W)$ $(0 < \lambda < \infty)$.

- **M1P2** Conditional on n, the PML NB locations follow a uniform distribution on W^*

where W^* is known and is the region of the nucleus interior unoccupied by nucleoli. In **M1P1** we base the number of PML bodies on the total volume of the nucleus so that we do not have to estimate $\nu(W^*)$ (the volume of the nucleus interior unoccupied by nucleoli).

5.2.1 Model 1 characteristics

The theoretical model characteristics are similar to those of a homogeneous Poisson process. Some of these are detailed in Chapter 2. We have that, under this model

$$E(X(W)) = \lambda\nu(W)$$

Furthermore

$$\mathrm{Var}(X(W)) = \lambda\nu(W).$$

Also for this model, we have that the $K(r)$-function, for the three-dimensional case, in the absence of nucleoli (or in the case where nucleoli exhibit CSR and can be modeled as points) is given by

$$K(r) \approx \frac{4}{3}\pi r^3$$

5.2.2 Fitting Model 1

The maximum likelihood estimate for the homogeneous Poisson process intensity parameter (which for Model 1 is equivalent to the parameter λ) is well-known and can be found for example in Stoyan & Stoyan (1994, section 13.2). Let $\widehat{\lambda}_r$ be the

MLE of the intensity parameter λ_r for replicate r observed in window W_r. Then the MLE estimator, $\widehat{\lambda}_{\text{rep}}$ for the intensity parameter for k iid replicates is given by

$$\widehat{\lambda}_{\text{rep}} = \frac{\sum_{r=1}^{k} n_r}{\sum_{r=1}^{k} \nu(W_r)} \tag{5.5}$$

For PPDS2 we have $k = 8$ replicates. Using (5.5), we obtain $\widehat{\lambda}_{\text{rep}} = 9.24 \times 10^{-6}$ for PPDS2. The variance of the estimator (5.5) is given by

$$\text{Var}_{\widehat{\lambda}_{\text{rep}}} = \frac{\lambda_{\text{rep}}}{\sum_{s=1}^{k} \nu(W_k)} \tag{5.6}$$

and an unbiased estimator for (5.6) is

$$\widehat{\text{Var}}_{\widehat{\lambda}_{\text{rep}}} = \frac{\widehat{\lambda}_{\text{rep}}}{\sum_{s=1}^{k} \nu(W_k)}. \tag{5.7}$$

Using (5.7), we estimate the standard error $\left(\sqrt{\widehat{\text{Var}}_{\widehat{\lambda}_{\text{rep}}}}\right)$ of the estimated $\widehat{\lambda}_{\text{rep}}$ for PPDS2 to being 1.02×10^{-12}.

5.2.3 Model 1 diagnostics

Under Model 1, stationarity inside W^* is implicitly assumed. We can therefore calculate some of the summary statistics defined in Chapter 2 and the $Q(r, \tilde{r})$-function defined in Chapter 3. Plots of the $Q(r, \tilde{r})$-functions for the spatial locations of PML NBs in the cell nuclei of PPDS2 compared to simulation envelopes produced by 100 simulated realisations from Model 1 are shown in Figures 5.1 and 5.2. The $Q(r, \tilde{r})$-function plots suggests that there are some values of $(r, \tilde{r})$ where the model does not fit the data quite as well as for other values. For example, for cells 1, 6, 7 and 8, for values of r and $\tilde{r} \geq 135$ units (see Appendix B for units interpretation), a high value of $Q(r, \tilde{r})$ is obtained for the PML NB points, compared to Model 1. The $Q(r, \tilde{r})$-function plots for cells 2, 3, 4, and 5 do not provide evidence against the appropriateness of Model 1.

Diggle (2003) states that the use of EDF plots of the data with simulation envelopes, as used for the CSR tests in Chapter 1, is strictly invalid as a formal approach for assessing model fit. Loosmore & Ford (2006) state that using simulation envelopes based on summary statistics such as the $K(r)$- and $G(r)$-functions is invalid for testing an observed point pattern against a specified model because it violates the assumptions of Monte Carlo methods and the type I error is greatly underestimated.

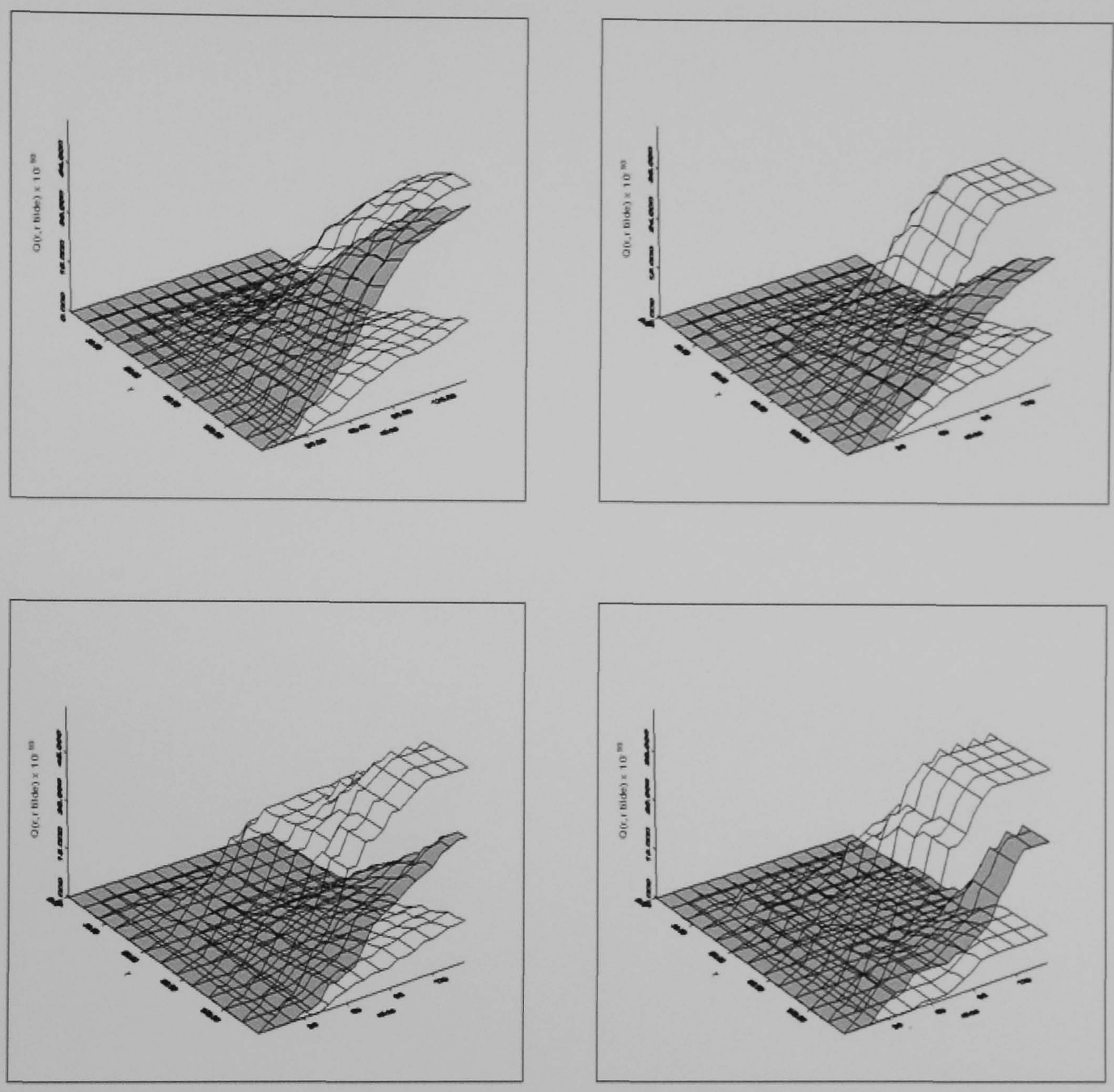


Figure 5.1: $Q(r, \tilde{r})$ function ($\times 10^{-10}$) of cells 1-4 of PPDS2 (grey surfaces) with simulation envelopes (transparent surfaces)

We adopt the following procedure, similar to that proposed by Diggle (2003, p89) to more rigorously assess the appropriateness of Model 1 for describing the spatial distribution of the PML bodies. We simulate m times from Model 1 and calculate the quantities

$$g_i = \int_0^T \left(\widehat{G}_i(t) - \overline{G}_i(t)\right)^2 dt \tag{5.8}$$

and

$$f_i = \int_0^T \left(\widehat{F}_i(t) - \overline{F}_i(t)\right)^2 dt \tag{5.9}$$

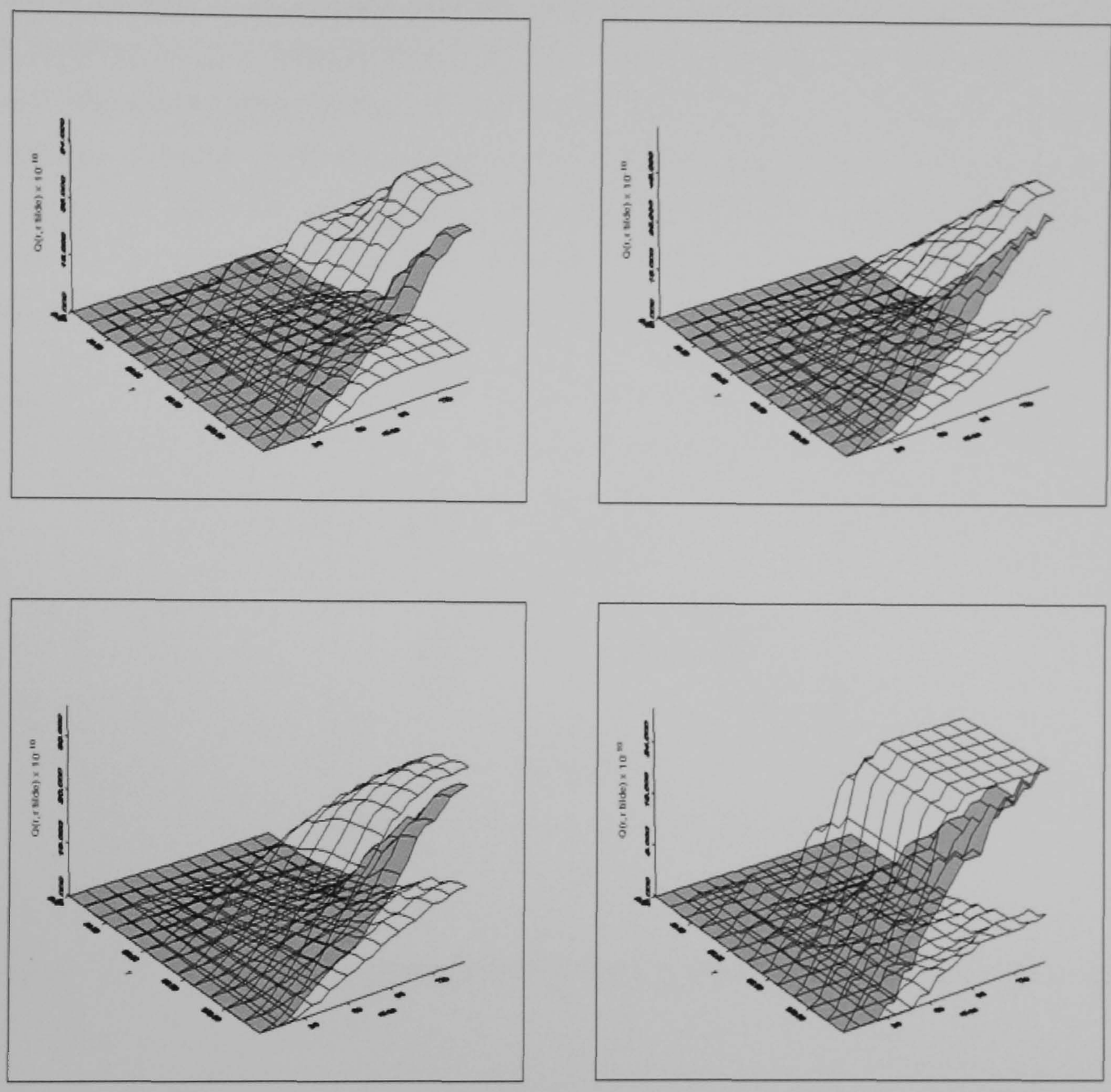


Figure 5.2: $Q(r, \tilde{r})$ function ($\times 10^{-10}$) of cells 5-8 of PPDS2 (grey surfaces) with simulation envelopes (transparent surfaces)

where

$$\overline{G}_i(t) = \frac{\sum_{j \neq i} \widehat{G}_j(r)}{m-1} \tag{5.10}$$

for $i = 2, \ldots, m$ are the EDFs of the nearest neighbour distances from simulations from Model 1 and $\widehat{G}_1$ is the EDF of the nearest neighbour distances for the data. Similarly for the empty space function $\widehat{F}_i(t)$ and $\overline{F}_i$ (see Chapter 2). We replace the integrals in (5.8) and (5.9) with summations so that

$$g_i = \sum_{t=1}^{S} \left(\widehat{G}_i(\delta T t) - \overline{G}_i(t \delta T) \right)^2 \tag{5.11}$$

for $\delta T = T/S > 0$ for some $S \in \mathbb{Z}^+$ and similarly for f_i. If g_1 is ranked the rth smallest of $\{g_1, \ldots g_m\}$, then the attained level of significance, p, for a two-tailed test is given by $p = \sup\{r/2(m+1), (m-r)/2(m+1)\}$, and similarly for f_1. We repeat the entire experiment M times and only accept the model if neither of the quantities (5.8) or (5.9) show signs of significant lack of fit. In these assessments we use $m = 100$, $M = 100$, $S = 100$ and $T = \text{Rad}(W_r)$, where $\text{Rad}(W_r)$ is the radius of the nucleus r. That is, $\text{Rad}(W) = \inf\{\|\eta - \xi\| : \eta, \xi \in \partial W_r\}$. Scatter plots of the attained significance levels are shown in Figures 5.3 - 5.5. For cells

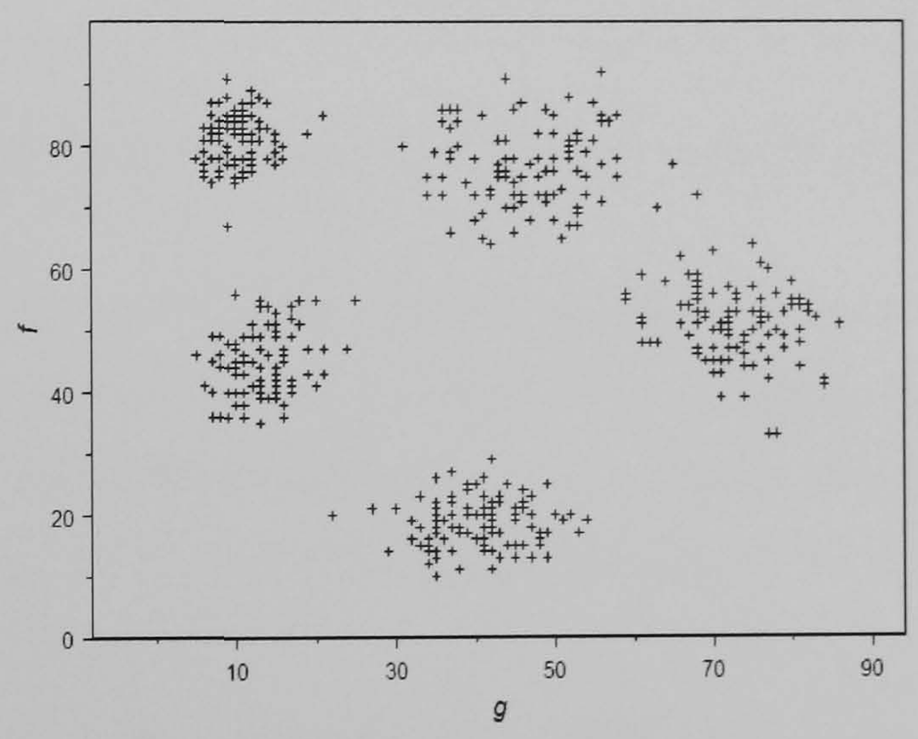


Figure 5.3: Cells 1 - 5 superimposed f_i and g_i function significance test results

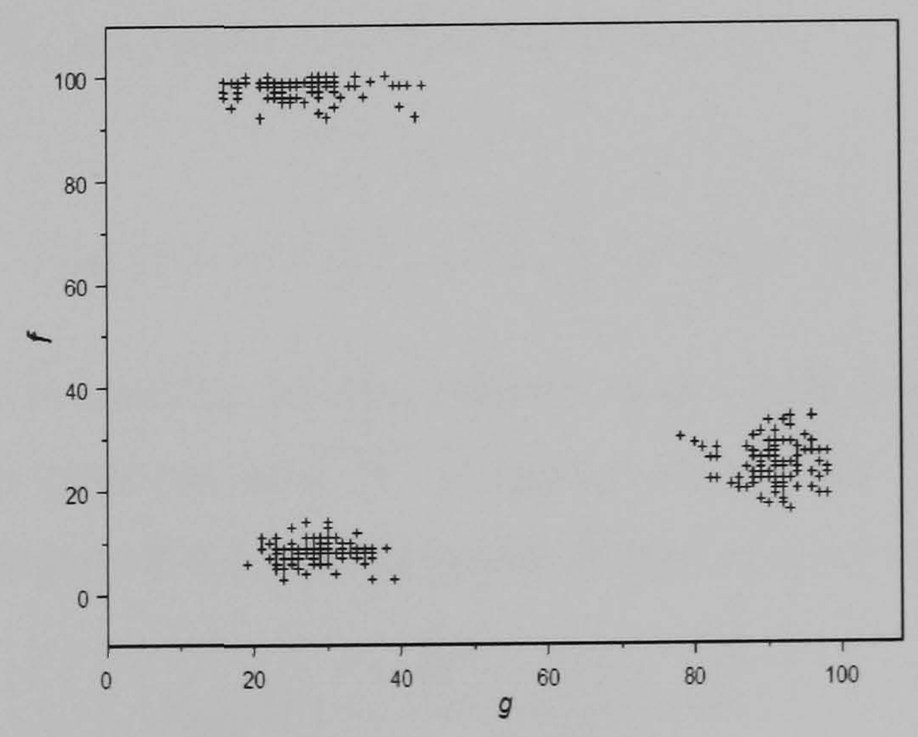


Figure 5.4: Cells 6 - 8 superimposed f_i and g_i function significance test results

6-8 there was evidence to reject Model 1 at the 5% level of a two-tailed test. The strongest evidence for rejecting the model came from cell 8 (see Figure 5.5). The

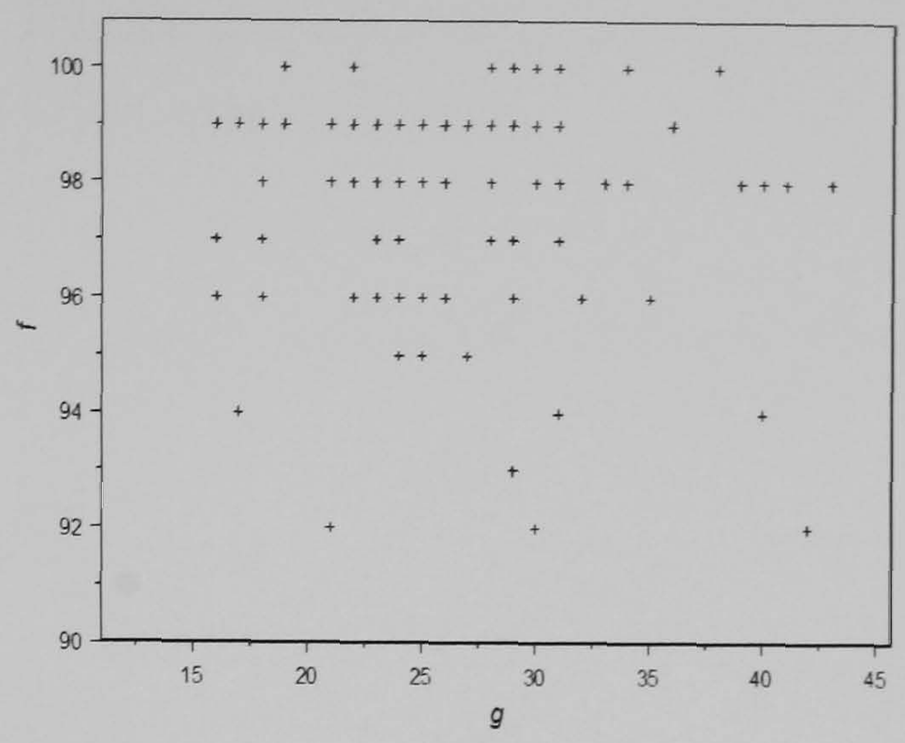


Figure 5.5: Cell 8 f_i and g_i function significance test results

results for cells 1-5 showed that for these cells there was not significant evidence to reject Model 1 at the 5% level of a two tailed test. The overall correlation of the f_i and g_i quantities for all of the cells was -0.31. From the results of the diagnostics tests, Model 1 appears to be a good initial candidate model for the PML data.

5.3 Model 2: Boundary effects model

As outlined at the beginning of this chapter, biologists speculate that there may exist a relationship between the spatial locations of PML and the location of the nuclear boundary. To gain quantitative insight into this idea we introduce a summary statistic $F_{\partial W}(r)$ which we define for spatial point process X in bounded $W \subset \mathbb{R}^d$, by

$$F_{\partial W}(r) = P(\|\eta - \partial W\| \leq r | \eta \in X). \tag{5.12}$$

The function $F_{\partial W}(r)$ is similar to the empty space function (see Chapter 2). For a homogeneous Poisson process and W a sphere of radius R it is straightforward to show from the properties of a homogeneous Poisson process that

$$F_{\partial W}(r) = \frac{R^3 - (R-r)^3}{R^3}.$$

Recall that cells 6-8 provided the strongest evidence to reject Model 1. To see if this is a result of boundary effects, we estimated (5.12) for cells 6-8 using the estimate

$$\widehat{F}_{\partial W}(r) = \frac{\sum_{\eta \in X \cap W} I_{(\|\eta - \partial W\| \leq r)}}{X(W)} \tag{5.13}$$

for W the cell nucleus, obtained as described in Chapter 1 and Appendix B. We subsequently formed simulation envelopes of (5.13) from Model 1, conditional on the number of events (see Figures 5.6-5.8).

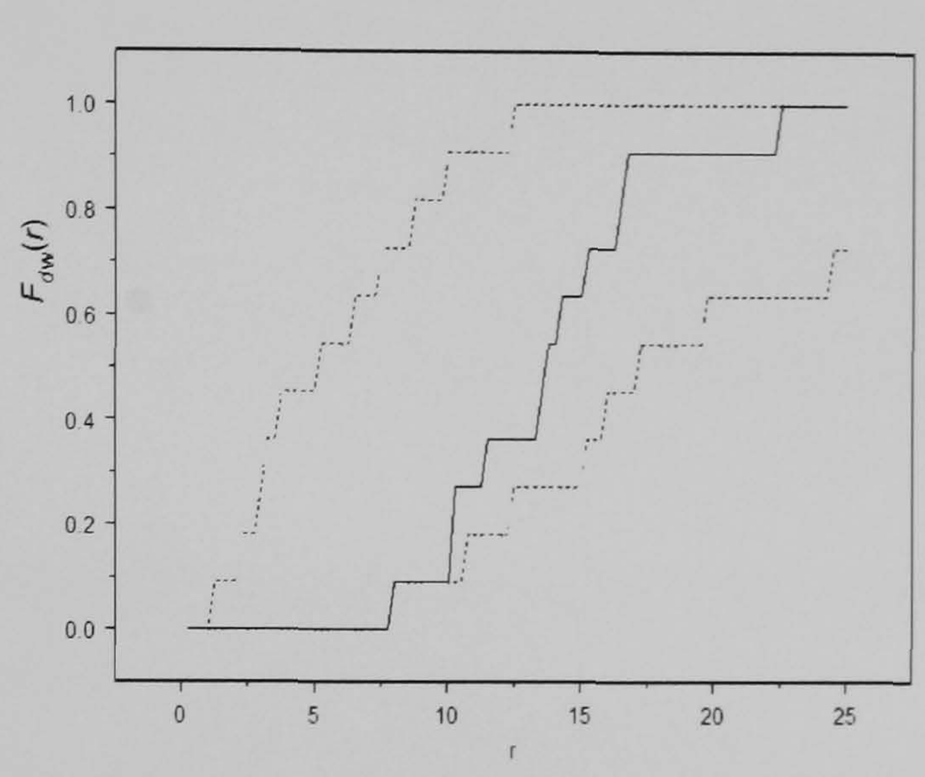


Figure 5.6: Cell 6 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1

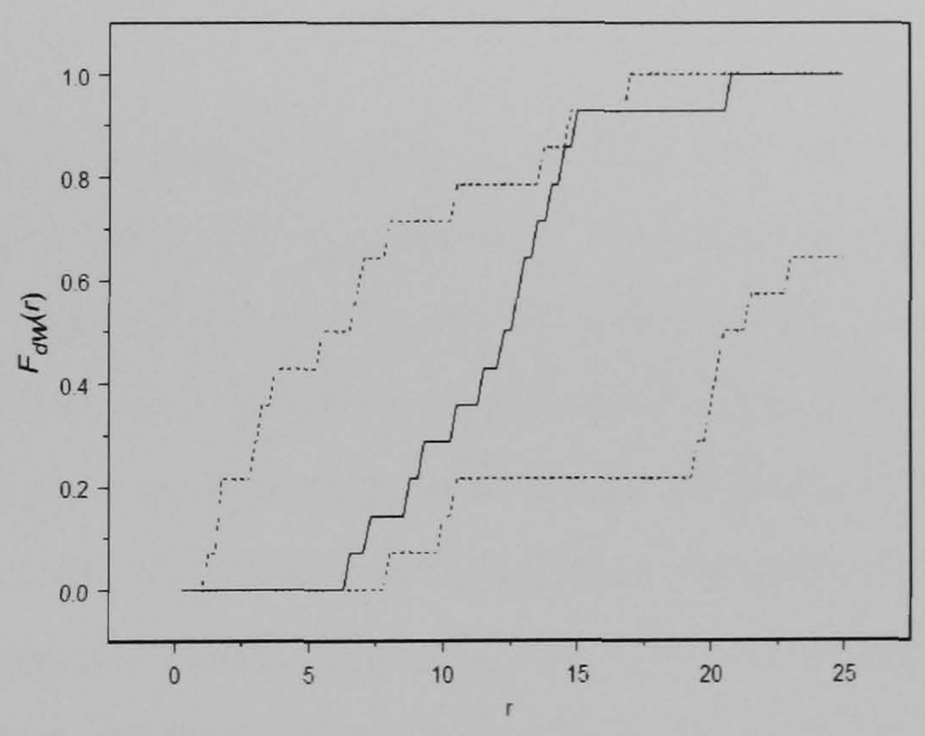


Figure 5.7: Cell 7 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1

Figures 5.6-5.8 provide some interesting insight into the spatial distribution of the PML NBs relative to the nuclear boundary. For cells 6-8 there is a lower probability of finding a PML body within a distance of 10 units of the boundary compared to the PML NBs being uniformly placed in the nuclear interior excluding that occupied

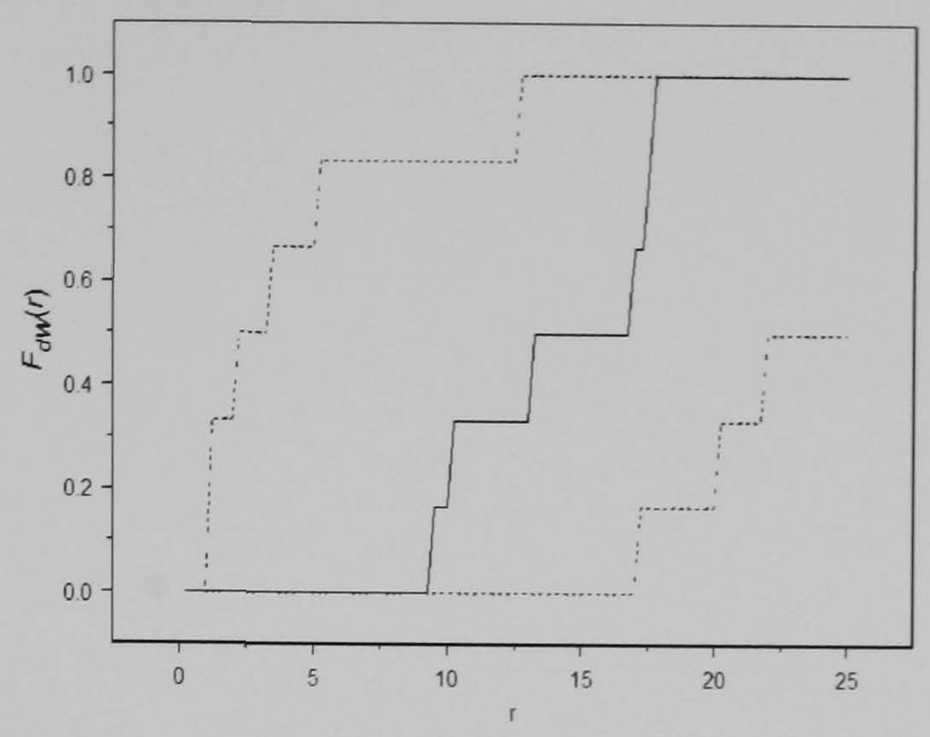


Figure 5.8: Cell 8 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1

by nucleoli (Model 1). Furthermore, we would generally expect to see more PML NBs within a distance of about 25 units of the boundary compared to that suggested by Model 1. These results provide some evidence for a non-uniform environment within the region excluding the nucleoli. Apart from cell 2, which shows greater signs of a CSR distribution, the result is not too dissimilar for the other cells of PPDS2 (see Figure 5.9).

We use these findings to take Model 1 a step further. Specifically, we now incorporate the idea of PML NB interaction with the nuclear boundary into Model 1 and as a result fit an inhomogeneous Poisson process to PPDS2. The inhomogeneous Poisson process is a flexible model that has been applied in diverse fields such as ecology, geology, seismology, and epidemiology (Rathbun & Cressie, 1994). While much application of the inhomogeneous Poisson process has been to fitting data in the time domain, there remains very few examples where inhomogeneous Poisson processes have been used to model data in $\mathbb{R}^d$ for $d \geq 3$ (Cressie, 1991). We now drop the notation W^* and always assume that the region of interest inside the nucleus excludes that occupied by nucleoli. We define Model 2 by the following postulates.

- M2P1 The event (PML body) locations are a realisation of a homogeneous Poisson process with intensity λ inside bounded $W \subset \mathbb{R}^3$.
- M2P2 Each event η is retained with probability

$$p(\eta) = 1 - \exp(-\kappa \|\eta - \partial W\|) \quad \kappa \in \mathbb{R}^+$$

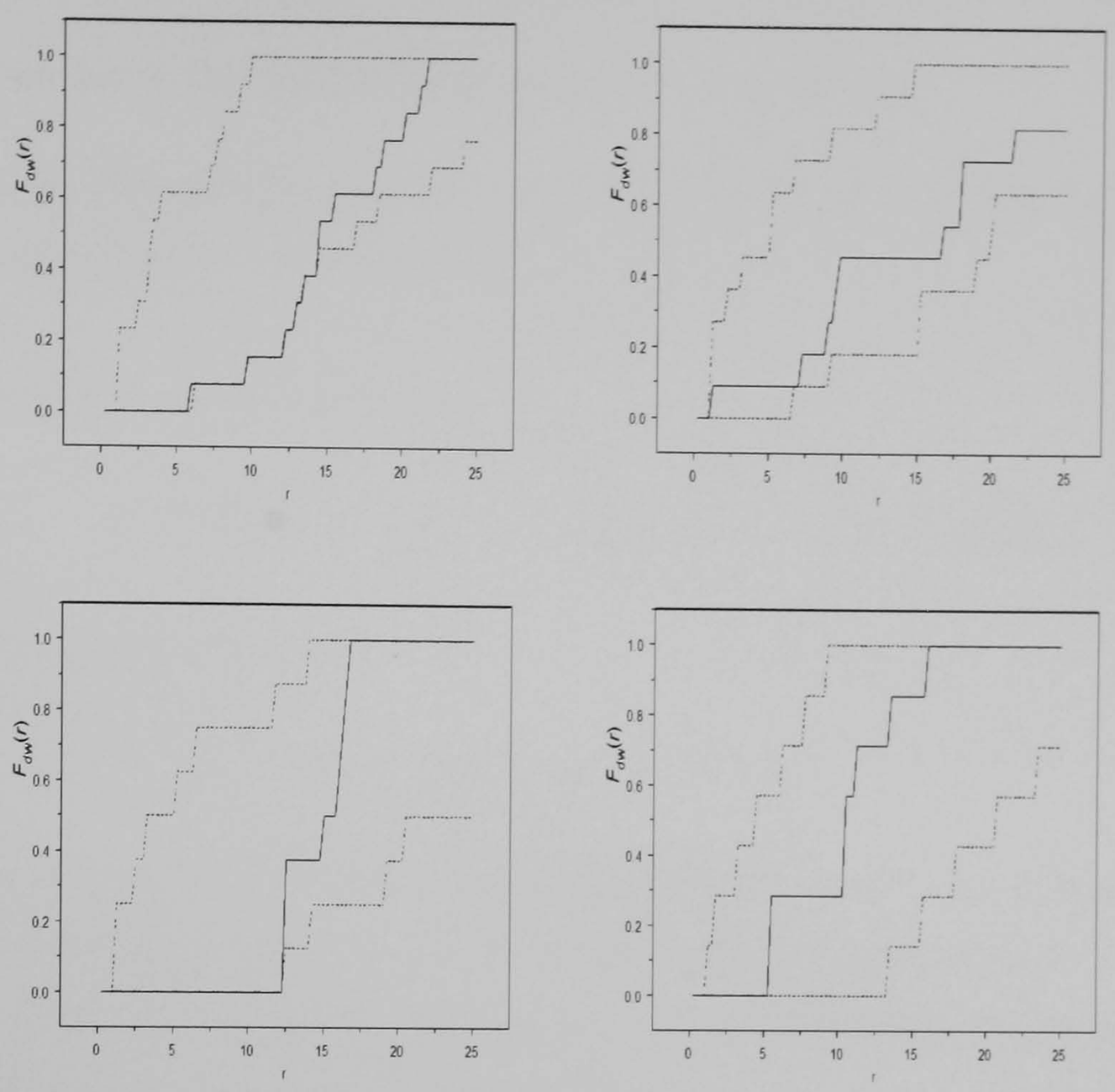


Figure 5.9: Cells 1-5 of PPDS2 estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1. Cell 1 is the top left and cell 2 the top right.

otherwise independently thinned with probability

$$1 - p(\eta).$$

The specifications of Model 2 implies that as $\|\eta - \partial W\| \to 0$, $p(\eta) \to 0$ which means that PML bodies are less likely to be observed close to the boundary. Evidence for this phenomenon was presented at the beginning of this section. Furthermore, as $\kappa \to 0$ Model 2 simplifies to Model 1.

5.3.1 Simulation and theoretical properties of Model 2

For Model 2 we have the following important proposition:

Proposition 6 *The spatial point process X, defined by M2P1-M2P2 is an inhomogeneous Poisson process. Furthermore,*

i) X has intensity measure given by

$$E(X(B)) = \int_B (1 - exp(-\kappa \|\eta - \partial B\|))\lambda\nu(d\eta) \tag{5.14}$$

ii) X has intensity function given by

$$\lambda(1 - exp(-\kappa\|\eta - \partial W\|)) \tag{5.15}$$

Proof. i) follows from Prekopa's theorem which is stated by Stoyan et al (1995). That is, the resulting spatial point process from an inhomogeneous Poisson with intensity measure Λ_b that has undergone a $p(\eta)$-thinning, is again Poisson with intensity measure

$$E(X(B)) = \int_B p(\eta)\Lambda_b(d\eta) \tag{5.16}$$

ii) Follows directly from the definition of the intensity function and intensity measure. ■

For some B we can calculate the expectation (5.14) explicitly.

Proposition 7 *For B a sphere of radius R, under Model 2 we have the following.*

$$E(X(B)) = \frac{4\lambda\pi}{3\kappa^3}(\kappa^3R^3 + 6exp(-\kappa R) + 6\kappa R - 3\kappa^2R^2 - 6)$$

Proof. We have

$$\begin{aligned} E(X(B)) &= \lambda \int_B (1 - \exp(-\kappa\|\eta - \partial W\|))\nu(d\eta) \\ &= 2\lambda \int_0^{2\pi} \int_0^{\frac{\pi}{2}} \int_0^R (1 - \exp(-\kappa(R - r)))r^2\sin(\phi)drd\phi d\theta \\ &= \frac{4\lambda\pi}{3\kappa^3}(\kappa^3R^3 + 6\exp(-\kappa R) + 6\kappa R - 3\kappa^2R^2 - 6) \end{aligned}$$

■

We have shown that Model 2 is an inhomogeneous Poisson process. We are able to

make use of Model 2 being an inhomogeneous Poisson process in the case where we need to fit data to the model (this follows in the next section), or we need to simulate realisations from the model. Specifically, we can adopt the rejection method due to Lewis and Shedler (1979), more precisely described as follows

- Simulate a homogeneous Poisson process with intensity $\lambda(1 - \exp(-\kappa\|\partial W\|))$
- Retain each point η with probability $\frac{1-\exp(-\kappa|\eta-\partial W\|)}{1-\exp(-\kappa\|\partial W\|)}$.

Where $\|\partial W\|$ is the maximum distance of any point in W to the boundary of W. For W a sphere of radius R, $\|\partial W\| = R$. For irregular windows, such as those provided in PPDS2, $\|\partial W\|$ can be computationally expensive and difficult to calculate. We will adopt this rejection method for simulating Model 2 throughout this chapter.

5.3.2 Fitting Model 2: parameter estimation and diagnostics

A variety of methods can be used to fit Model 2 to data in the general case. Minimum contrast methods have been widely adopted in the spatial point process literature due to the difficulty of being able to write down closed form expressions for quantities such as the maximum likelihood (see for example Stoyan & Stoyan, 1994, section 15.9 for further discussions). Informally, the general concept behind the minimum contrast approach is to select parameters that minimise the 'difference' between a theoretical characteristic with that estimated from the data. As an example, suppose we wish to fit the parameter κ of Model 2 separately. A minimum contrast approach for fitting the parameter in symmetric W (for example in Chapter 1 we stated that some cell nuclei are spherical) would entail carrying out the relatively simple minimisation (with respect to κ) of the discretised, adjusted Mean Integrated Squared Error function, given by

$$\sum_{s=0}^{u} \left(\frac{\widehat{F}_Y(\delta s + r_{min})}{F_Y(\delta s + r_{min})} - 1 \right)^2,$$

Where $F_Y(r)$ denotes the probability that an event is within distance r of the boundary. Furthermore, $\widehat{F}_Y(r)$ is the empirical $F_Y(r)$ calculated from the point pattern and $\delta = \frac{r - r_{min}}{u}$. The function $F_Y(r)$ is of particular use here compared to other characteristics as it only depends on the model parameter κ, and we can obtain its closed form expression as follows:

$$F_Y(r) = P(\|\eta - \partial W\| \leq r) \tag{5.17}$$

$$= \frac{1}{\tau}\lambda \int_0^r (1 - \exp(-\kappa t))dt$$
$$= \frac{1}{\tau}\lambda \left(r + \frac{\exp(-\kappa r)}{\kappa} - \frac{1}{\kappa} \right) \tag{5.18}$$

where the normalising constant τ is equal to

$$\lambda \left(\|\partial W\| + \frac{\exp(-\kappa \|\partial W\|)}{\kappa} - \frac{1}{\kappa} \right).$$

This approach for determining κ separately is plausible as the location of events is determined purely by their distance to the boundary. A simulation study showed that this minimum contrast type approach for estimating κ typically yielded a biased estimate (see Table 5.1). We will see later that a combination of different model parameters and a more sophisticated model fitting approach when considering the PML NB data, greatly reduces the variance (relative to the parameter values) of the model parameter estimates (see Table 5.5 in the next section). The simulation study to obtain the results presented in Table 5.1 involved simulating 1000 times from Model 2, inside the unit sphere, using various values for κ and λ. The value of κ and λ were chosen in such a way that, a particular expected number of events would be obtained, when simulating from the model. The parameter κ was subsequently estimated from each of the simulated realisations using the minimum contrast approach.

Table 5.1: Simulation assessment for fitting κ to Model 2

λ	κ	Expected no. events	bias of $\hat{\kappa}$	variance of $\hat{\kappa}$
10	1.2	10	0.15	2.99
10	3.2	20	3.79	17.57
10	8	30	13.57	47.21
20	5	50	5.18	47.54
40	5.1	100	2.66	57.38

Fitting Model 2 to PPDS2 is non-trivial compared to fitting data in a regular window. We will now detail two approaches that we consider for fitting Model 2 to PPDS2. The first approach is minimum contrast and requires the minimisation of the function

$$\sum_{i=1}^{U} (\hat{\lambda}(\eta_i) - \lambda(1 - \exp(-\kappa \|\eta_i - \partial W\|)))^2$$

with respect to λ and κ. Furthermore, to take into account the k iid replicates we extend this to

$$\frac{1}{n}\sum_{r=1}^{k} n_r \sum_{i=1}^{U}(\hat{\lambda}(\eta_i) - \lambda(1 - \exp(-\kappa\|\eta_i - \partial W\|)))^2.$$

This approach can be very cumbersome for the sizes of U (the number of uniformly selected points that have been classified as empty space inside the nucleus, using the image data) that we consider. Recently, Waagepetersen (2007) fit an inhomogeneous Neyman-Scott process to a point pattern using minimum contrast estimation based on the $K_{\text{inhom}}(r)$-function and concluded that the minimum contrast approach has a potential loss of efficiency compared to the MLE approach. However, the technique is a very simple and fast alternative to MLE (Waagepetersen, 2007).

The problem of fitting Model 2 is eased since, as shown above, the postulates of Model 2 specify an inhomogeneous Poisson process. Using the standard approaches such as that provided by Cressie (1991, section 8.5), we can write down the form of the log-likelihood function, $L(X;\theta = (\lambda, \kappa))$ (shortened to $L(\theta)$) for Model 2 for a single replicate:

$$L(\theta) = \sum_{i=1}^{n} \log(\lambda(\eta_i))) - \int_W \lambda(\eta)d\eta \tag{5.19}$$

Furthermore, utilising the quadrature scheme

$$\int_W \lambda(\eta)d\eta \approx \sum_{j=1}^{U} \lambda(\eta_j)\Delta V \tag{5.20}$$

as an approximation for the integral, we can write (5.19) as

$$\begin{aligned} L(\theta) = &\sum_{i=1}^{n} \log(\lambda(1 - \exp(-\kappa\|\eta_i - \partial W\|))) \\ &- \sum_{j=1}^{U} \lambda(1 - \exp(-\kappa\|\eta_j - \partial W\|))\Delta V. \end{aligned} \tag{5.21}$$

Writing (5.21) as

$$\begin{aligned} L(\theta) = &\sum_{i=1}^{n} \log(\lambda) + \\ &\sum_{i=1}^{n} \log(1 - \exp(-\kappa\|\eta_i - \partial W\|)) - \\ &\lambda \sum_{j=1}^{U}(1 - \exp(-\kappa\|\eta_j - \partial W\|))\Delta V \end{aligned}$$

we can determine (if it exists) the maximum likelihood estimate in the standard way as detailed for example by Hogg & Craig (1959). We have that

$$\frac{\partial L}{\partial \lambda} = n\lambda^{-1} - \sum_{j=1}^{U}(1 - \exp(-\kappa\|\eta_j - \partial W\|))\Delta V$$

$$\frac{\partial L}{\partial \kappa} = \sum_{i=1}^{n} \frac{\|\eta_i - \partial W\|\exp(-\kappa\|\eta_i - \partial W\|)}{1 - \exp(-\kappa\|\eta_i - \partial W\|)} - \lambda \sum_{j=1}^{U} \|\eta_j - \partial W\|\exp(-\kappa\|\eta_j - \partial W\|)\Delta V$$

and the following solution at the log-likelihood stationary point

$$2\lambda = \sum_{i=1}^{n} \frac{\|\eta_i - \partial W\|\exp(-\kappa\|\eta_i - \partial W\|)}{1 - \exp(-\kappa\|\eta_i - \partial W\|)} \Big/ \sum_{j=1}^{U} \|\eta_j - \partial W\|\exp(-\kappa\|\eta_j - \partial W\|)\Delta V + \frac{n}{\sum_{j=1}^{U}(1 - \exp(-\kappa\|\eta_j - \partial W\|))\Delta V} \tag{5.22}$$

We can therefore substitute this solution into the above two equations and solve the equations

$$\frac{\partial L}{\partial \lambda} = 0, \quad \frac{\partial L}{\partial \kappa} = 0.$$

Finally, the stationary point is assessed to ensure it is a maximum.

There are two possibilities that can be adopted for considering replication. The first is that all replicates have the same set of Model 2 parameters. For this case we extend our likelihood equations as required and obtain for k iid replicates that (5.21) becomes

$$L(\theta) = \sum_{r=1}^{k}\sum_{i=1}^{n} \log(\lambda) + \sum_{r=1}^{k}\sum_{i=1}^{n} \log(1 - \exp(-\kappa\|\eta_{ir} - \partial W\|)) - \lambda \sum_{r=1}^{k}\sum_{j=1}^{U}(1 - \exp(-\kappa\|\eta_{jr} - \partial W\|))\Delta V \tag{5.23}$$

As before, we can determine (if it exists) the maximum likelihood in the standard way. We now have

$$\frac{\partial L}{\partial \lambda} = kn\lambda^{-1} - \sum_{r=1}^{k}\sum_{j=1}^{U}(1 - \exp(-\kappa\|\eta_{jr} - \partial W\|))\Delta V$$

$$\frac{\partial L}{\partial \kappa} = \sum_{r=1}^{k}\sum_{i=1}^{n} \frac{\|\eta_{ir} - \partial W\| \exp(-\kappa \|\eta_{ir} - \partial W\|)}{1 - \exp(-\kappa \|\eta_{ir} - \partial W\|)} - \lambda \sum_{r=1}^{k}\sum_{j=1}^{U} \|\eta_{jr} - \partial W\| \exp(-\kappa \|\eta_{jr} - \partial W\|) \Delta V$$

and we have the following solution for the log-likelihood curve for the replicated data stationary point

$$2\lambda = \sum_{r=1}^{k}\sum_{i=1}^{n} \frac{\|\eta_{ir} - \partial W\| \exp(-\kappa \|\eta_{ir} - \partial W\|)}{1 - \exp(-\kappa \|\eta_{ir} - \partial W\|)} \Big/ \sum_{r=1}^{k}\sum_{j=1}^{U} \|\eta_{jr} - \partial W\| \exp(-\kappa \|\eta_{jr} - \partial W\|) \Delta V + \frac{kn}{\sum_{r=1}^{k}\sum_{j=1}^{U}(1 - \exp(-\kappa \|\eta_{jr} - \partial W\|))\Delta V} \tag{5.24}$$

We have now used the result that Model 2 is an inhomogeneous Poisson process to describe how the standard MLE approach, that involves optimising (5.19) can be adjusted to apply to fitting three-dimensional replicated data, that is presented in the form of PPDS2. That is, we have a finite set of points, with each point categorised as one of event (for example PML body or nucleoli), empty space inside W (nuclear interior), or the boundary of W. To the best of our knowledge, a similar adjustment has not been put forward. The proceeding part of this section is dedicated to exploring through simulation, the feasibility of using this adjusted MLE approach to fit Model 2 to data that is similar to PPDS2.

The adjusted MLE procedure requires the integral in the log-likelihood expression (5.19) to be estimated. The approximation error obtained in adopting the quadrature scheme that we put forward, was assessed by simulating uniformly M times in the unit sphere and estimating the integral (5.19). The theoretical value of the integral (with parameters $\lambda = 29$ and $\kappa = 15.5$) was evaluated to 100.8 (see Table 5.2). To assess the entire adjusted MLE model parameter fitting approach (that utilises the quadrature scheme), we simulated 30 realisations of Model 2 with parameters $\lambda = 29$ and $\kappa = 15.5$ inside the unit sphere (and hence $E(X(B)) = 100$) in order to mimic 30 replicates with a high number of events. We fitted the simulated replicates to Model 2 by optimising the log-likelihood function (5.23) for the replicated inhomogeneous Poisson data using a grid search.

The grid search (carried out over a wide range, with the maximum values of the

Table 5.2: Simulation results for assessing the accuracy of the quadrature scheme

M	integral estimate	%error
10	88.67	12.03
100	101.7	-0.90
1000	101.29	-0.48
10000	101.3	-0.49
100000	100.66	0.14
200000	100.73	0.064
210000	100.73	0.068

grid search 5 times the true values) gave the optimal values $\lambda = \mathbf{29.72}$ and $\kappa = \mathbf{14.725}$. We also calculated the average number of events of the 30 replicates. The average was 102.87, which provides evidence for correct simulation and theoretical calculation of the expected number of events. We compared the grid search approach to fitting Model 2 to the same simulated data, but instead, using the theoretical maximum likelihood values (that is, using (5.24)) and obtained the estimates $\lambda = \mathbf{30.02}$ and $\kappa = \mathbf{14.734}$.

We repeated the model fitting via a grid search approach 1000 times to obtain the mean and variance, $\hat{\sigma}^2$ of the parameter estimates. The results are provided in Table 5.3. The standard error reported in 5.3 is the quantity given by $\hat{\sigma}/\sqrt{1000}$.

Table 5.3: Simulation results for fitting Model 2 to 1000 simulated realisations of Model 2

	κ	λ
mean	15.48	28.97
variance ($\hat{\sigma}^2$)	1.47	0.44
Standard Error	0.038	0.02

Rathbun & Cressie (1994) show that the maximum likelihood estimator is consistent and asymptotically normal, and asymptotically efficient as $W \to \mathbb{R}^d$ provided that a set of six conditions are met. These conditions impose constraints both on the intensity function and W (Rathbun & Cressie, 1994). Computing the inverse of the Fisher information matrix with entries

$$\mathcal{I}_{ij}(\theta) = E\left(\frac{\partial}{\partial\theta_i}\ln L(X;\theta)\frac{\partial}{\partial\theta_j}\ln L(X;\theta)\right)$$

for $\theta = (\theta_1, \theta_2) = (\lambda, \kappa)$ can provide the asymptotic variance of the parameters λ and κ. However evaluating $\mathcal{I}(\theta)^{-1}$ for Model 2 with PPDS2 is non-trivial. Adopting the use of Monte Carlo simulation to form numerical approximations is a possible approach. Also, bootstrapping has been recognised as a practical tool for estimating the sampling distribution of a broad class of parameter estimates for spatial point pattern models (Solow, 1989). Cowling et al (1996) use bootstrapping to obtain confidence regions for the intensity of an inhomogeneous Poisson process. However, Snethlage (1999) shows that it is not always necessary to use bootstrapping methods in such analysis. Specifically, Snethlage (1999) derives the confidence intervals for the estimate of an inhomogeneous Poisson intensity function that has been estimated using a non-parametric kernel estimation method.

In light of the results presented in Table 5.3, we applied a grid search, followed by a local stochastic search around the extrema of the log-likelihood function (5.23), to find the Model 2 parameters, θ^*_{M2}, that best describe the spatial locations of the PML NBs in the replicated cell nuclei that form PPDS2. We obtained the estimates $\theta^*_{M2} = \{\hat{\lambda} = 4.28 \times 10^{-5}, \hat{\kappa} = 0.022\}$ for PPDS2. Recall that for Model 2, the combination of λ and κ dictate the number of PML NBs that we expect to find in a cell nucleus. As an example, for cell 1 of PPDS2, the optimal parameters θ^*_{M2} yield an expected number of PML NBs of 11 (which is 2 less than the observed number of PML NBs for this cell nucleus).The parameter κ determines the probability of finding a PML NB within a certain distance of the nuclear boundary. Using Pseudo-Monte Carlo integration we approximate that for cell 1 of PPDS2, the optimal value of κ for the replicated data results in a probability of finding a PML NB within 13 units of the nuclear boundary ($1.1\mu m$, which is in the region of 5-10% the length of a typical nucleus) is 0.096, which is not too dissimilar to the empirical $F_{\partial w}(r)$ curve plotted in Figure 5.9.

5.3.3 Model 2 diagnostic tests on PPDS2

In the previous section, we assessed the appropriateness of Model 1 by utilising Monte Carlo based significance tests that entailed estimating the quantities (5.8) and (5.9) for the data, and ranking these estimates against estimates obtained from simulated data. The outcome of these tests showed that only 2% of the tests based only on the empty space function (just the f_i), provided evidence to reject Model 1. The same tests were repeated for Model 2. In this case, only 1.88% of these tests provided significant evidence to reject the model (see also the histogram in Figure

5.10).

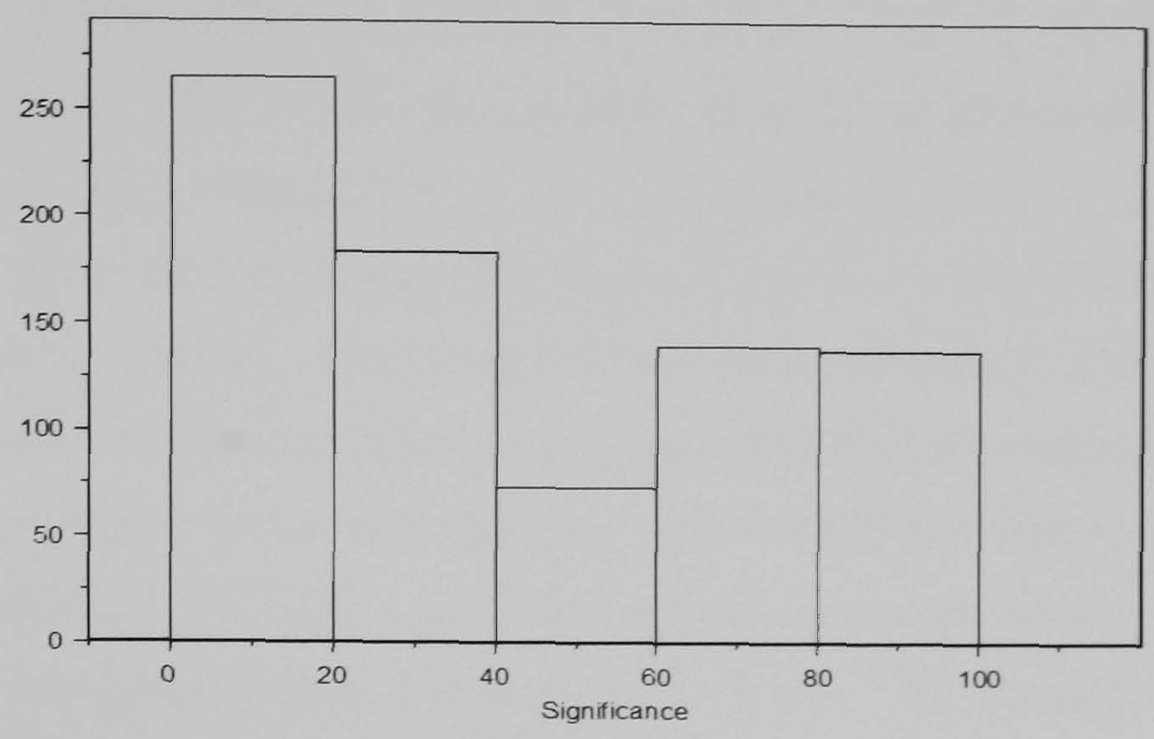


Figure 5.10: Histogram of cells 1 - 8 F-function significance levels against simulations from Model 2

We cannot fully rely on analysis based on the empty space function for assessing the appropriateness of a non-stationary model. Also, given the relative difficulty of estimating the $K_{\text{inhom}}(r)$-function for the PPDS2 data, we would not advocate adjusting the Monte Carlo test that uses the quantity f_i (in equation (5.9)) to using the $K_{\text{inhom}}(r)$-function. We are also cautious with respect to using the $K_{\text{inhom}}(r)$-function given that the theoretical form of $K_{\text{inhom}}(r)$ is the same for Model 2 as it is for a homogeneous Poisson process.

For the case where the fitted model is an inhomogeneous spatial Poisson process, a powerful diagnostic tool can be obtained by transforming to a Poisson process on the real line, with uniform intensity 1 on an interval (Baddeley et al, 2005). A mathematically rigorous discussion on transforming spatial point processes into Poisson processes is provided by Schoenberg (1999) and a more brief discussion, specific to inhomogeneous Poisson point processes by Cressie (1991, section 8.5.1). We will now outline how we utilise the method of transforming data to a Poisson process on the real line as a diagnostic tool for assessing the appropriateness of Model 2. Similarly to Diggle (1990), for a particular nuclei, W_j, we order the PML NBs $\xi_i, i = 1 \ldots n$ in increasing distance from the centre of mass of the nucleus, η_{cm}, given by

$$\eta_{\text{cm}} = \left(\frac{1}{U} \sum_{s=1}^{U} \eta_s^{(1)}, \frac{1}{U} \sum_{s=1}^{U} \eta_s^{(2)}, \frac{1}{U} \sum_{s=1}^{U} \eta_s^{(3)} \right) \quad \eta_s \in W_j$$

Denoting the distances by $\delta_1, \ldots, \delta_n$ and defining the quantities

$$t_i = \int_{b^{(3)}(\eta_{\mathrm{CM}}, \delta_i) \cap W_j} \hat{\lambda}(1 - \exp(-\hat{\kappa}\|\eta - \partial W_j\|))d\eta \quad i = 1, \ldots, n \tag{5.25}$$

then under Model 2, the t_i are the points of a homogeneous Poisson process on $\mathbb{R}^+$. Furthermore, the quantities $u_i = t_i/t_n$ and $g_i = t_{i+1} - t_i$ are independent random samples from the uniform distribution on (0,1) and exponential distribution respectively (Diggle, 1990). Departure from unit intensity suggests a misspecified spatial trend, while departure from the exponential distribution of the inter-arrival times is evidence of interpoint interaction. We calculated the t_i defined by (5.25) for all of the cell nuclei of PPDS2 using the usual, as used above quadrature scheme. EDF plots of u_i are provided in Figures 5.11- 5.13. The results of cell 7 (see Figure 5.11) are of particular interest as this cell contains the highest PML NB count. A

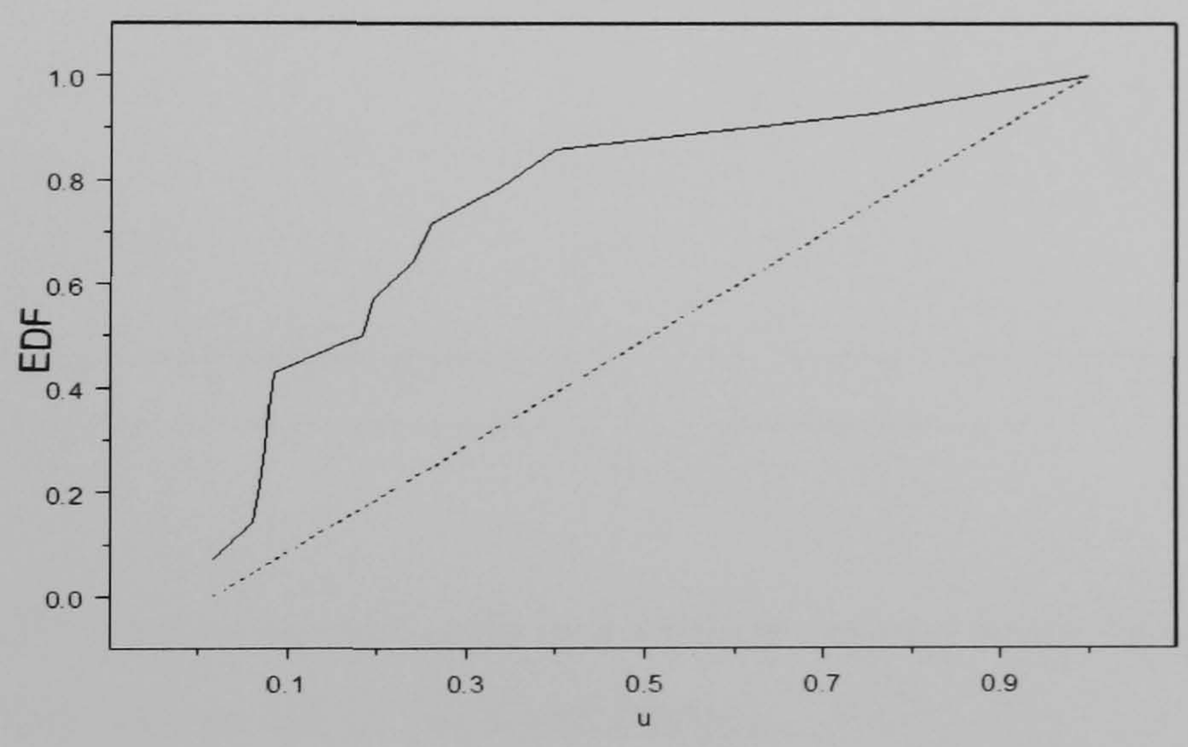
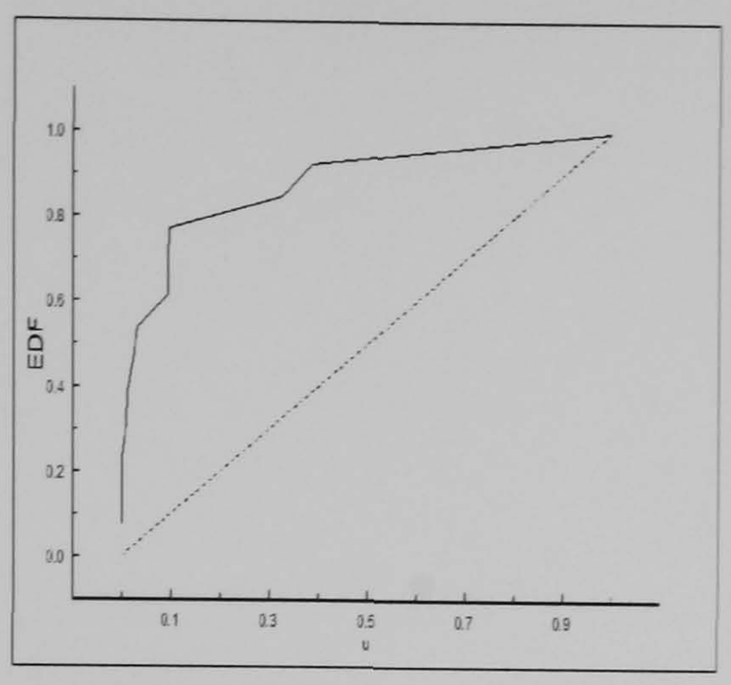
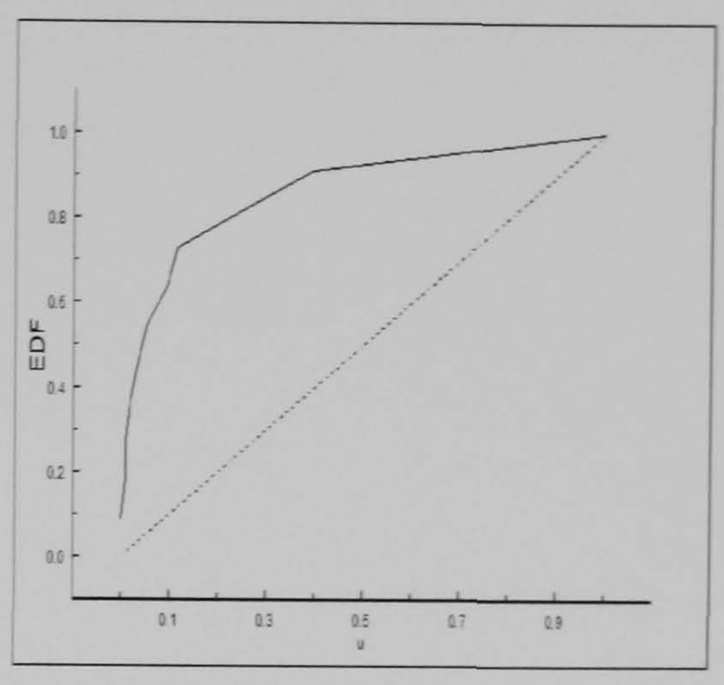
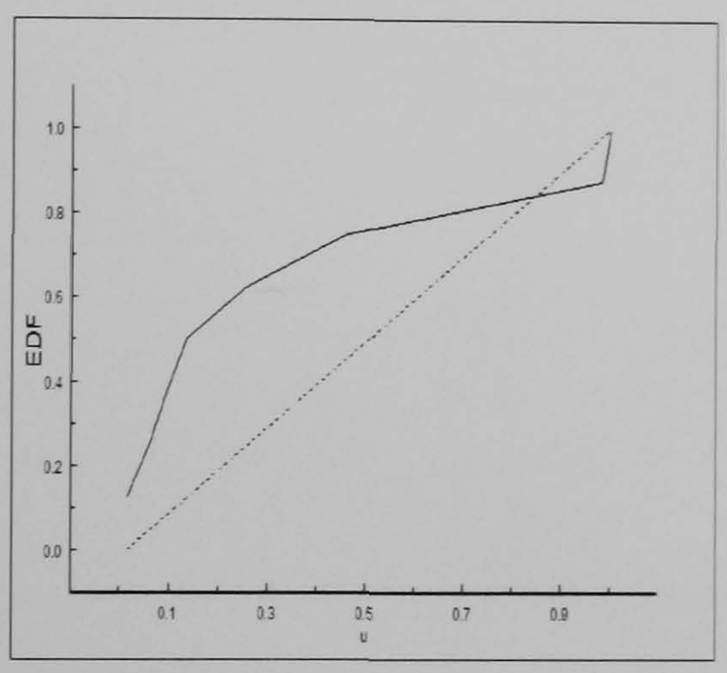


Figure 5.11: EDF plot of u_i for cell 7 (black curve) with the theoretical CDF of a uniform distribution on (0,1) (dashed curve)

two-tailed Kolmogorov-Smirnov test for cell 7 provided a p-value of 0.039 for the null hypothesis that the u_i are uniform on (0,1) (see Table 5.4). A similar test for the null hypothesis that the g_i have an exponential distribution with rate $\hat{\rho}$ estimated from the g_i gave a p-value of 0.002. Hence there is reason to doubt the form of the intensity function for describing the spatial distribution of the PML NBs in cell nucleus 7. Furthermore, for nucleus 7 there is evidence to suggest possible interaction between the PML NBs. More specifically, the g_i appeared more dispersed compared to an exponential distribution, indicating signs of clustering. The results obtained for the other cells suggest that the intensity function for Model 2 is appropriate for describing the spatial distribution of the PML NBs in cells 3 and 8, but not the







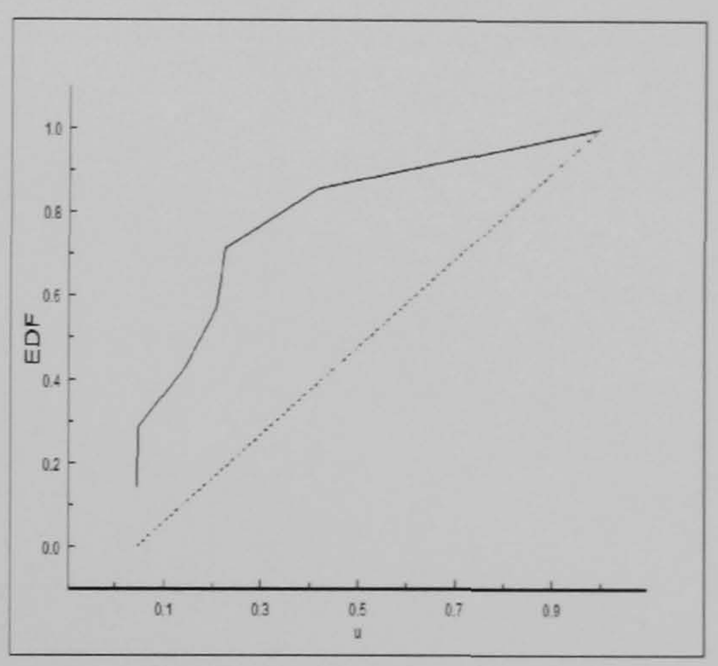
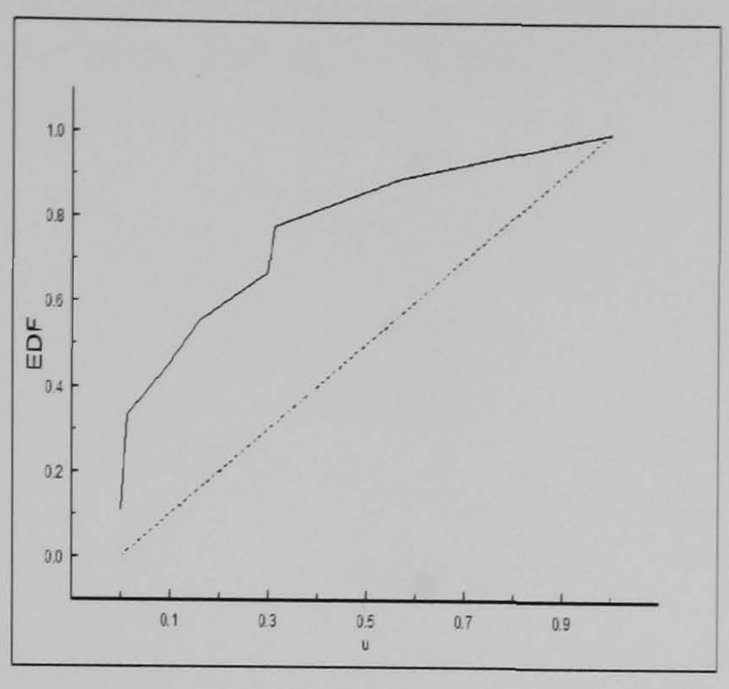


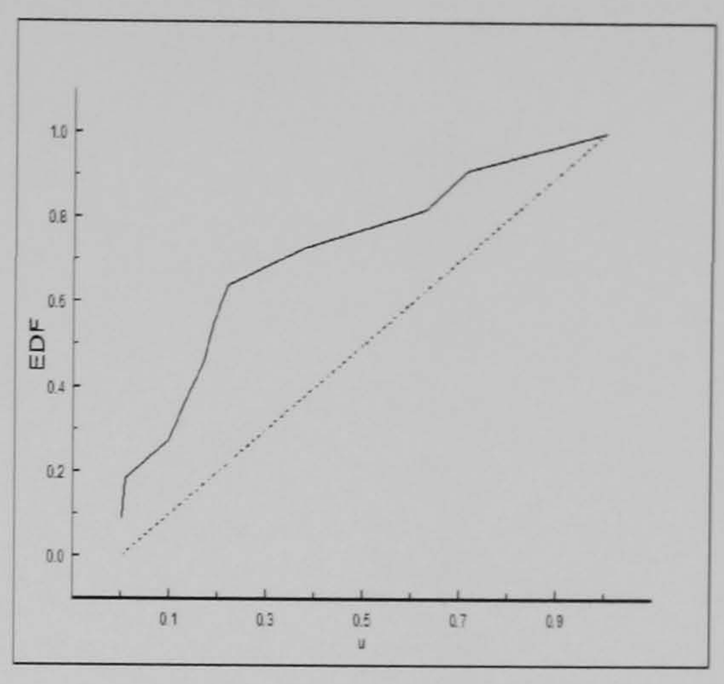
Figure 5.12: EDF plot of u_i for cells 1-4 (black curve) with the theoretical CDF of a uniform distribution on (0,1) (dashed curve)

other cells of PPDS2. Furthermore, there appears to be some evidence of PML NB interaction in cells 1 and 2. We estimated the log-likelihood of Model 2, L_{M_2}, at the optimal parameters θ^*_{M2}, to being -994.46. The log-likelihood of Model 1 at the MLE, $\widehat{\lambda_{\mathrm{MLE}}}$ of the intensity parameter for k iid replicates is given by

$$N\log(\widehat{\lambda_{\mathrm{MLE}}}) - \widehat{\lambda_{\mathrm{MLE}}} \sum_{i=1}^{k} \nu(W_i) \tag{5.26}$$

where $N = \sum_{i=1}^{k} n_i$, is the total number of events for all replicates and $\nu(W_i)$ is the volume of replicate i. Evaluating (5.26) at the MLE of the intensity parameter of Model 1 we obtained a log-likelihood, L_{M_1}, of -1007.4. In classical statistical analysis we can test formally whether there is a significant difference between the likelihoods, in the case of nested models by using the result that twice the difference





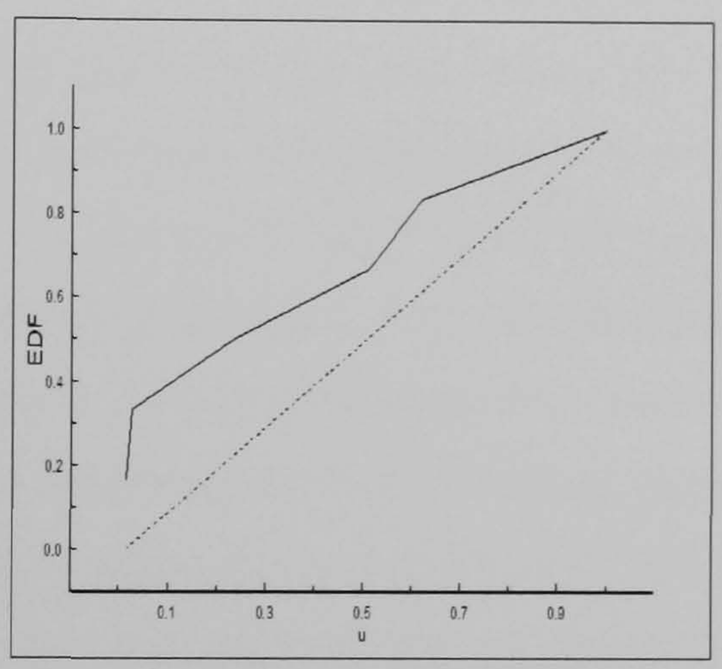


Figure 5.13: EDF plot of u_i for cells 5,6, and 8 (black curve) with the theoretical CDF of a uniform distribution on (0,1) (dashed curve)

of loglikelikelihoods is asymptotically chi-squared. This result is used by Diggle (1990) for comparing two sets of parameters of an inhomogeneous Poisson process. Such a result may not hold for comparing Model 1 against Model 2 as Model 1 is obtained by setting a parameter of Model 2 (κ) to a value on the boundary of the parameter space. For non-nested models, methods such as the simulation and bootstrap techniques detailed by Kapetanios & Weeks (2003) can be adopted. Cressie (1991) has cited Lin'kov (1985) to show that, for an inhomogeneous Poisson process defined on $[0, T]$, a test based on the log-likelihood ratio is asymptotically ($T \to \infty$) most powerful for testing the null hypothesis of one set of parameters versus another.

As noted above, there are two ways of extending Model 2 to handle replicated patterns. The first approach that we have detailed above, is to assume that all cell

Table 5.4: p-values for KS tests for the appropriateness of Model 2

Cell	p-value for $u_i \sim$ Uniform(0,1) test	$p-$value for $g_i \sim$ Exponential test
1	0	0.0008
2	0.0002	0.048
3	0.18	0.45
4	0.047	0.84
5	0.026	0.54
6	0.029	0.83
7	0.039	0.002
8	0.52	0.83

nuclei share the same model parameters and hence we model the nuclei jointly by optimising a log-likelihood function for replicated data. The second approach assumes that the parameters of Model 2 vary with the cell. Note that this would not in this case contradict our test result that the spatial distribution of the PML NBs is generated by the same spatial point process since we have shown that the $K_{\text{inhom}}(r)$ function is the same for all inhomogeneous Poisson processes. We estimated on a cell by cell basis, the optimal parameters for Model 2. The parameter estimates are reported in Table 5.5 ($\hat{\Lambda}$ is an estimate of the intensity measure). From the discus-

Table 5.5: Model 2 parameter estimates for nuclei in PPDS2

Cell	no. pml	$\hat{\kappa}$	$\hat{\lambda}$	$\hat{\Lambda}$	Standard Error($\hat{\lambda}$)	Standard Error($\hat{\kappa}$)
1	13	0.0102	1.0×10^{-4}	13.07	8.34×10^{-5}	8.96×10^{-4}
2	11	0.0107	9.0×10^{-5}	11.02	8.55×10^{-5}	1.06×10^{-3}
3	8	0.0107	5.0×10^{-5}	8.09	8.08×10^{-5}	1.02×10^{-3}
4	7	0.0107	5.0×10^{-5}	6.66	7.27×10^{-5}	9.30×10^{-4}
5	9	0.0101	8.0×10^{-5}	9.05	8.99×10^{-5}	1.05×10^{-3}
6	11	0.0102	9.0×10^{-5}	11.07	8.90×10^{-5}	9.99×10^{-4}
7	14	0.0107	1.2×10^{-4}	13.75	8.98×10^{-5}	1.06×10^{-3}
8	6	0.0107	5.0×10^{-5}	5.86	8.39×10^{-5}	1.07×10^{-3}

sions (at the end of the previous section) on methods for obtaining the standard errors for the estimates of the parameters of an inhomogeneous Poisson model, it appears that the standard errors for the estimates for Model 2 are hard to come by for data in the form of PPDS2. We obtained empirical estimates of the standard errors by simulating 100 times from Model 2 (inside the appropriate cell nucleus)

and repeating the model fitting procedure for each of the simulated realisations. From Table 5.5 we see that different cells have different parameters. Since some of the parameters appear quite similar, it is not clear if these are genuine differences or differences arising for example from experiment measurement error. Nevertheless, we can extend the definition of Model 2 to have the additional postulate

- M2P0 $(\log(\kappa), \log(\lambda)) \sim$ Z

where Z is some multivariate distribution that respects the range of the parameters κ and λ. Hence Z is for example a bivariate lognormal distribution. The model that is specified by M2P0-M2P2 is in fact a Cox process, since it is an inhomogeneous Poisson process with a random intensity measure. Furthermore, for this type of approach we are allowing for variation amongst cells. This is supported by some earlier results, which showed for example that some cells exhibit much great departure from CSR than others. When adopting a model specified by P0M2-P2M2 we have to bear in mind that we only have 8 nuclei in PPDS2. Although the Imperial College centre for structural biology has produced more data, we have been constrained by our cautious approach that entails analysing nuclei that are as similar as possible.

For the avoidance of over-fitting a model, data recently obtained from Imperial College London centre for structural biology (detailed in Appendix B as PPDS2(b)) was used as PPDS2 data, set aside (from the modeling process) for carrying out further model diagnostics. Estimates of the $F_{\partial W}(r)$ function for the new data in PPDS2(b) are encouraging (see Figure 5.14) in the sense that they provide some evidence for the need for a model that captures possible PML NB interaction with the nuclear boundary. In particular, on visual inspection of the graphs in Figure 5.14, cells 2, 4, 5, and 6 of PPDS2(b) provide some evidence that the PML NBs tend to be located further away from the nuclear boundary compared to if the PML NBs were uniformly placed in the cell nucleus. Further diagnostics were performed on Model 2 by calculating the quantities provided in Table 5.4 for the data in PPDS2(b) (the data that was not used in the modeling process). The results of the additional diagnostic tests are shown in Table 5.6. The results in Table 5.6 provide supporting reason to question the validity of the form of the intensity function of Model 2 (when applied to the cells in PPDS2(b)). Cells 4 and 5 provide evidence for interaction between the PML NBs.

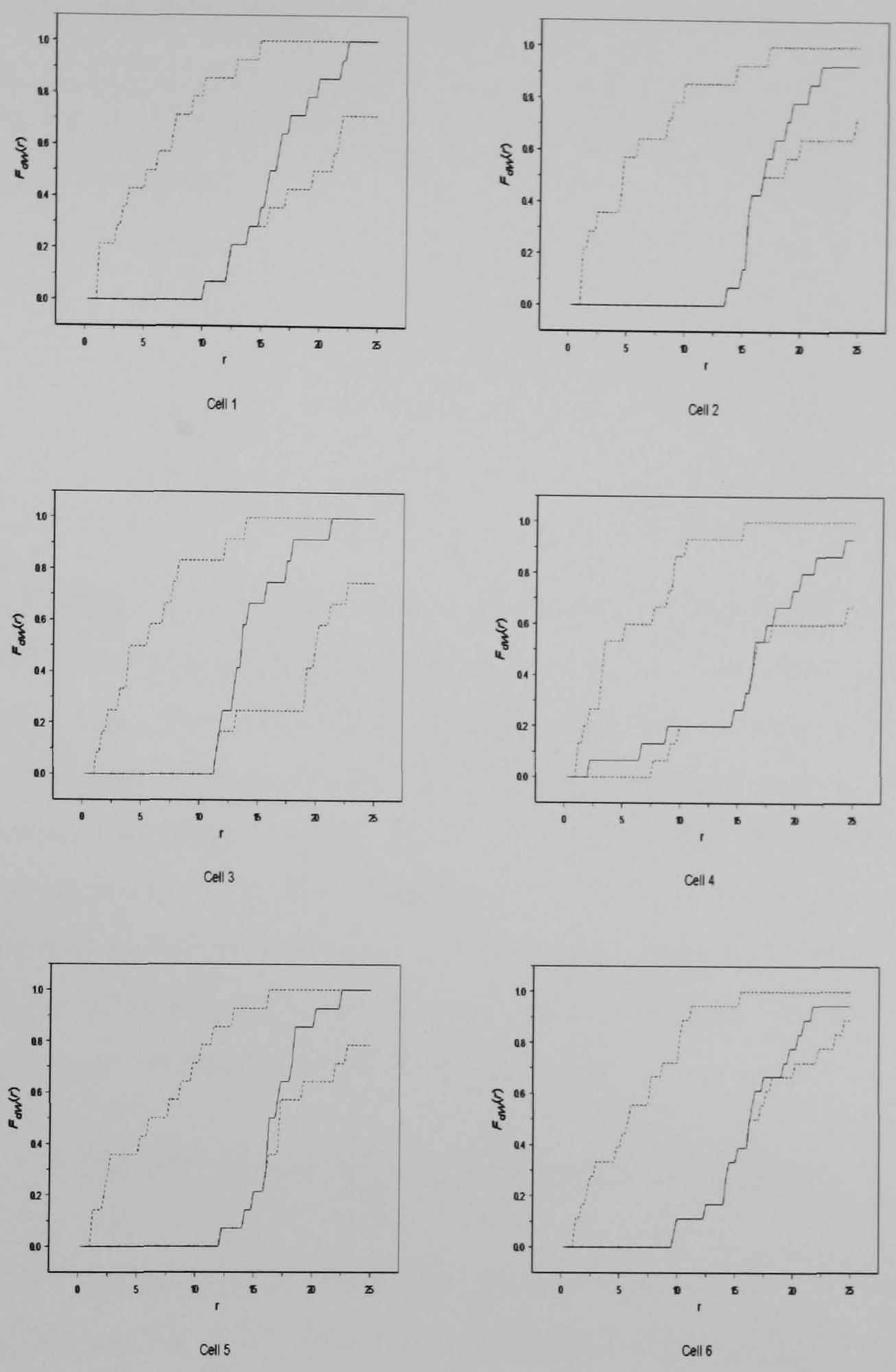


Figure 5.14: Cells 1-6 of PPDS2(b) estimate of $F_{\partial W}(r)$ with simulation envelopes from 99 simulations from Model 1.

5.4 Model 3: Boundary effects and PML NB interaction

So far we have fit a simple homogeneous model to PPDS2, where we treat PML NBs as uniformly distributed in the region W^*, of the nuclear interior. Recall, that is, if $W \subset \mathbb{R}^3$ is the nuclear interior, then $W^* \subset W$ is the region unoccupied by nucleoli. We have then extended this model to fit an inhomogeneous Poisson process defined on W^*, with an intensity function that captures PML NB interaction with

Table 5.6: p-values for KS tests for the appropriateness of Model 2 for PPDS2(b) PML NB data

Cell	p-value for $u_i \sim$ Uniform(0,1) test	$p-$value for $g_i \sim$ Exponential test
1	0.006	0.668
2	0.001	0.474
3	0.012	0.167
4	0	0.0433
5	0.0015	0.430
6	0.0001	0.0027

the nuclear boundary. Diagnostic tests on Model 2 revealed that for cells 1, 2, and 7 of PPDS2 there was perhaps evidence to support the existence of interaction between the PML NBs. We now study a model that incorporates the characteristics of the previous two, but with the additional feature of being able to capture possible interaction between the PML bodies. More precisely, let X be a finite point process defined on a bounded set $A \subset \mathbb{R}^d$. Suppose also that X has density $f_X(V;\theta)$ with respect to a Poisson point process on A with unit intensity ($X_{\text{poi}(1)}$); we consider a class of Markov models where the characterising density function (we drop the subscript X from now onwards), $f(V;\theta)$ is given by

$$C(\theta)f(V;\theta) = \alpha \prod_{i=1}^{n} \beta(\eta_i) \prod_{i \neq j} \Psi_R\left(\|\eta_i - \eta_j\|\right) \tag{5.27}$$

for $V \subset A$ and $C(.)$ a normalising constant. Furthermore, $n < \infty$ is the cardinality of V, denoted card(V), and $\eta_i \in V$. That is V belongs to $\Omega = \bigcup_{i=0}^{\infty}\{x \subset \mathbb{R}^3 : \text{card}(x) = i\}$, the set of all finite point configurations. $\theta = (\alpha, \kappa, R) \in (0,\infty) \times (0,\infty) \times (0,\infty)$ belongs to the three-dimensional parameter space, Θ, of the model.

Markov models of this kind, that model interaction in an inhomogeneous environment, have been studied by Ogata and Tanemura (1986), Stoyan & Stoyan (1998), and Jensen & Nielsen (2000). Hahn et al (2003) have introduced a new class of models called locally scaled point processes for modelling inhomogeneous Markov processes. Recently Prokešovaá et al (2006) carried out analysis of this class of models.

For this model the spatially varying intensity function $\beta(.)$ in (5.27) has the form

$$\beta(\eta_i) = \left(1 - \exp\left(-\kappa\|\eta_i - \partial W\|\right)\right). \tag{5.28}$$

This is the same as the intensity function for Model 2. Furthermore, the function $\Psi_R(u)$ (which we refer to as a pair-potential function (see for example Cressie, (1991) for use of this nomenclature) is given by

$$\Psi_R(u) = \begin{cases} 0 & u < R \\ 1 & u \geq R \end{cases} \tag{5.29}$$

This model restricts PML bodies to being at least a distance R apart. A possible physical interpretation of this is that PML are bodies which occupy a volume of space. Furthermore, an assumption of this model is that we assume that the PML NBs cannot coalesce. However, the model can be relaxed to non-strict inhibition. This enables the model to capture the idea of PML bodies occupying various volumes of space. One possible approach for incorporating non-strict inhibition is as given by Diggle (2003), where the pair-potential function is defined as

$$\Psi_R(u) = \begin{cases} 0 & u < R \\ v_1 - v_2(u - R) & R \leq u < v_3 \\ 1 & u \geq v_3 \end{cases} \tag{5.30}$$

This allows an element of attraction in the model. Although it is appealing to have this extra degree of flexibility, we are cautious to not over-fit by making the parameter space too large, given the low PML body counts. We therefore do not consider this extra complication.

5.4.1 Estimating the parameters of Model 3 using MLE

The normalising constant in (5.27) is given by

$$C(\theta)^{-1} = \sum_{n=0}^{\infty} \int_B \dots \int_B f(V;\theta)\nu(d\eta_1)\dots\nu(d\eta_n) \tag{5.31}$$

The complexity of evaluating the normalising constant makes computing quantities such as the likelihood more difficult. However, similar to as derived by Diggle (2003), for any $\theta_0 \in \Theta$, we have in the case where n is fixed,

$$\begin{aligned} C(\theta)^{-1} &= \int_B \dots \int_B f(V;\theta)\frac{C(\theta_0)f(V;\theta_0)}{C(\theta_0)f(V;\theta_0)}\nu(d\eta_1)\dots\nu(d\eta_n) \\ &= C(\theta_0)^{-1}E_{X|\theta=\theta_0}[f(V;\theta)/f(V;\theta_0)] \end{aligned} \tag{5.32}$$

Then the likelihood function, $l(\theta)$ is given by

$$l(\theta) = C(\theta)f(V;\theta) = \frac{C(\theta_0)f(V;\theta)}{E_{X|\theta=\theta_0}[f(V;\theta)/f(V;\theta_0)]} \tag{5.33}$$

Therefore, for a given value of θ_0, for which we can calculate $C(\theta_0)$ directly and simulate a realisation of the model with parameters θ_0, we can in principle obtain an empirical approximation to $C(\theta)$. An example of a case when we can calculate an approximation for $C(\theta_0)$ is when $\kappa >> 1$ and $R << 1$. In this case, for n fixed and bounded set B, by (5.31) $C \approx n\alpha\nu(B)$. However, Diggle (2003) goes on to warn that numerical instability may cause the method to fail unless θ is close to θ_0.

For k iid replicates, using this empirical approximation to the expectation term in (5.33), we can obtain an approximation to the maximum likelihood estimate of θ by maximising the function

$$\hat{\mathrm{L}}(\theta) = \sum_{i=1}^{k} \left(\log f(x_i;\theta) - \log \sum_{j=1}^{p} \frac{f(x_i;\theta)/f(\tilde{x}_{ij};\theta_0)}{p} \right) \tag{5.34}$$

where $x_1, \ldots, x_k$ are the replicates and $\tilde{x}_{i1}, \ldots, \tilde{x}_{ip}$ are simulated realisations from the model in nucleus i with $\theta = \theta_0$.

The realisations $\tilde{x}_{i1} \ldots \tilde{x}_{ip}$ can be simulated using Ripley's (1977) method for simulating Markov point processes on the bounded Borel set $B \subset \mathbb{R}^d$ when n is fixed. The spatial birth-and-death process (see for example Berthelsen & Møller, 2002(b), for a definition) provides a framework for this method (Cressie, 1991). Briefly, the technique works by considering the ratio (Papangelou conditional intensity. See Chapter 2)

$$\frac{f(V \bigcup \xi;\theta)}{f(V;\theta)} = \beta(\xi)\Pi_{\eta_i \neq \xi}\Psi(\|\xi - \eta_i\|) \tag{5.35}$$

and we note that, since $0 \leq \Psi(.) \leq 1$ and $0 < \beta(.) \leq 1$,

$$P(\xi) = \beta(\xi)\Pi_{\eta_i \neq \xi}\Psi(\|\xi - \eta_i\|) \tag{5.36}$$

is a probability. The simulation algorithm is initiated with n events uniformly placed in W. One of the events is then deleted at random. A singleton, ξ is then placed uniformly in W, and is accepted with probability $P(\xi)$. Otherwise ξ is rejected, and this step is repeated until a drawn ξ is accepted. This depletion and replacement step is repeated m times. Ripley (1987) suggests that $m = 10n$ iterations are sufficient for convergence. However, Ripley (1979) suggests that $4n$ iterations are sufficient in practice. As in Chapter 3 for simulating a Strauss process, quicker convergence can be obtained by carrying out the depletion-replacement steps in a systematic order (Diggle, 2003).

5.4.2 Perfect simulation of Markov models through coupling from the past (CFTP)

Recently perfect or 'exact' simulation has gained much attention. as it allows the exact simulation of Markov point processes. In a review on applications to stochastic geometry, Møller (2001) describes perfect simulation as one of the most important and exciting developments in stochastic simulations. We adopt perfect simulation for model fitting. More specifically, we use the simple dominated Coupling From The Past algorithm (CFTP) based on spatial birth-and-death processes as outlined by Berthelsen & Møller (2002), Ambler & Silverman (2004), and Kendall & Moller (2000) in order to simulate from the model specified by (5.27)-(5.29) (Model 3). A sufficient condition for applying dominated CFTP based on spatial birth-and-death processes and Metropolis-Hastings algorithms is that $f(.)$ is locally stable. That is we require that for some constant $K > 0$

$$f(V \bigcup \xi;) \leq K f(V;), \quad \forall V \in \Omega \quad \text{and} \quad \xi \in \mathbb{R}^3 \backslash \{X\} \tag{5.37}$$

This condition clearly holds for $f(.)$ as specified by (5.27)-(5.29).

Before detailing the CFTP algorithm that we adopt, we first recall some properties of the spatial birth-and-death process. For this process, if at time t the events of the process X_t are located at $\{\xi_1, \ldots, \xi_n\}$ then the probability that an event is born in a bounded Borel set B during the time interval $[t, t+\delta)$ is

$$\delta \int_B \mu(\eta; \{\xi_1, \ldots, \xi_n\}) \nu(d\eta) + O(\delta)$$

where μ is the birth rate defined by

$$\mu(\eta; \{\xi_1, \ldots, \xi_n\}) \equiv \lim_{\nu(d\eta) \to 0, \delta \to 0} \frac{P(\text{birth in d}\eta \text{ during} [t, t+\delta) | \xi_1, \ldots, \xi_n)}{\delta \nu(d\eta)}$$

Furthermore, if events at time t are located at $\{\xi_1, \ldots, \xi_n, \eta\}$ then the probability that event η dies during the interval $[t, t+\delta)$ is $\delta \mu^c(\eta, \{\xi_1, \ldots, \xi_n\}) + O(\delta)$ where μ^c is the death rate defined as

$$\mu^c(\eta, \{\xi_1, \ldots, \xi_n\}) \equiv \lim_{\delta \to 0} \frac{P(\text{death of event } \eta \text{ during} [t, t+\delta) | \xi_1, \ldots, \xi_n)}{\delta}$$

Consider a spatial birth-and-death process $\{X_t : t \in [t_0, \infty)\}$ with birth rate $\mu \geq 0$ and death rate $\mu^c > 0$ both measurable functions so that the integral $J(V) = \int \mu(\eta; V) d\rho(\eta)$, for some arbitrary diffuse probability measure ρ. is finite for all $V \in \Omega$. Kendall & Moller (1999) state that the process X is a stationary

jump process that evolves as follows. Suppose that we condition on the event that $X_t = x = \{\xi_1, \ldots, \xi_n\}$ and we let $E_0, \ldots, E_n$ be exponentially distributed random variables that are mutually independent and independent of past history with respective means $1/J(x), 1/\mu^c(\xi_1, x\backslash\xi_1), \ldots, 1/\mu^c(\xi_n, x\backslash\xi_n)$ (for $J(x) = 0$ we set $E_0 = \infty$). The next transition occurs at time $t + E$, where $E = \sup\{E_0, \ldots . E_n\}$. If $E = E_0$ then $X_{t+E} = x \cup \{\xi\}$ represents birth of a point ξ with density $\nu(\cdot, x)/J(x)$. If $E = E_i$ with $i > 0$ then $X_t = x\backslash\{\xi_i\}$ represents death of the point ξ_i.

The CFTP algorithm for simulating X, from Model 3 can be outlined as follows. Let $\ldots Z_{-2}, Z_{-1}, Z_0$ denote the jump chain of $D = \{D_t : t \in \mathbb{R}\}$, a time stationary and reversible spatial birth-and-death process with equilibrium distribution such that $D_t \sim X_{\text{poi}(1)}$

- Simulate $Z_0 \sim X_{\text{poi}}$ with intensity K
- Generate backwards the jump chain $Z_0, \ldots, Z_{-T_0}$ via a spatial birth-and-death process with birth rate μ by making a backwards birth with probability $\frac{K\mu}{(K\mu+n)}$. Furthermore, generate the associated marks s_i $\sim$Uniform(0,1) for forward births $Z_i = Z_{i-1} \cup \eta_i$
- Set $Y_{-T_0} = \emptyset$, where $T_0 = \inf\{i \in \mathbb{Z}^+ : Z_{-i} = \emptyset\}$ and define recursively $Y_{-T_0}, \ldots Y_0$ using the rule

$$Z_i = Z_{i-1} \cup \eta_i \Rightarrow Y_i = \begin{cases} Y_i \cup \eta_i & s_i \leq \frac{\Psi(Y_{i-1} \cup \eta)}{\Psi(Y_i)} \\ Y_{i-1} & \text{otherwise} \end{cases}$$

$$Z_i = Z_{i-1} - \eta_i \Rightarrow Y_i = Y_{i-1} - \eta_i$$

- Return $Y_0 \sim X$

Note that in practical simulation-based application to the confocal microscoy data used, we re-scale with $0 < K \leq 1$ to cope with the low PML intensity. That is. the cell nucleus volume is very large, and simulating $X_{\text{poi}(1)}$ events sometimes results in generating many events. This can lead to computer memory problems.

We can incorporate perfect simulation into a Metropolis-Hastings MCMC routine, to obtain estimates of Model 3 parameters for replicated point pattern data. Such a procedure can be outlined as follows:

- Begin with the parameters set θ_t

- Propose a new set of parameters, θ_{t-1}, where $\alpha_{t+1} \in \theta_{t+1} \sim \Upsilon(\theta_t)$ and similarly for κ_{t+1} and $R_{t+1} \leq R_{\text{Threshold}}$
- Set $\theta_{t+1} = \theta_t$ with probability $\min\left\{1, \frac{L(\theta_t)\pi(\theta_t)}{L(\theta_{t+1})\pi(\theta_{t+1})}\right\}$
- Repeat the previous two steps m times

Where $\pi(.)$ denotes the prior probability and Υ some symmetric proposal distribution that respects the range of the parameters. As a means of saving computational burden, at the early stages of our model fitting approach we set $\hat{R} = \sup\{\|\eta_{il} - \eta_{jl}\| : \eta_{il}, \eta_{jl} \in X_l, i \neq j, l = 1 \ldots k\}$ as an initial estimate for the parameter R. Furthermore, this is clearly the maximum value, $R_{\text{Threshold}}$, that R can be for the replicated data.

5.4.3 Estimating the parameters of Model 3 via path sampling

Note that we cannot use perfect simulation to estimate the likelihood using (5.34) as there is no practical method for perfect simulation conditional on n (Berthelsen & Møller, 2003). However, path sampling, or 'thermodynamic integration' (see for example Gelman & Meng, 1998) in the terminology used in statistical physics, allows for the estimation of the ratio of normalising constants. Berthelsen & Møller (2003) have demonstrated how path sampling can be incorporated with perfect simulation to estimate ratios of normalising constants for a Strauss process. We incorporate a path sampling approach for estimating the ratio of the posterior densities $\frac{L(\theta_0)\pi(\theta_0)}{L(\theta_1)\pi(\theta_1)}$ in the Metropolis Hastings routine for finding the parameters of Model 3. The objective is to avoid the need for having to specify an initial set of parameters that are an approximation to the true parameters being sought.

The path sampling routine we adopt is as follows; we wish to calculate the ratio $C(\theta_1)/C(\theta_0)$ for $\theta_1, \theta_0 \in \Theta$. We begin by selecting a continuous path in the parameter space that links θ_1 and θ_0

$$\theta(t) = (\alpha(t), \kappa(t), R(t)) = (\theta^{(1)}(t), \theta^{(2)}(t), \theta^{(3)}(t)) \quad \text{for} \quad t \in [0, 1]$$

with $\theta(0) = \theta_0$ and $\theta(1) = \theta_1$. Defining

$$\dot{f}_w(V; \theta) = \frac{\partial \log f(V; \theta)}{\partial \theta^{(w)}}$$

and

$$d\theta^{(w)}(t) = \frac{d\theta^{(w)}(t)}{dt}$$

we use the result

$$\begin{aligned}\log\left(\frac{C(\theta_1)}{C(\theta_0)}\right) &= \int_0^1 E_{\theta(t)}\left(\frac{d\log f(V;\theta(t))}{dt}\right)dt \\ &= \int_0^1 E_{\theta(t)}\left(\sum_{w=1}^{3} d\theta^{(w)}(t)\dot{f}_w(V;\theta(t))\right)dt \qquad (5.38)\end{aligned}$$

From (5.38), we use the path sampling estimator, $\widehat{E}(\theta_1,\theta_0)$, for $\log\left(\frac{C(\theta_1)}{C(\theta_0)}\right)$ given by

$$\widehat{E}(\theta_1,\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\sum_{w=1}^{3} d\theta^{(w)}(t_i)\dot{f}_w(\widetilde{X}_i,\theta(t_i)) \qquad (5.39)$$

where the t_i are random draws from Uniform(0,1). As stated by Gelman & Meng (1998), (5.39) is a consistent and unbiased estimator for $C(\theta_1)/C(\theta_0)$ provided that the sample average converges to its population average. Combining the path sampling with the Metropolis-Hastings algorithm, for k iid replicates we replace the Hastings ratio with

$$\frac{\pi(\theta_1)}{\pi(\theta_2)}\prod_{i=1}^{k}\left(e^{\widehat{E}(\theta_1,\theta_2)}\frac{f(X_i;\theta_1)}{f(X_i;\theta_2)}\right)$$

Besides the result

$$\frac{\partial \log f}{\partial \alpha} = \frac{1}{\alpha}$$

the other partial derivatives of f are unknown or have non-trivial closed form and thus we use the standard numerical approximation

$$\frac{\partial \log f}{\partial \kappa} \approx \frac{\log f(V;(\kappa+\Delta,\alpha,R)) - \log f(V;(\kappa-\Delta,\alpha,R))}{2\Delta}$$

and similarly for R. We generalise to using any permitted values of the parameters by using the path

$$\theta^{(w)}(t) = \theta_0^{(w)} + (\theta_1^{(w)} - \theta_0^{(w)})t \quad w = 1,\ldots,3$$

and hence $d\theta^{(w)}(t)$ is simply

$$\theta_1^{(w)} - \theta_0^{(w)}.$$

For a more detailed background on this path sampling approach, see Section 2 of Gelman & Meng (1998).

We assessed the model-fitting process that utilises the path-sampling methodology by constructing a dataset that is similar to PPDS2 (and is a realisation of Model 3) and then attempted to fit Model 3 to the simulated data. Using perfect simulation, we simulated Model 3 with the parameters $\alpha = 100, \kappa = 0.004$ and $R = 1.5$ inside the sphere of radius 10 (see Figure B.5 in Appendix B). This was done 10 times (to form 10 replicates). The parameters were selected by trial and improvement to give the ideal scenario of a high number of PML bodies. For the replicate (shown in Figure B.5), an average of 36 events were produced. We also formed some points that mimicked locations in the nuclear interior (as provided by PPDS2 and PPDS3), by simulating uniformly inside the sphere of radius 10. In this case, we are of course in a position to be able to calculate exactly the distance of any point to the boundary of the spherical sampling window.

The parameter estimates for the data simulated from Model 3 were obtained as follows. Let $\tilde{x}_1, \ldots, \tilde{x}_k$ be the simulated realisations of Model 3. Our estimate for R, $\hat{R}$, as detailed earlier is given by

$$\hat{R} = \inf\{\|\eta_{ij} - \eta_{ik}\| : \eta_{ij}, \eta_{ik} \in \tilde{x}_i, i = 1 \ldots k\} \tag{5.40}$$

Using (5.40) we obtained the value $\hat{R} = 1.54$ and in this trial, kept this as the overall estimated value for R. The parameters α and κ were estimated using the Metropolis-Hastings routine outlined earlier (without a threshold for R since this was fixed already) with 8000 steps (parameter proposals). We adopted uniform priors and the proposal distributions

$$\begin{aligned} \alpha_{t+1} &\sim |N(\alpha_t, 10)| \\ \kappa_{t+1} &\sim |N(\kappa_t, 2)|, \end{aligned} \tag{5.41}$$

where the variance of the Normal distributions in (5.41) have been chosen to be large in comparison to the magnitudes of the parameters α and κ, and hence resulting in a 'noisy' proposal distribution, in the sense that we do not restrict the stochastic optimisation routine to a small region of the parameter space. Using this approach with starting parameters $\alpha = 1, \kappa = 1$ and $\Delta = 1 \times 10^{-7}$ and $n = 200$ in (5.39) we obtained the optimal parameters $\theta^* = \{\alpha = 112, \kappa = 0.00052, R = 1.54\}$. The correlation between the accepted values of α and κ in the Metropolis-Hastings algorithm was found to be -0.16. In this study the modelling procedure appeared to work

to a certain extent. However, we are not entirely convinced that α had converged after the 8000 proposal steps (and hence are not very confident in these particular preliminary results). Several further experiments with different proposal and prior distributions were carried out in order to fine-tune this sophisticated procedure to yield results that we have much more confidence in. The study was repeated but this time using independent Gaussian priors with mean 100 and variance 20 for parameter α and mean 0.004 with variance 10 for parameter κ. Furthermore, the proposal distributions

$$\begin{aligned} \alpha_{t+1} &\sim |N(\alpha_t, 20)| \\ \kappa_{t+1} &\sim |N(\kappa_t, 5 \times 10^{-8})|, \end{aligned} \tag{5.42}$$

were chosen in such a way that a desirable parameter acceptance rate was obtained.

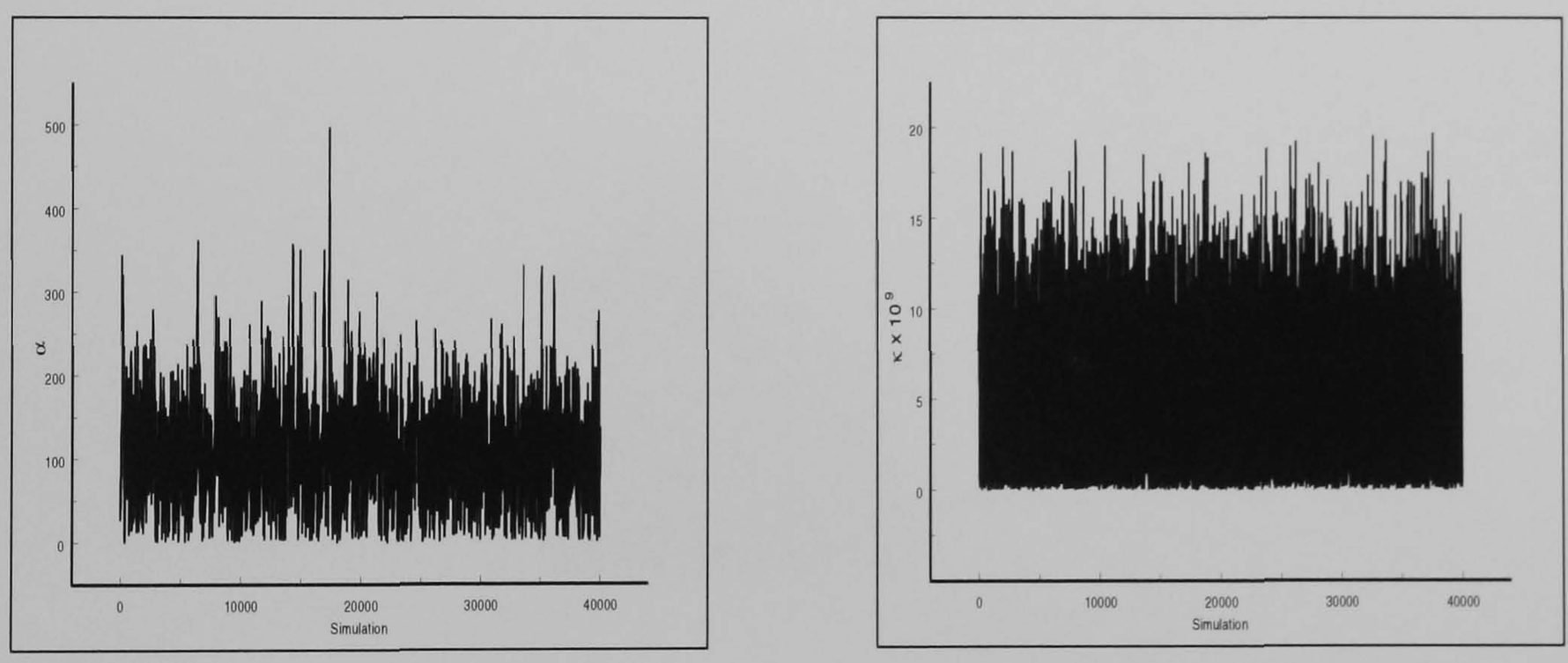


Figure 5.15: The graph on the left shows the value of α at a particular simulation (the mth transition in the Metropolis-Hastings procedure) and similarly for the parameter κ on the right.

Following 40000 steps, the accepted values for both parameters appeared to have converged in distribution (see Figures 5.15-5.17) and we obtained the estimates $\hat{R} = 1.54$, $\hat{\kappa} = 5.3 \times 10^{-9}$ and $\alpha = 108.9$ with starting parameters $\alpha = 1$ and $\hat{\kappa} = 1 \times 10^{-9}$. This demonstrates that if we wish to adopt this procedure then a good prior specification for the parameters and fine-tuning of the various parameters in the algorithm is paramount.

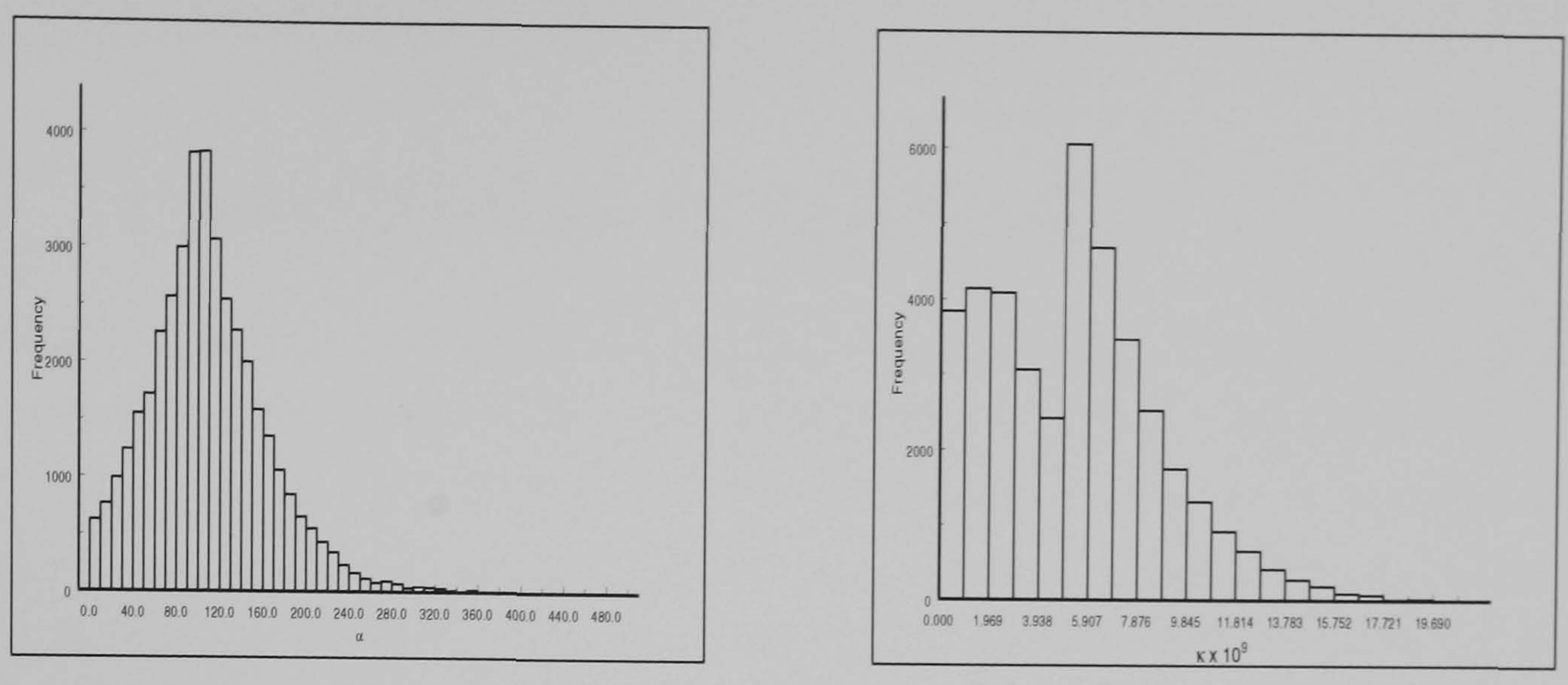


Figure 5.16: Histograms of the values plotted in Figure 5.15.

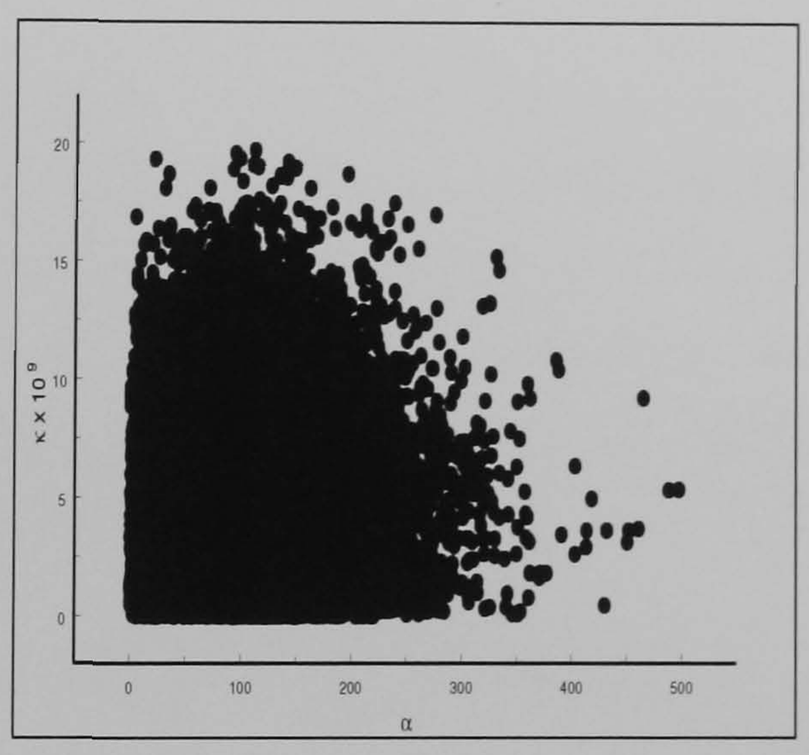


Figure 5.17: A scatter plot of the values plotted in Figure 5.15. The correlation between the accepted values of α and κ was found to being -0.0077

Berthelsen & Møller (2006) have introduced a technique for simulating from a posterior density when the likelihood is given by an unnormalised density, with an unknown partition function. This avoids the need for estimating the ratio of normalising constants, while at the same time making use of perfect simulation. However, a critical issue with the algorithm is the requirement for specifying an auxiliary distribution that should approximate the likelihood. This is of course a similar problem to specifying a θ_0 as in (5.34). We therefore do not have a robust method for

fitting Model 3 to PPDS2 that does not require reasonable prior knowledge of the appropriate parameters.

5.5 A log-Gaussian Cox process

The final model that we consider comes from a flexible class of models that have been used extensively in spatial point process modelling and have well-known theoretical properties. Suppose that we have a Poisson process with intensity measure Λ. If Λ is a realisation of a random measure, M (a function of a random set, taking finite values on bounded sets), so that the conditional distribution of X given $M = \Lambda$ follows a Poisson process with intensity measure Λ, then X is said to be a Cox process driven by M. As stated by Møller & Waagepetersen (2002), Cox processes are like inhomogeneous Poisson processes for aggregated point patterns. Usually in applications M is unobserved, and so we cannot distinguish a Cox Process X from its corresponding Poisson process $X|M$ when only one realisation of $X \cap W$ is available.

Although we stated informally that a Cox process may be a suitable extension of Model 2 for modeling the spatial distribution of PML bodies, Møller & Waagepetersen (2001) formally discuss what should be considered when deciding which of these two models are appropriate. That is, whether M should be random or deterministic. In many specific Cox processes, such as the one we are about to define, M is specified by a positive spatial process $Z = \{Z(\eta) : \eta \in \mathbb{R}^d\}$, so that

$$M(B) = \int_B Z(\eta) d\nu(\eta).$$

We then say that X is driven by the random intensity surface Z. In particular, X is said to be a log-Gaussian Cox process (LGCP) if any linear combination

$$\sum_i a_i \log(Z(\eta_i))$$

follows a normal distribution. That is, for $Y(\eta) = \log(Z(\eta))$, the joint distribution of any finite vector $(Y(\eta_1), \ldots, Y(\eta_n))$ is Gaussian. Furthermore, under the assumption of stationarity and isotropy, the distribution of Y is specified by the mean, $m(\eta)$ and covariance function $c(r) = \mathrm{Cov}(Y(\eta_1), Y(\eta_2))$. Where $r = \|\eta_1 - \eta_2\|$. The endless possibilities of a covariance function makes LGCPs highly flexible models. A popular covariance function (see for example Moller et al, 1998, Benes et al, 2002, and Diggle, 2003) that yields a tractable model, is the exponential covariance

function

$$c(\eta, \xi) = \sigma^2 \exp\left(-\|(\eta - \xi)/\kappa\|\right) \tag{5.43}$$

where $\kappa > 0$, $\sigma^2 > 0$. Other covariance functions can be found in Moller et al (1998). We note that the intensity and pair correlation function completely determine the distribution of the process. Moreover, the intensity is given by

$$\rho(\eta) = \exp(m + \sigma^2/2). \tag{5.44}$$

Furthermore, the pair correlation function is given by

$$\begin{aligned} g(\eta, \xi) &= \exp(c(\eta, \xi)) \\ &= \exp(\sigma^2 \exp(-\|(\eta - \xi)/\kappa\|)) \end{aligned} \tag{5.45}$$

(Moller & Waagepetersen, 2003). We can rewrite this as

$$g(t) = \exp\left(\sigma^2 \exp\left(-\frac{t}{\kappa}\right)\right)$$

Moreover, using the relationship between the $g(r)$ and $K(r)$-functions, we can write down the following expression for the $K(r)$-function

$$K(r) = 4\pi r^2 \int_0^r t^2 \exp\left(\sigma^2 \exp\left(-\frac{t}{\kappa}\right)\right) dt \tag{5.46}$$

We shall attempt to fit a LGCP with constant mean $m(\eta) = E(Y(\eta)) = m$ and an exponential covariance function to PPDS2. Benes et al (2002) discuss a discretisation approach for a likelihood function for fitting a Cox process in the two-dimensional case. The procedure appears highly computationally intensive and non-trivial to implement for the case where the observation region (the cell nuclei) boundary is represented as a convex hull. A fast and relatively straightforward to implement alternative for fitting Cox processes is the widely used minimum contrast approach (see for example Møller et al, 1998, Møller & Waagepetersen, 2004, and Guan, 2007). We adopted the minimum contrast approach to individually fit the LGCP to the cell nuclei in PPDS2. Specifically, to determine κ and σ we optimised the expression

$$\int_r ((K(r))^p - (\hat{K}(r))^p)^2 dr. \tag{5.47}$$

via a grid search with $p = 1$. Since the log-Gaussian Cox process is stationary, we can estimate the intensity in the usual way, to obtain

$$\hat{\rho} = \frac{x(B)}{\nu(B)} = \exp(m + \sigma^2/2) \tag{5.48}$$

Rearranging (5.48) for m and replacing σ with an estimate $\hat{\sigma}$, we can obtain the estimate

$$\hat{m} = \ln(\hat{\rho}) - \hat{\sigma}^2/2 \tag{5.49}$$

for m. We could not evaluate the integral in (5.46) explicitly, and hence adopted numerical integration (the trapezium rule). The $K(r)$-function in (5.47) was estimated using a proportion of volume weighting edge-correction (see Chapter 4). It is important to adopt an edge-correction here to reduce estimation bias. This is because a theoretical value of a characteristic is being compared to an estimated value of the characteristic. Using the model-fitting procedure that has just be outlined, we obtained the parameters $\kappa = \epsilon$, $\sigma = \epsilon$ and $m = -11.1$ for $\epsilon << 1$ for each of the 8 cell nuclei of PPDS2. These results should not be surprising since Brix (1999) stated that for a particular class of Cox processes, minimum contrast estimation based on the $K(r)$-function can be numerically unstable and yield parameter estimates that are on the boundary of the parameter space when there is not sufficient data to estimate the $K(r)$-function well. Brix (1999) reported that the same problem did not seem to occur when using the pair correlation function. We are faced with the limitation that the pair correlation function is difficult to estimate for this dataset. Furthermore, our investigation of edge-correction in three-dimensions in Chapter 4 showed that the edge-correction adopted, particularly in this case is only valid for small distance. A similar observation is made by Diggle (2003, p50). We prefer the border method in this particular case, however, we are restricted by the low PML count (see Chapter 4). As a suggestion, a possible way round these problems is to consider various other summary statistics, for which we may not have the closed forms for but we can approximate them from simulated data. Hence we are left to minimise the expression

$$\int_{\theta} ((\tilde{S}(\theta))^p - (\hat{S}(\theta))^p)^2 d\theta. \tag{5.50}$$

where $\hat{S}$ and $\tilde{S}$ are some arbitrary, but descriptive summary statistic estimated from the data and simulated data respectively. However, simulating a LGCP inside an irregular three-dimensional body is non trivial. Although we are able to simulate a LGCP inside a brick, we are not able to prove that our approximation for simulating inside a convex hull is sufficiently accurate.

More precisely, to simulate a LGCP inside a brick, we use the approach described by Moller & Waagepetersen (2001). That is, we simulate a Gaussian vector $\tilde{Y}_i =$

$Y(c_i)$, where we divide W into indexed disjoint cells C_i. with index set I. So that $W = \bigcup_{i \in I} C_i$ and $c_i \in C_i$ is the centre of C_i. We form

$$\tilde{Y}_i = \Gamma Q + \mu^{(k)},$$

where $\mu = (m, \ldots, m)^T$ is a k-dimensional vector of means. Also, Γ a vector of standard normals and Q the lower triangle Cholesky decomposition of the covariance matrix. We then simulate independent Poisson distributions X_i with intensity $\exp(\tilde{Y}(c_i))$ inside the C_i. To obtain a simulation of the process inside a cell nuclei (for which the boundary is described by a convex hull) we use a rejection algorithm. That is, we embed the nuclei boundary W inside a brick $W_1 \supset W$ with dimensions determined by points on the convex hull. We then simulate a LGCP inside W_1 and only accept events that lie inside the convex hull.

We tested this model-fitting process by simulating a LGCP with parameters $m = 0.0045, \kappa = 1.2$ and $\sigma = 0.3$ inside the brick with edges determined by $(x_{min}, x_{max}) = (0, 2), (y_{min}, y_{max}) = (0, 4), (z_{min}, z_{max}) = (0, 6)$. Hence we have that $E(X(B)) \approx 50$ since for B a brick with vertices at $(x_{min}, y_{min}, z_{min})$, $(x_{max}, y_{max}, z_{max})$ and so forth we have

$$\begin{aligned} E(X(B)) &= \int_B \rho(\eta)\nu(d\eta) \\ &= \int_{z_{min}}^{z_{max}} \int_{y_{min}}^{y_{max}} \int_{x_{min}}^{x_{max}} \exp(m + \sigma^2/2) d\eta^{(1)} d\eta^{(2)} d\eta^{(3)} \\ &= (z_{max} - z_{min})(y_{max} - y_{min})(x_{max} - x_{min}) \exp(m + \sigma^2/2) \end{aligned} \tag{5.51}$$

On attempting to fit parameters to the simulated data using the same minimum contrast approach as done for the cell nuclei of PPDS2, we obtained the estimates $m = 0.06, \kappa = 1 \times 10^{-7}$ and $\sigma = 1 \times 10^{-7}$. These values are not acceptably 'close' to the true ones. Indeed it appears that the model fitting approach is not picking up the covariance structure of the model. However, again this is not entirely surprising, given our discussions earlier, which suggested that the minimum contrast approach has been reported to being numerically unstable. A simulation study revealed that increasing m in the LGCP model (gradually from 0.0045 to 0.9) continued to yield values of κ and σ that were close to the boundary of the parameter space. Figure 5.18 was used as a visual confirmation that the issue of obtaining parameter estimates that are not 'close' to the true values was not a result of $K(r)$-function estimation error. To obtain the curves in Figure 5.18 we estimated the $K(r)$-function for the simulated data using the same Monte Carlo integration method for the edge-correction term, as done for the PPDS2 cells and compared this with the theoretical

value. We obtained the simulation envelopes in Figure 5.18 from 100 simulated realisations of the LGCP with parameters $m = 0.0045, \kappa = 1.2$ and $\sigma = 0.3$.

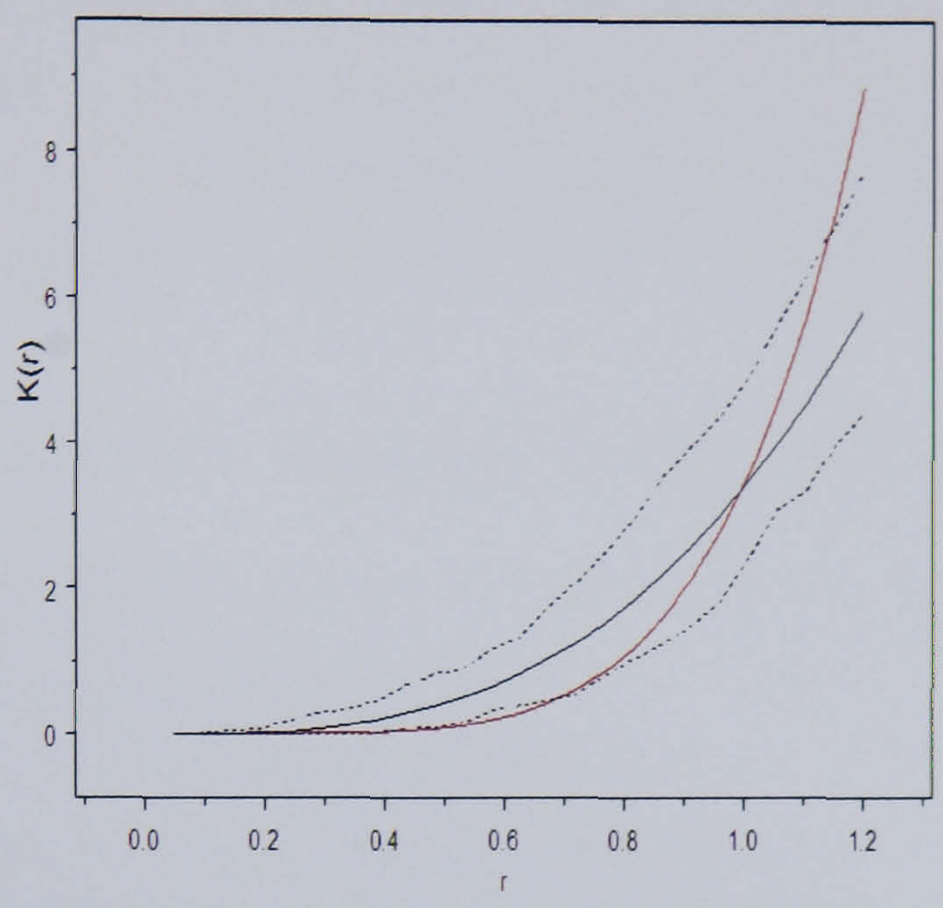


Figure 5.18: Simulation envelopes for the $K(r)$-function of 100 simulated LGCPs with parameters $m = 0.0045, \kappa = 1.2$ and $\sigma = 0.3$ (dashed curves) with the average of the simulations (black curve) and the theoretical curve (red).

5.6 Model selection

At the beginning of this chapter we carried out experiments which suggested that the replicates that make up PPDS2 are similar in terms of PML body spatial distribution. This was found to be consistent with the cell nuclei being biologically similar in terms of being at the same stage of the cell cycle. We have used the idea of cell similarity as a premise for fitting some models to PPDS2 in a replicated data framework. We have discussed four possible models for PPDS2. These being

- Model 1, which is an homogeneous Poisson process with nucleoli exclusion zones.
- Model 2 is an extension of model 1 and incorporates boundary effects. We showed that Model 2 is an inhomogeneous Poisson process.
- Model 3 incorporates the idea of PML body interaction and has a similar intensity surface to Model 2.

- Model 4 is a LGCP with an exponential covariance function.

We saw through a rigorous diagnostics process that Models 1 and 2 appear suitable for some of the cell nuclei but not for others. When discussing the idea of comparing model likelihoods, we implicitly touched upon the idea of model selection for the case when one has two or more competing models. Model selection is non-trivial, we will briefly discuss the use of pattern recognition and partial Bayes factors in choosing between competing spatial point process models.

5.6.1 Model selection through pattern recognition methods

Raghavan et al (1998) have proposed a pattern recognition based scheme to classify a spatial point pattern. The results of experiments that they carried out suggested that a supervised pattern recognition scheme is successful at classifying spatial point patterns. Furthermore, as expected, the misclassification error is increased as the difference between the processes increases. Dučinskas, K. and Šaltytė (2001) have used Quadratic Discriminant Analysis (QDA) (see, for example, Webb, 1999) to classify a stationary Gaussian random field, for the one-dimensional case. Experimental evaluation of Raghavan et al's (1998) method suggested that this method was unsatisfactory for application to PPDS2. The key issue with applying the method was that it appeared to yield an unacceptable misclassification probability for various spatial point patterns with a low number of events (similar to PML body counts).

5.6.2 Model selection via partial Bayes factors

Walsh & Raftery (2005) have discussed the idea of using partial Bayes factor for choosing between competing spatial point processes. The technique uses quantities similar to those in the computer experiments approach given by Raghavan et al (1998). The partial Bayes factor approach outlined by Walsh & Raftery (2005) involves calculating the Bayes factor for competing hypotheses H_1 against H_0, which is the ratio of the posterior to the prior odds for H_1 against H_0. The Bayes factor is the ratio of integrated likelihoods

$$BF = \frac{\int P(x|\theta_1, H_1)\pi_1(\theta_1|H_1)d\theta_1}{\int P(x|\theta_0, H_0)\pi_2(\theta_0|H_0)d\theta_0}$$

where θ_i is a vector of parameters associated with hypothesis H_i and $\pi_i(\theta_i|H_i)$ denotes the prior distribution. However, BF may be difficult to compute. As a result, Walsh & Raftery (2005) suggest the use of a partial Bayes factor, which, is defined

as the ratio of integrated likelihoods for a summary statistic, or a vector of several summary statistics, S, rather than for the complete data. That is

$$\begin{aligned} PBF &= \frac{\int P(S|\theta_1, H_1)\pi_1(\theta_1|H_1)d\theta_1}{\int P(S|\theta_0, H_0)\pi_2(\theta_0|H_0)d\theta_0} \\ &= \frac{I_1(S)}{I_0(S)}. \end{aligned} \tag{5.52}$$

The Monte Carlo estimate for PBF is

$$\hat{I}_i(S) = \frac{1}{K}\sum_{j=1}^{K} \widehat{P}(S|\theta_{(i)}^{(j)}, H_i) \tag{5.53}$$

To estimate $\widehat{P}(S|\theta_i^j, H_i)$, we simulate say $m = 100$ events from H_i with parameters $\theta_i^{(j)}$, and calculate the summary statistics $S_i^{(j)}$. We then use a kernel density estimator to obtain $\widehat{P}(S|\theta_{(i)}^j, H_i)$. The issue of bandwidth selection for the kernel density estimation, described in section 5.1 re-emerges in this instance. In our implementation we adopted a method suggested by Turlach (1993). Figure 5.19 shows the results of a pilot study that involved simulating 1000 times from Models 1 and 2 and using the average nearest event distance and average event to boundary distance as S in (5.53), to assess how well the PBF approach distinguishes between Models 1 and 2. The results in Figure 5.19 appear to be promising.

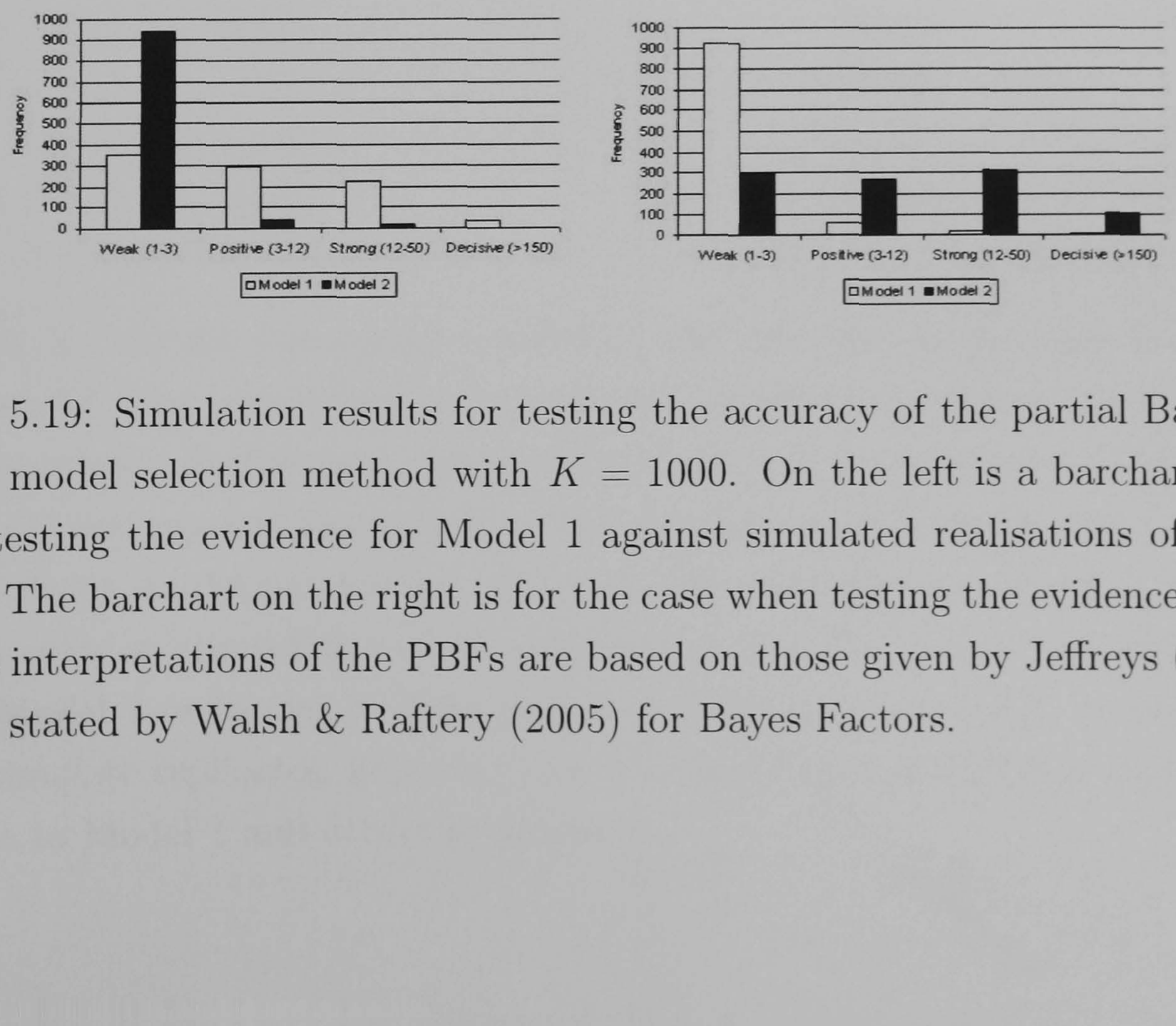


Figure 5.19: Simulation results for testing the accuracy of the partial Bayes factor (PBF) model selection method with $K = 1000$. On the left is a barchart of PBFs when testing the evidence for Model 1 against simulated realisations of Models 1 and 2. The barchart on the right is for the case when testing the evidence for Model 2. The interpretations of the PBFs are based on those given by Jeffreys (1961) and as also stated by Walsh & Raftery (2005) for Bayes Factors.

In section 5.2.3 we found that, for cells 1-5 of PPDS2, there was no evidence to reject the hypothesis that Models 1 and 2 are appropriate for describing the spatial distribution of PML NBs. We adopted the partial Bayes factors approach for assessing which of Model 1 or Model 2 best describes the spatial distribution of PML NBs in cells 1-5 of PPDS2 in terms of inter-PML NB distance mean and the mean PML body to nuclear boundary distance. We consider these particular quantities as they are consistent with some of the PML NB distribution features, that we are interested in capturing. In this experiment, we set $m = 1000$, with kernel density estimation as described earlier. In summary, using the guide proposed by Jeffreys (1961) for interpreting Bayes factors and the results of the simulation studies for how well the PBF method works for data of type PPDS2, we have classified the level of evidence in favour of Model 1 over Model 2. The results are provided in Table 5.7. They suggest that for cells 1, 2, 5 and 6 there is more evidence to favour Model 2 (boundary effects model) over Model 1 (no boundary effects model). While for cells 3, 4, 7, and 8 there is greater evidence for Model 1 as being more appropriate for describing the spatial distribution of the PML compared to Model 2. The results

Table 5.7: Model selection using partial Bayes factors

Cell	PBF For H_1 : Model 1	Evidence for Model 1 compared to Model 2
1	0.27	Weak
2	0.75	Weak
3	5.55	Positive
4	27.18	Strong
5	0.63	Weak
6	1.18	Weak
7	12.04	Positive
8	19.42	Strong

in Table 5.7 do not necessarily contradict the outcome of our tests for similarity between the data (see section 5.1) as Models 1 and 2 have approximately the same $K(r)$-functions. Furthermore, given that Model 1 is a special case of Model 2, the results do not oppose the possibility that the data come from the same Model, but with differing model parameters. However, the information in Table 5.7 does raise some doubts on the validity of the assumption that the data are all realisations of either Model 1 or Model 2. Note that if we were to adopt a Cox process then we could simulate replicates, whereby some simulated realisations have similar spatial features to Model 1 and others to Model 2.

Chapter 6

Marked spatial point pattern analysis

Up until now we have focused on the analysis of the spatial distribution of PML NBs as a means of obtaining quantitative evidence for biological ideas regarding their function within the nucleus. We have applied *univariate* point process theory for comparing the spatial locations of PML NBs (nuclear bodies) with CSR and have attempted to fit a variety of spatial point process models to PPDS2. We have also seen from some of the biological literature cited in Chapters 1 and 5 that several studies have suggested that PML NBs may have a preference for certain subnuclear locations.

The notion of PML NB interaction with its surroundings is one of great importance. Lanctôt et al (2007) have reported that gene expression is mediated by interactions between chromatin and protein complexes. Dellaire & Bazzet-Jones (2004) have proposed that PML NBs are dynamic sensors of cellular stress, that associate with regions of DNA damage. Borden (2002) has reported that PML NBs tend to be near certain nuclear compartments such as Cajal/coiled bodies, cleavage bodies, and splicing speckles. Knowledge of the locations of PML NBs in relation to other structure may provide clues to PML NB function. Furthermore, its relative location may give insight into what specific targets it regulates (Borden, 2002). Following this, most reported strategies for assessing PML NB functions in essence are designed to answer the following questions: what nuclear structures are the bodies near to, what other macromolecules localize with the body, and what are the effects of disrupting the body (Borden, 2002)?

Recently, the Imperial College centre for structural biology has been able to provide images obtained via confocal microscopy, that yield data (such as PPDS3) that is similar to PPDS2 but at the same time, providing the spatial locations of other bodies including RNA Polymerase II (see Figure 6.1 as an example). Borden (2002) states that general transcription factors such as RNA Polymerase II do not colocalize with PML bodies. This is in line with the experimental results obtained by Xie & Pombo (2006) which supported the view that although PML NBs are present in transcriptionally active areas, they are not generally sites of polymerase II assembly. In this chapter, we carry out multivariate (or equivalently, *marked*) point

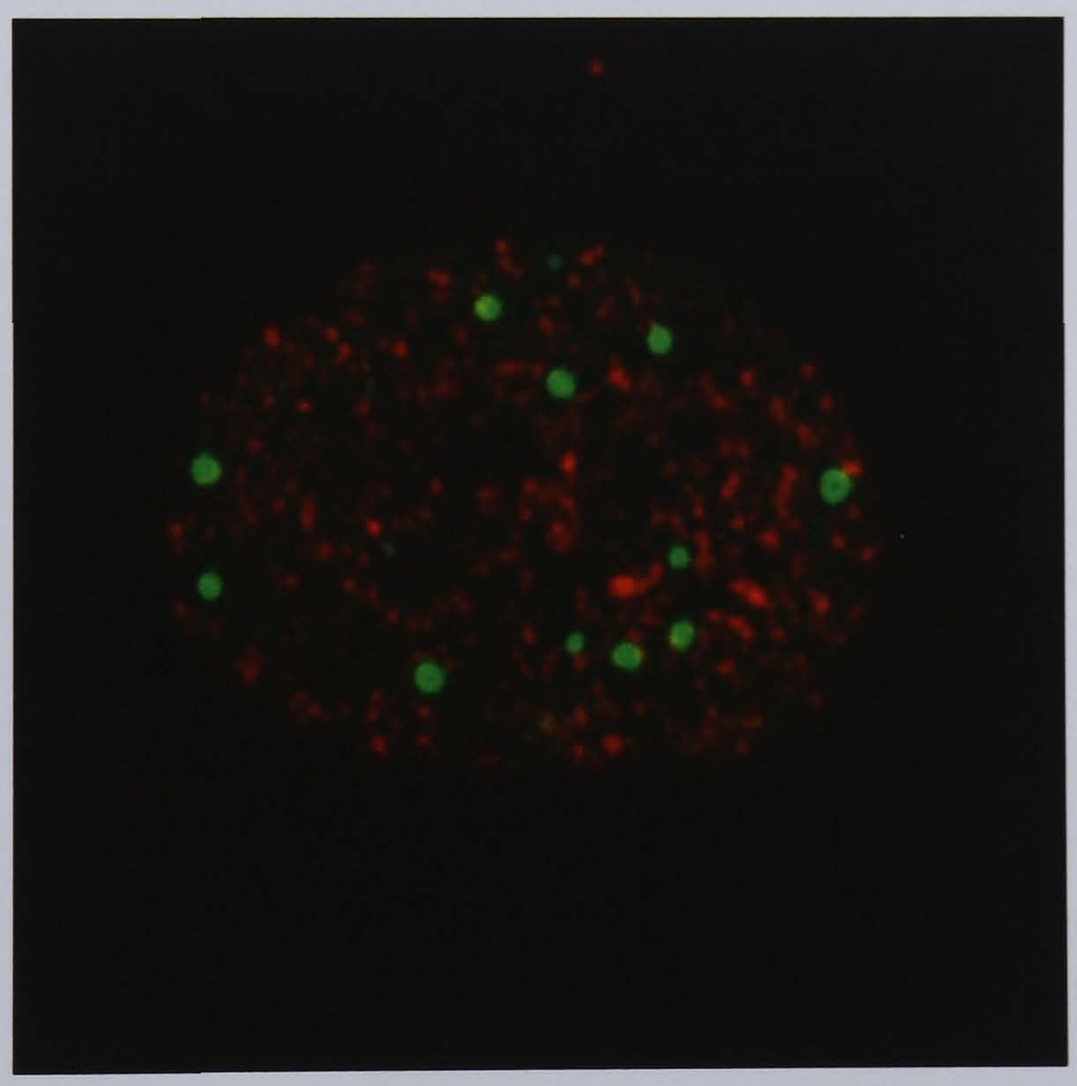


Figure 6.1: A two-dimensional projection of cell 4 of PPDS3 (see also Appendix B) produced by the Imperial College centre for structural biology. The bright green dots are PML NBs. The bright red dots are RNA Polymerase II

pattern analysis so that we can provide some statistical evidence for biological ideas regarding PML NBs association with other nuclear bodies. We also explore the relationship between PML NB 'size' with its positioning in the nuclear interior. We begin with a background to point process theory for multiple event types, and end with a discussion on how the models from the previous chapter can be extended to models for *marked* spatial point process.

6.1 Some preliminaries for marked spatial point processes

6.1.1 Definitions

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A *marked spatial point process* $X^{[m]}$ on $S \subset \mathbb{R}^d$ is a point process on the product space $S \times \mathbb{M}$, that is, a measurable mapping of $(\Omega, \mathcal{F})$ into $(\Phi, \mathcal{N})$. Here $\mathbb{M}$ is the mark space, assumed to be a Polish space (a separable complete metrisable topological space) and Φ is the set of locally finite counting measures on $S \times \mathbb{M}$, so that for all B in Borel sets of S and for all C in Borel sets of $\mathbb{M}$ and $x(B \times C)$ is the number of events in B whose marks belong to C. $\mathcal{N}$ is the smallest σ-algebra generated by sets of the form $\{x \in \Phi : x(B \times C) = n\}$, for all B in Borel sets of S and for all C in Borel sets of $\mathbb{M}$ and all $n \in \mathbb{N}$. The mark spaces that we are interested in include $\mathbb{M} = \{0, 1\}$ and $\mathbb{M} = \mathbb{R}^+$, although some of the theory provided applies to the more general case where $\mathbb{M} = \mathbb{R}^d$.

A marked spatial point process defined over $(\Omega, \mathcal{F}, P)$ induces a probability measure, $P_{X^{[m]}}(Y) = P(X^{[m]} \in Y)$, on $(\Phi, \mathcal{N})$ for all $Y \in \mathcal{N}$. As done previously, we shall drop the subscript $X^{[m]}$ in the probability measure $P_{X^{[m]}}(Y)$ unless clarity is required. As stated by Stoyan et al (1995), it is always possible to interpret a marked point process as an ordinary point process in $\mathbb{R}^d \times \mathbb{M}$.

The *mark distribution*, $Z_{X^{[m]}}(t)$ of a (real-valued marks) marked spatial point process $X^{[m]}$ is the probability that the mark of a 'typical' point is less than t (Stoyan & Stoyan, 1994). That is, $Z_{X^{[m]}}(t)$ is a probability measure on $\mathbb{M} \times \mathcal{M}$ for $\mathcal{M}$, the Borel σ-algebra of $\mathbb{M}$. The mark distribution can be thought of as a Palm distribution (see Stoyan et al (1995), section 4.4 for further details).

A multivariate spatial point pattern is a special case of a marked spatial point pattern, where there is a finite number of marks, each representing an event-type (Cressie, 1991). Hence, for example in the case where we set $\mathbb{M} = \{0\}$ we have a univariate spatial point process, while by setting $\mathbb{M} = \{0, 1\}$, we are implicitly considering a bivariate spatial point process. More formally, a *bivariate spatial point process*, $X^{[2]} = (X_1, X_2)^T$ on $S \subset \mathbb{R}^d$ is a measurable mapping of $(\Omega, \mathcal{F})$ into $(\Phi, \mathcal{N})$, where here Φ is a set of locally finite bivariate counting measures and $\mathcal{N}$ is the smallest σ-algebra generated by sets of the form $\{x \in \Phi : (x^{(1)}(B_1), x^{(2)}(B_2)) = (n_1, n_2)\}$ for all Borel sets B_1, B_2 and all $n_1, n_2 \in \mathbb{N} \cup \{0\}$.

A bivariate spatial point process may be used to model, for example, the locations of two different types of subnuclear bodies in the nucleus, while a marked spatial point process may be used to model the volume of one type of subnuclear body in the nucleus. In this chapter we use a marked spatial point pattern to model the dependency of PML NB location on PML size.

Similarly to the univariate case, we make use of the Palm distribution. Let m_ξ denote the event type of an event at point ξ, Λ_i the intensity measure of the spatial point process of type i, and define the *reduced Campbell measure*, $\mathcal{C}_i$ with respect to X_i by

$$\mathcal{C}_i(B \times Y) \equiv \int_\Phi \int_B I_{(x-\xi:m_\xi=i\})\in Y x^{(i)}(d\eta) P_{X^{[2]}}(dx)$$

For Borel set B and $Y \in \mathcal{N}$. It follows from the justifications provided by Cressie (1991, section 8.6.1) that there exists uniquely determined measures on $(\Phi, \mathcal{N})$ such that

$$\mathcal{C}_i(B \times Y) = \int_B P^!_{X_i,\xi}(Y)\Lambda_i(d\xi).$$

$P^!_{X_i,\xi}$ is the reduced Palm distribution of X_i with respect to ξ. For a stationary bivariate spatial point process $P^!_{X_i,\xi} \equiv P^!_{X_i,O}$, since the reduced Palm distribution does not depend on ξ.

6.1.2 Stationarity and isotropy

Similar to as mentioned by Stoyan (2006), Euclidean motions of marked point processes are defined as transforms that move the events but the marks are unchanged. Hence, if analogous to the univariate case, we have the marked spatial point process $X^{[m]} = \{[\eta_1, m_1], [\eta_2, m_2], \ldots, [\eta_u, m_u]\}$ with $\mathbb{M} = \{m_1, \ldots\}$, then the translation of $X^{[m]}_\xi = X^{[m]} + \xi$, is given by $X^{[m]}_\xi = \{[\eta_1 + \xi, m_1], [\eta_2 + \xi, m_2], \ldots, [\eta_u + \xi, m_u]\}$. Furthermore, rotations act on marked spatial point processes by rotating the points but not altering the marks. A marked spatial point process $X^{[m]}$ is said to be *stationary* if for all $\xi \in \mathbb{R}^d$, the translated process $X^{[m]}_\xi$ has the same distribution as $X^{[m]}$. That is $P_{X^{[m]}_\xi}(Y) = P_{X^{[m]}}(Y)$ for all $Y \in \mathcal{N}$. An analogous definition of an *isotropic marked spatial point process* follows from replacing the translation operation to a rotation about the origin.

6.1.3 Moment measures and summary statistics for marked spatial point patterns

It is generally not difficult to extend the definitions of the moment measures from Chapter 2 and most of the summary statistics from Chapters 2 and 3 to the marked spatial point pattern case.

The intensity (or mean) measure, Λ, for marked spatial point process $X^{[m]}$ is defined by

$$\Lambda(B \times C) = E(X^{[m]}(B \times C)) = \int x(B \times C)P(dx)$$

for Borel sets B and C.

When $X^{[m]}$ is stationary with intensity λ, we have (Stoyan, 2006) that the intensity measure of $X^{[m]}$ is given by

$$\Lambda = \lambda\nu(B)Z_{X^{[m]}}$$

The second-order factorial moment measure, $\alpha_m^{(2)}(.)$, of the marked spatial point process $X^{[m]}$ is given by

$$\alpha_m^{(2)}(B_1 \times B_2 \times C_1 \times C_2) = \int \sum_{[\eta_1;m_1],[\eta_2;m_2]} I_{(\eta_1 \in B_1)} I_{(\eta_2 \in B_2)} I_{(m_1 \in C_1)} I_{(m_2 \in C_2)} P(dx)$$

For a stationary and isotropic bivariate spatial point process $X^{[2]} = (X_1, X_2)^T$ with intensity $(\lambda_1, \lambda_2)^T$, the reduced second moment measure of X_1 and X_2 is given by

$$\lambda_j \mathcal{K}_{ij}(B) = \int_\Phi x^{(i)}(B) P^!_{X_j,0}(dx^{(i)}) \quad i,j = 1,2$$

for all B in the Borel σ-algebra of $S \subset \mathbb{R}^d$. The bivariate version of the $K(r)$-function, $K_{ij}(r)$ of a stationary marked point process was first introduced by Hanisch & Stoyan (1979) and has recently been adopted by Diggle et al (2006) for modeling a bivariate spatial point pattern. Heuristically, letting λ_k denote the intensity of events of type X_k, $K_{(ij)}(r)$ is the expected number of events of type j that are within distance r of an event of type i. Formally, if $B = b(o,r)$ is the closed ball of radius r centred at the origin, then

$$K_{ij}(r) = \mathcal{K}_{ij}(b(o,r)), \quad r \geq 0.$$

For n_1 the number of type 1 events, and n_2 the number of type 2 events, the $K_{ij}(r)$-function can be estimated inside the bounded window $W \subset \mathbb{R}^d$ using the estimator

$$\widehat{K}_{ij}(r) = \frac{\sum_{\eta \in X_1} \sum_{\xi \in X_2} \omega(\eta, \xi)^{-1} I_{(\|\eta - \xi\| \leq r)}}{\nu(W)\widehat{\lambda}_1 \widehat{\lambda}_2} \tag{6.1}$$

as suggested by Hanisch & Stoyan (1979). The estimate for the intensity parameter, λ_i is given by $\lambda_i = n_i/\nu(W)$. The function ω is an edge-correction factor such as the proportion of the surface area of the ball centred at η, passing through ξ. For relative ease of calculation and efficiency, we prefer the edge-correction $\omega(\eta, \xi) = \nu(W \cap b(\eta, \|\xi - \eta\|))$ (see Chapter 4 on edge-correction). To adopt this preferred form of edge-correction, we utilise the quadrature approximation

$$\nu(W \cap b(\eta, \|\xi - \eta\|)) \approx \frac{\nu(W)}{U} \sum_{i=1}^{U} I_{(\eta_i \in W \cap b(\eta, \|\xi - \eta\|))}$$

6.2 Bivariate spatial point process analysis of PPDS3

We estimated the $K_{12}(r)$-function using (6.1) and the preferred form of edge-correction for cell 4 of PPDS3. The data used to produce Figure 6.2 suggest that, heuristically, for this particular cell, we would expect to observe fewer RNA Polymerase II bodies within a distance of 0.5 units of the typical PML, compared to if the RNA Polymerase II exhibited CSR. From the biological literature, it appears that the inhibition (and perhaps more generally, spatial relationship between RNA Polymerase II and PML NBs) is driven by the biological function of PML NBs within different cell nuclei (see for example the results obtained by Mikecz et al, 2000). Also, it may be interesting to compare these results with the findings of Xie & Pombo (2006) who reported that PML bodies contain no detectable RNA polymerase II, but are often surrounded by them at a distance greater than 25 nanometers. From the detail provided in section B.1 of Appendix B, we estimate that 0.5 units is approximately 41 nanometers. Following the investigation on edge-correction provided in Chapter 4, we are cautious towards interpreting the results provided in Figure 6.2 for large values of r. We therefore also consider the outcome of simulation studies. Specifically, a plot of $K_{12}(r)$-functions with simulation envelopes, as shown in Figure E.18 (see Appendix E), can provide further insight. The $K_{12}(r)$-function simulation envelopes for cell nucleus k of PPDS3 was obtained by simulating 100 independent realisations of a homogeneous Poisson process, $\{\tilde{x}_{1k}, \ldots \tilde{x}_{100k}\}$, inside the convex

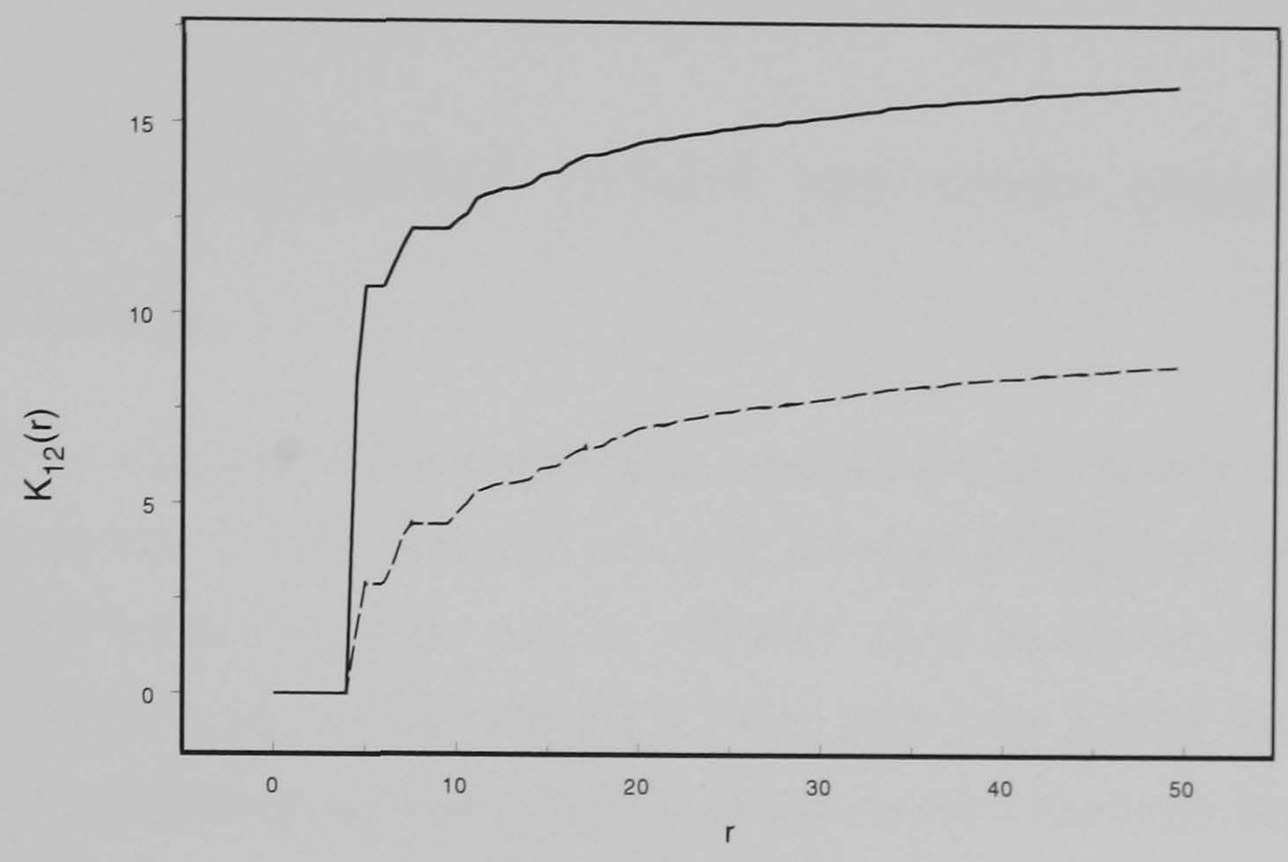


Figure 6.2: The $K_{12}(r)$-function plot for the PML (type 1) and RNA-Polymerase II (type 2) in cell nucleus 4 of PPDS3. The blocked curve is K_{12} and dashed curve is K_{21}

hull, representing the nuclear boundary of cell k. Each simulated realisation, $\tilde{x}_{jk}$, of the homogeneous Poisson process had a conditional number of events, n_k, where n_k is the number of RNA Polymerase II found inside cell nucleus k. For each $\tilde{x}_{jk}$ we estimate the $K_{12}(r)$-function without applying an edge-correction, where 1 is an event of type PML and 2 is an event belonging to $\tilde{x}_{jk}$. We hence obtain 100 estimates of the $K_{12}(r)$-function, $\{K_{12_{1k}}(r), \ldots, K_{12_{100k}}(r)\}$ for each cell k. The upper and lower simulation envelopes for each cell nucleus are respectively $\inf\{K_{12_{1k}}(r), \ldots, K_{12_{100k}}(r)\}$ and $\sup\{K_{12_{1k}}(r), \ldots, K_{12_{100k}}(r)\}$. The results presented in Figure E.18 suggest that, apart from cell 4, generally, for a wide range of r, there are fewer RNA Polymerase II bodies within distance r of the typical PML body compared to RNA Polymerase II bodies randomly scattered inside the nucleus. The results for cell 4 are consistent with those obtained for small r (relative to the nuclear magnitude), for the other cells in PPDS3. However, on the contrary to the other cells, for larger r, the PML in cell nucleus 4 typically tend to have a much greater number of RNA Polymerase II bodies within distance r, compared to RNA Polymerase II randomly placed inside the nucleus. The data obtained from the simulation studies (that were used to produce the charts in E.18) also confirmed that there are enough nuclear bodies in

cell 4 of PPDS3 for one to make legitimate observations for features of the spatial point pattern at distances of 0.5 units. This is owed to the high number of RNA Polymerase II bodies (there are 72 RNA Polymerase II bodies in cell 4 of PPDS3).

6.3 Marked spatial point process analysis of PPDS2

In Chapter 5 we discussed some possible spatial point process models for PPDS2. In the final sections of this chapter we will be discussing how one of these models can be extended to incorporate marks. Before this final step, we will now provide an exposition on how we 'mark' the PML NBs and gain initial insight into the mark distribution by analysing an appropriate spatial point process characteristic.

As mentioned above, the extension of other popular characteristics to the multivariate case is generally not difficult (for discrete marks). For the general marked case, the empty space function, $F(r)$ of the marked spatial point process $X^{[m]}$ is the cumulative distribution function of the distance from a randomly selected origin to the nearest event in $X^{[m]}$. That is

$$F(r) = P\left(X^{[m]} \bigcap (b(0,r) \times \mathbb{M}) \neq \emptyset\right)$$

Also, let B be a Borel subset of $\mathbb{M}$ with $Z_{X^{[m]}}(B) > 0$. We define the nearest neighbour function for events with marks in B by

$$G_B(r) = P^!_{X^{[m]},0}(X^{[m]} \bigcap (b(0,r) \times \mathbb{M}) \neq \emptyset)$$

for $r \geq 0$. van Lieshout (2004) has recently introduced a J-function for marked spatial point patterns. The J-function with respect to mark set B, J_B is given by

$$J_B(t) = \frac{1 - G_B(t)}{1 - F(t)}$$

for all $t \geq 0$ and $F(t) < 1$. For an independently marked Poisson process, $G_B(t) = F(t)$ for all t and so $J_B \equiv 1$. Values greater than 1 are a sign of inhibition, while values less than 1 are a sign of clustering.

van Lieshout (2004) proves the following for X a stationary point process on $\Re^d$ with intensity $0 < \lambda < \infty$. If X is *randomly labelled* with mark distribution Z on mark space $\mathbb{M}$ and if $X^{[m]}$ is the marked point process obtained, then for all

$r \geq 0$ with $F(r) < 1$, the J-function with respect to a Borel mark set $B \subset M$ with $Z_{X^{[m]}}(B) > 0$ is given by

$$J_B(r) = J_X(r)$$

where $J_X(r)$ is the J-function of X and where the marked spatial point process $X^{[m]}$ is said to have the *random labelling* property if the marks of the events are conditionally iid given the event locations.

For each PML NB in PPDS2, we calculated an approximate PML body length from the image data used to produce PPDS2. This was done by measuring the maximum distance between any two points, of the points that have been classified as being a part of that PML NB in the image processing stage. That is, in the data provided by the Imperial College London centre for structural biology, each PML NB j is described as a set of points $\{\eta_{1j}, \ldots \eta_{uj}\}$ (see Appendix B). The 'length' of PML NB j was calculated as $\inf\{\|\eta_{ij} - \eta_{sj}\| : i, s = 1, \ldots, u\}$. We use these lengths to assign marks to the PML NBs in PPDS2. The J-function plots for the cell nuclei of PPDS2 is shown in Figures 6.6-6.7. Figures 6.6-6.7 suggests that, since the marked and unmarked $J(r)$-functions are not too dissimilar, we would generally not necessarily expect to observe the PML NBs placed in the nuclear interior, in such a way that depends on their relative sizes (in terms of 'length'). Note also that we found that the proportions of the PML body length to nuclear length, denoted by z_π was consistent with the theoretical proportions provided in the biological literature (see Chapter 1 and Figure 6.3). All of the PML NBs in PPDS2 were pooled and we calculated the correlation between z_π and the proportion of PML NB distance to the boundary to nuclear length. We obtained a correlation of 0.09, suggesting that the two are not strongly correlated. The lack of correlation between the length of the PML and distance to boundary, provides some evidence for random labelling with respect to PML size. This is consistent with the results obtained using the marked J-function. Hence, these results would not support for example, a view that larger PML NBs are found closer to the nuclear periphery or more internally.

6.4 A marked inhomogeneous Poisson process model for PPDS2

Some of the models that we have encountered can be defined so that we are able to adopt them for the multivariate case. For general marked point processes, the marks

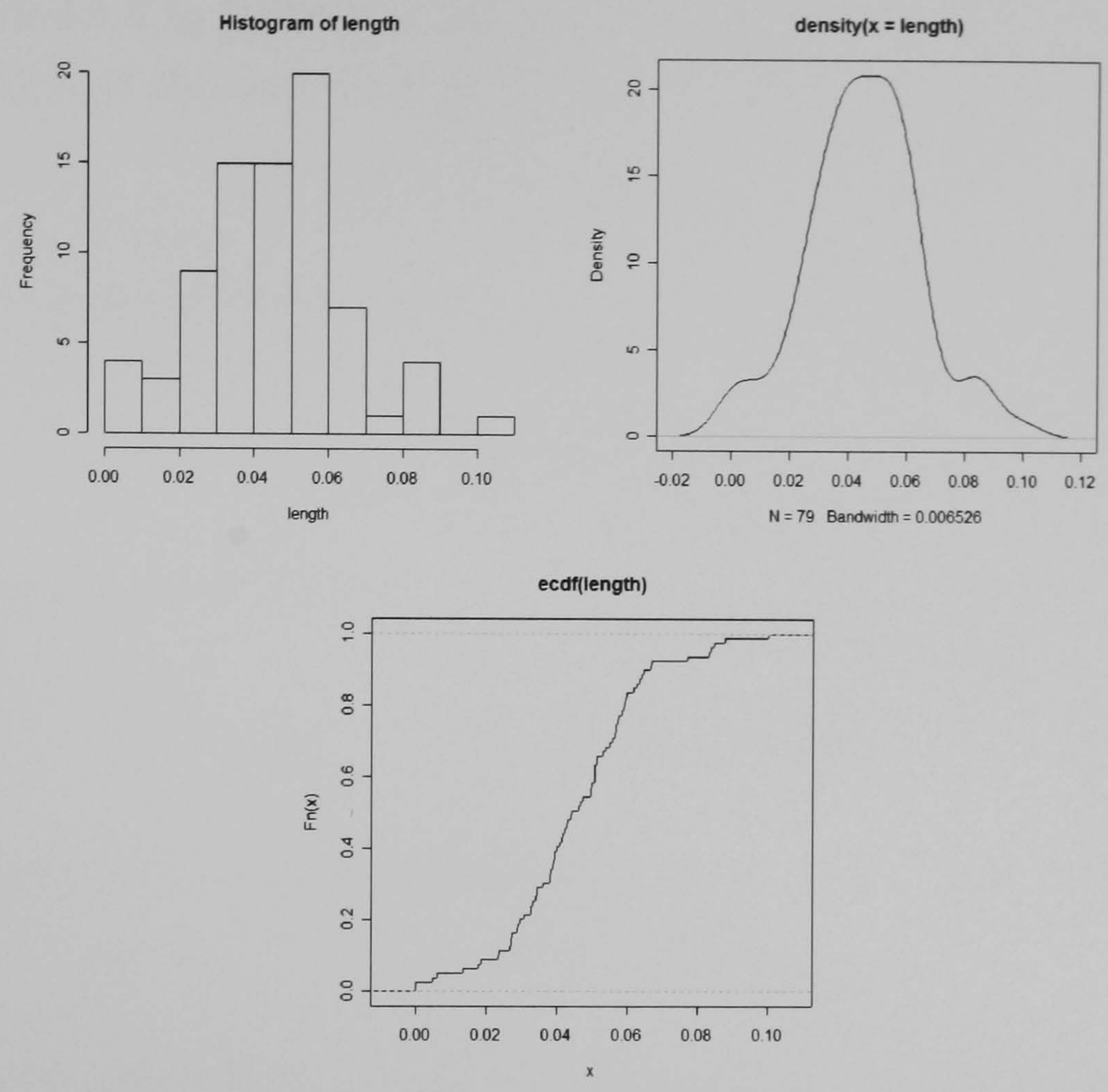


Figure 6.3: Histogram, density and CDF of z_π for all PML in PPDS2

can be modeled conditionally on the observed point pattern. Cressie (1991) states that the events of the point process can be viewed as locations of an irregularly spaced lattice and so the marks can be modeled using the theory of spatial models on lattices.

The final step in the model fitting discussions endeavours to identify each PML NB uniquely by assigning a mark to the PML NB. Each PML NB is now assigned a length. Consider a final model, Model 5, that is defined through the postulates of Model 2 (see Chapter 5, section 5.3), but not necessarily with the same intensity function (as we found that for some cells, this had perhaps been misspecified). Suppose that we now have the additional postulate:

- M5P1 Each PML η is randomly assigned a proportional length $z_\pi \sim Z$. That is, Z is a random variable that assigns to each PML NB, the mark

$$z_\pi = \frac{\text{PML NB length}}{\text{nuclear length}}$$

.

We can fit Model 5 to PPDS2 as follows. We first note that for a single replicate, the likelihood, l, of the data $\mathcal{D}$, is given by

$$l = p(\mathcal{D})p(\mathcal{M}|\mathcal{D})$$

For k iid replicates the likelihood l_{Rep} is given by

$$l_{\text{Rep}} = \prod_{j=q}^{k} p(\mathcal{D}_j)p(\mathcal{M}|\mathcal{D}_j)$$

and the log-likelihood is given by

$$l = \sum_{j=q}^{k} \log\left(p(\mathcal{D}_j)\right) + \log\left(p(\mathcal{M}|\mathcal{D}_j)\right)$$

We can therefore fit the marks separately to the models proposed in the previous chapter.

Formal tests on the data (see Figure 6.4) suggest that a normal distribution is a plausible model for the marks distribution Z. A KS test for the null hypothesis that the marks follow a normal distribution with mean 0.045 and standard deviation 0.019 provided a p-value of 0.92. Caution is required when choosing the mark

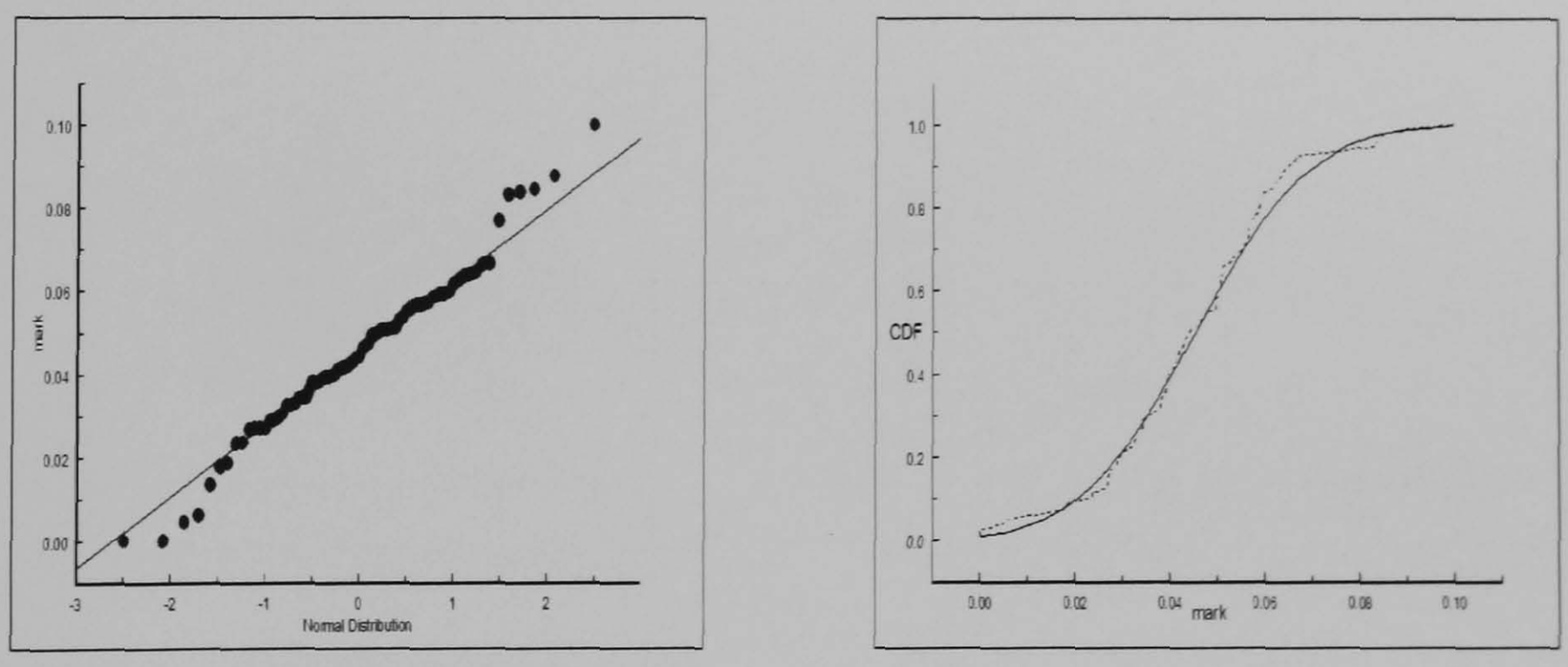


Figure 6.4: Q-Q plot of the PML NB marks (left). The points are approximately linear, suggesting possible normality. The graph on the right shows the empirical CDF of the PML NB marks (dashed curve) with the CDF of a normal distribution with mean 0.045 and standard deviation 0.019 (black line).

space since physical restrictions mean that, realistically, the mark space (that the

z_π belong to) is $A \subset (0,1)$ (and not for example $\mathbb{R}^+$ as implied by a normal distribution). This is because the PML NBs cannot be longer than the nucleus or have zero length. Hence, it is more appropriate to adopt a truncated normal distribution for Z. More precisely, Z has a normal distribution and lies within the interval (0,1). We estimated the mean and variance of the truncated (0,1) normal distribution, for the PML NB marks, to being (respectively) 0.045 and 0.019 (see for example Barr and Sherrill, 1999 for detail on the parameter estimation). The diagnostic plots

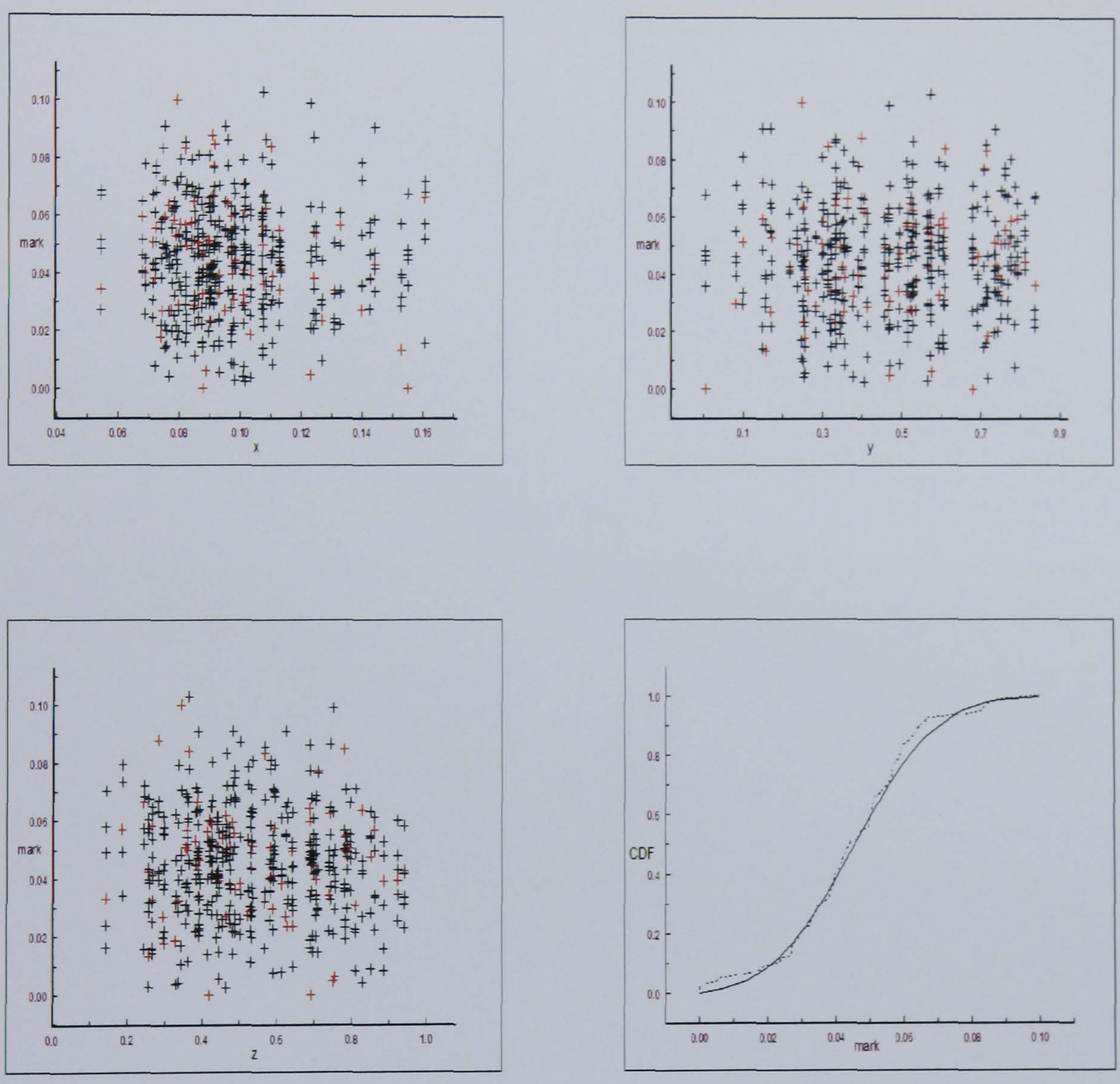


Figure 6.5: The graph on the bottom right shows the empirical CDF of the PML NB marks (dashed curve) with CDF of a truncated normal distribution with mean 0.045 and standard deviation 0.019 (black line). The scatter plots are of PML NB centroid x, y, z coordinates (divided by nuclear length) against simulated realisations of a truncated (0,1) normal distribution with mean 0.045 and standard deviation 0.019 (the modeled marks distribution). The red crosses represent the data and the black crosses are for the simulated marks.

presented in Figure 6.5 suggest that the truncated normal model that has been put forward for the PML NB marks distribution is a plausible one. Finally, by using this model for the marks distribution as Z in M5P1, and by letting M5P1 be an additional final postulate of Model 2, we obtain a marked spatial point process model for the spatial distribution of PML NBs.

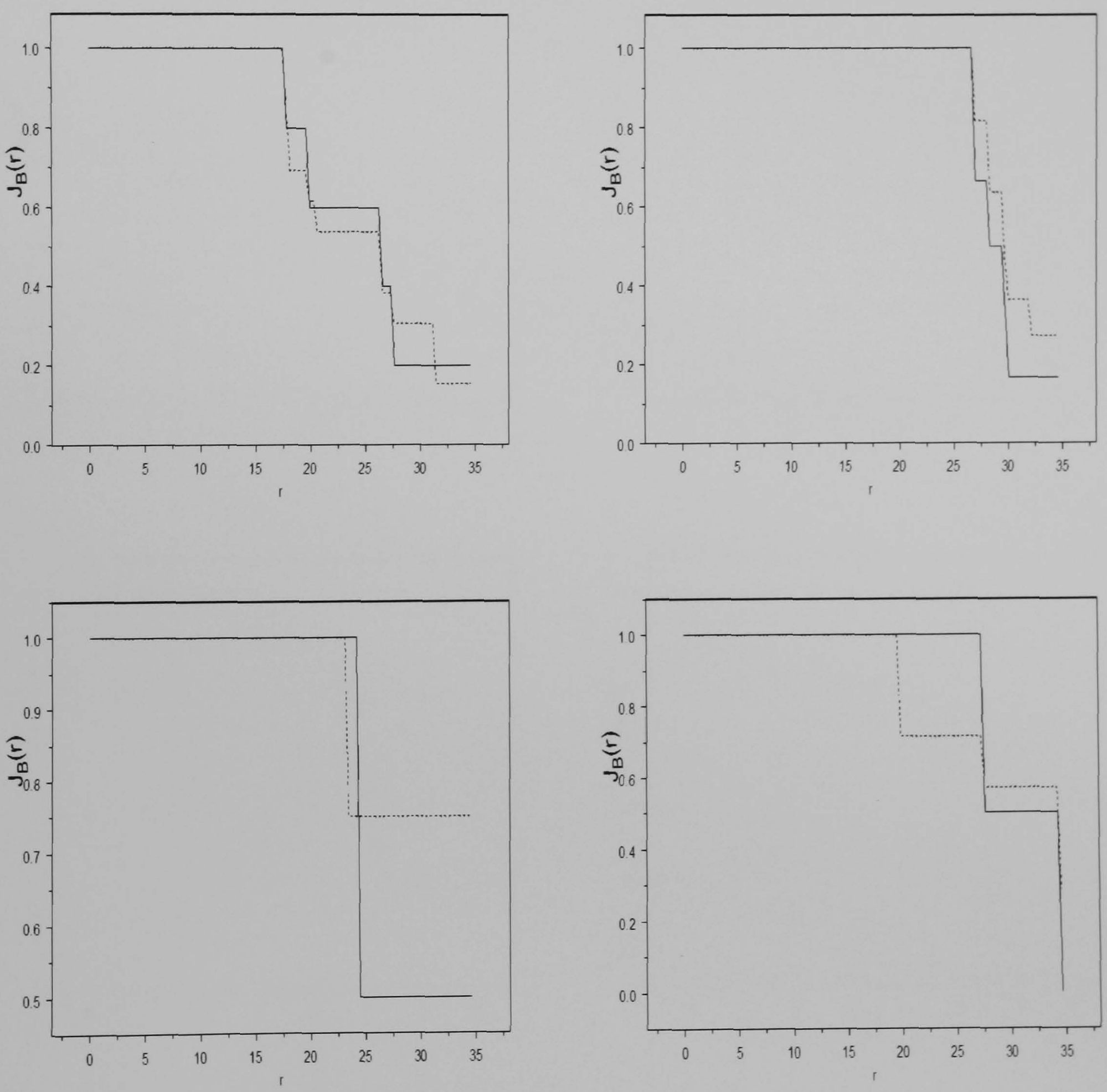
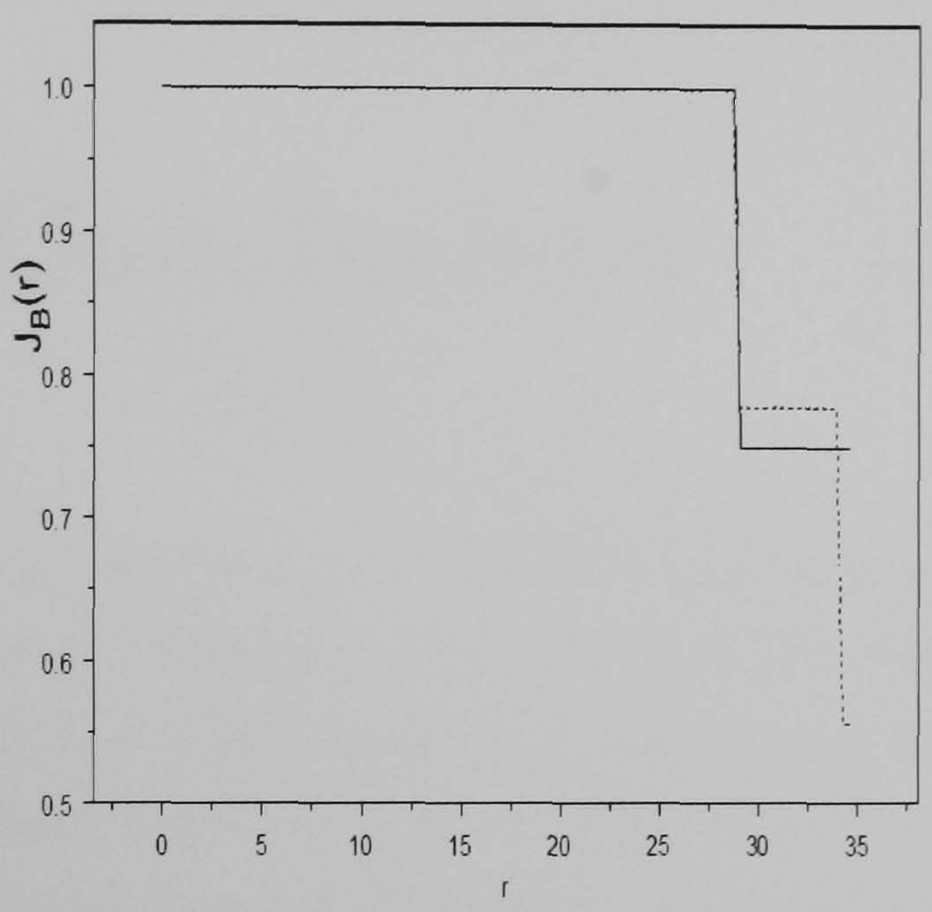
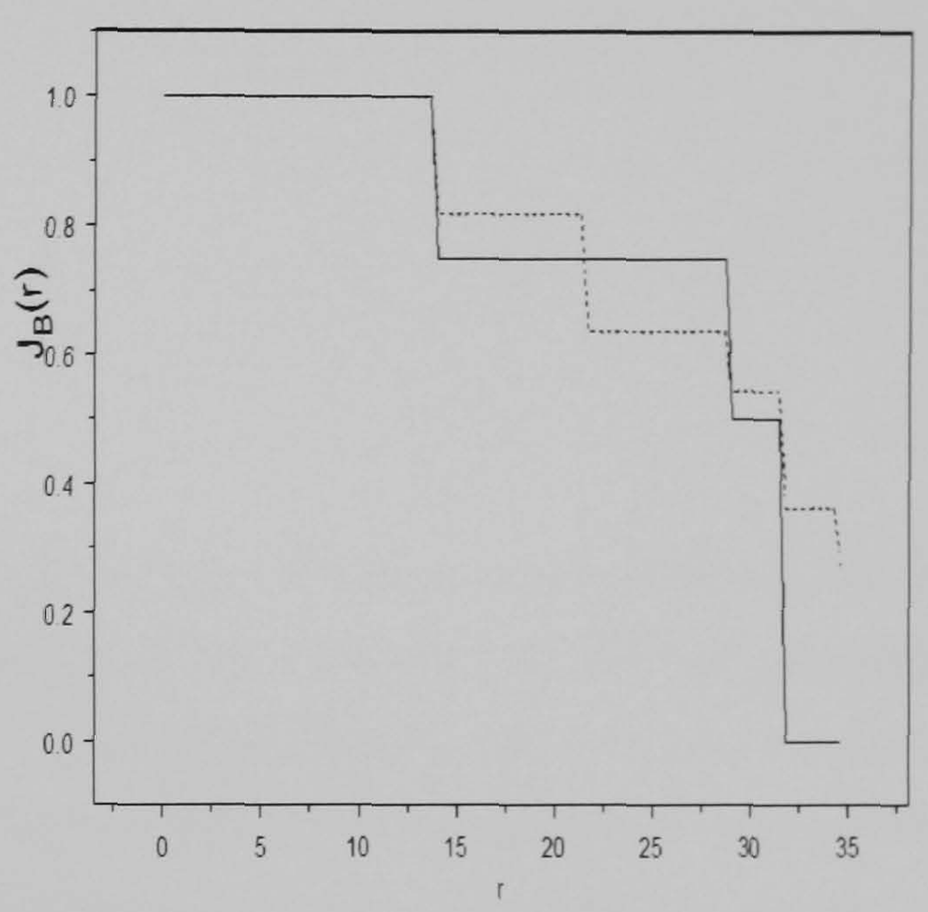
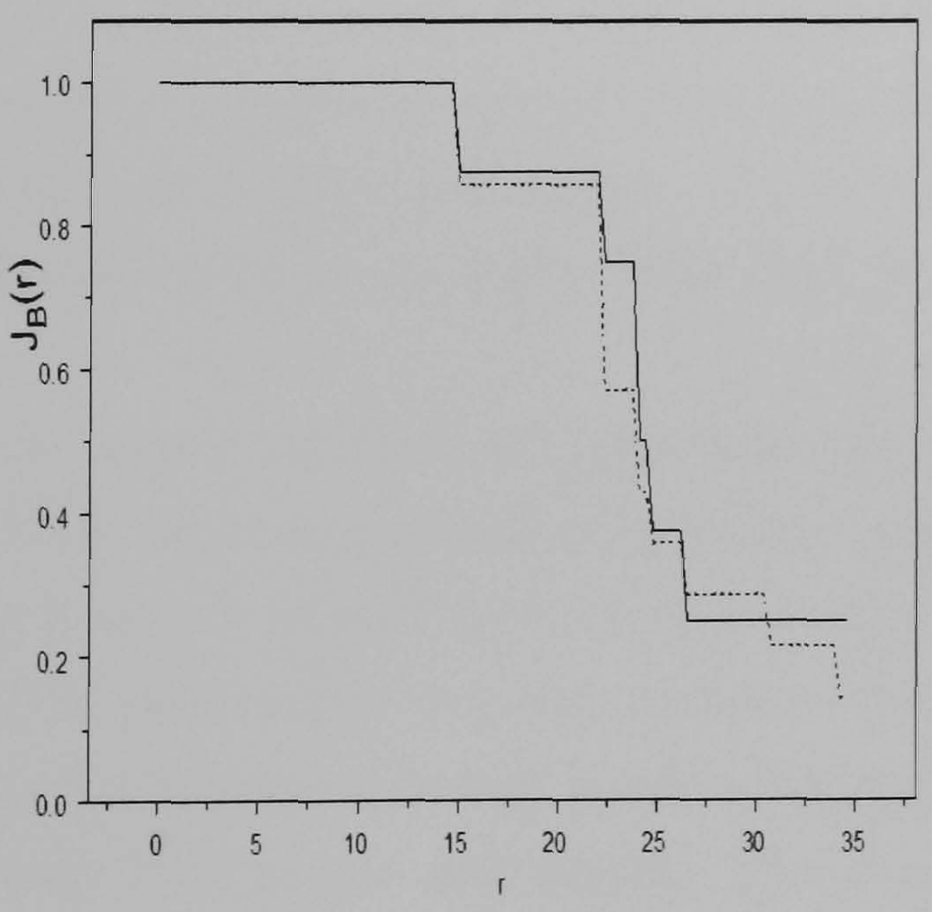


Figure 6.6: Marked (Black curve) and unmarked (dashed line) $J(r)$-functions for PPDS3 cells 1-4. The mark set $B = [0.01, 0.04]$







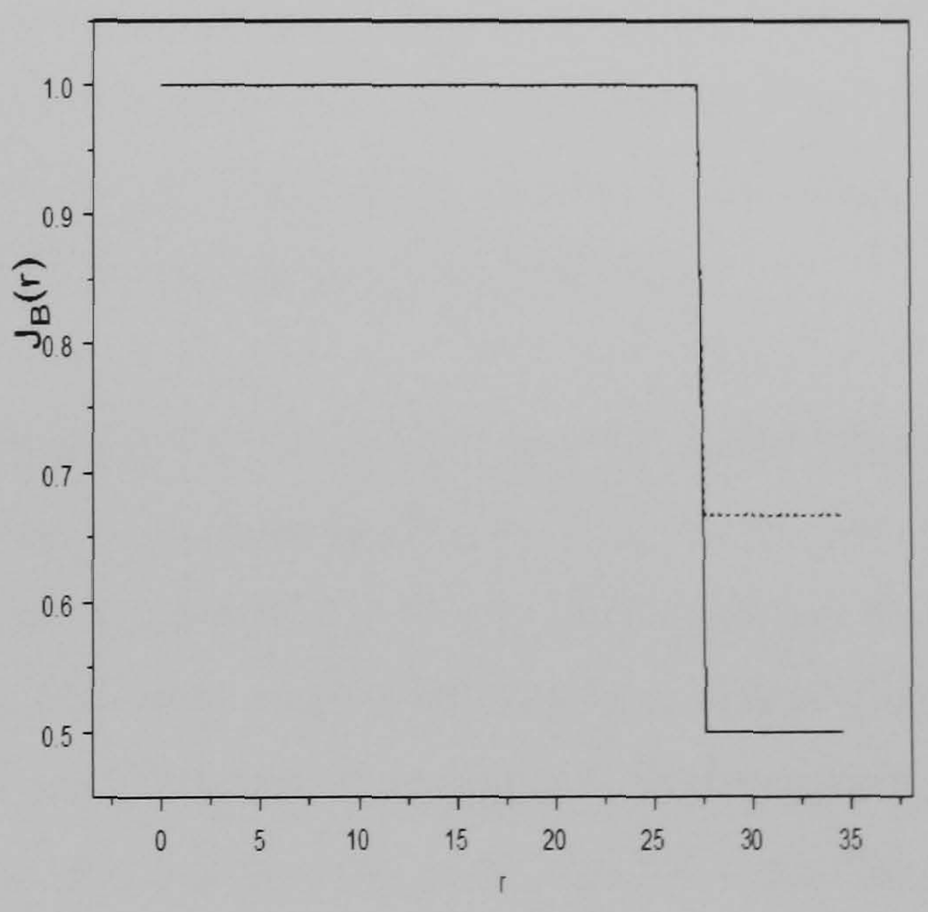


Figure 6.7: Marked (Black curve) and unmarked (dashed line) $J(r)$-functions for PPDS3 cells 5-8. The mark set $B = [0.01, 0.04]$

Chapter 7

Conclusions

The function of PML NBs has been a topic of intense debate in the field of cell biology. Although PML bodies have been implicated in a variety of cellular processes such as DNA repair and tumor suppression (Ching et al, 2005). little is known about their function and biochemical activities. The quantitative analysis of the spatial distribution of PML NBs and its association with other nuclear bodies is an important approach for unlocking some of the mysteries surrounding PML NBs. Through recent developments in confocal microscopy and imaging software, we have been able to obtain spatial point pattern data that describes the spatial locations of PML NB centroids within cell nuclei. As a result, we have been able to carry out various spatial analysis on PML data, within a spatial point process framework.

Like other authors, we have asserted that testing for complete spatial randomness (CSR) is an important step in any spatial point process analysis. The most important aspect of such a test is the insight into the level and direction of departure from CSR, that a spatial point pattern exhibits. CSR tests performed on five MRC-5 cell nuclei showed different levels of departure from CSR for the spatial distribution of PML NBs in the cell nuclei. The results for one particular cell showed that there was not enough evidence to reject the hypothesis of CSR. If the distribution of PML NBs within that cell nucleus exhibits CSR then this means that the PML NBs do not have any form of spatial preference within that particular cell nucleus. The CSR tests provided evidence to reject the hypothesis of CSR for the spatial distribution of PML NBs in the remaining cell nuclei. Furthermore, the PML NBs appeared to exhibit large scale clustering, or, informally speaking, appeared more 'grouped together' over large regions of the nuclear interior, compared to CSR. Also, the PML NBs exhibited small scale repulsion. We suggest that the various outcomes of the

CSR tests are perhaps a result of other influential factors within the cell nuclei. For example, it has been noted in the cell biology literature that nuclear compartments associate (Lanctôt et al, 2007). In particular, PML bodies associate with specific regions of high transcriptional activity (Ching et al, 2005). Moreover, as supported by the work of Dellaire & Bazzet-Jones (2004), the stability of PML bodies is affected by cell stress and changes in chromatin structure (Ching et al, 2005).

CSR tests generally rely on the use of summary statistics. We have discussed several which are 'distance-based'. Including for example the $F(r)$-, $G(r)$-, $g(r)$- and $J(r)$-functions. Estimation of these 'distance-based' summary statistics inside a bounded sampling window is often biased by edge effects. We have reported that edge effects can be ignored in some Monte Carlo based significance tests. However, in situations like that described in Chapter 5, where we wish to compare a theoretical value of a distance-based summary statistic with one that has been estimated in a bounded region, an edge-correction should be adopted. We have outlined and investigated alternative approaches for edge-correcting, and have introduced two new edge-corrections. In particular, we have introduced a probabilistic edge-correction for the $K(r)$-function and a new probabilistic border method. The probabilistic edge-correction for the $K(r)$-function was found to be expensive to compute, but gave a similar outcome to simulation-based CSR tests on PML data. By using various methods for assessing the difference between a theoretical and estimated value of a summary statistic, we also found that the probabilistic edge-correction for the $K(r)$-function generally outperformed the standard weighting factor edge-correction method. Simulation studies were used to show that the probabilistic border method in some cases yielded a small estimation bias compared to the standard border method.

The need to compare a theoretical summary statistic with one that has been estimated with or without an edge-correction has motivated us to deriving explicit closed form expressions for the $K(r)$-functions of a simple inhibition process and Matérn cluster process. Further to the work on edge-effects, as is the case for the PML data that has been analysed throughout, we have demonstrated through various simulation studies that there is generally a greater need for edge-correcting when analysing three-dimensional data, compared with the two-dimensional case. We have advocated a $RELBIAS$ function as a good tool for comparing the severity of edge effects in three-dimensions with 2 and have derived the theoretical $RELBIAS$ obtained when estimating the $K(r)$-function without adopting an edge-correction.

We believe that it is imperative when analysing point patterns to adopt a variety of summary statistics in order to obtain as much information as possible with regards to the spatial distribution of the point pattern. Advocates of this statement include Baddeley et al (1993), who state that no one of the $F(r)$-, $G(r)$-. or $K(r)$-functions should be regarded as "superior to the others, nor as characterizing a point pattern". In this thesis, we have provided some additional point process tools: one of the most important being a third-order characteristic, the $Q(r, \tilde{r})$, that addresses some of the shortcomings of the lower-order characteristics. We believe that this function has important application when requiring a more detailed description of a point pattern, or a more accurate, second opinion is required in certain hypothesis tests, such as those used to determine if a point pattern comes from a particular spatial point process model. Although estimation of the $Q(r, \tilde{r})$-function is computationally more expensive compared to lower-order characteristics such as the $K(r)$-function, we have observed through simulation studies that hypothesis tests based on the $Q(r, \tilde{r})$-function provided competitive results, in terms of power, compared to those based on the powerful, but more difficult to implement Q^2 statistic introduced by Grabarnik & Chiu (2002). We have derived the closed form expressions for the $Q(r, \tilde{r})$-function of some popular spatial point processes and have also discussed estimation of the $Q(r, \tilde{r})$-function and how bootstrapping can be used to assess the estimation uncertainty.

The relatively limited application of the $Q(r, \tilde{r})$-function in this research has been one of the consequences of having a low PML NB count. The biology literature reports cell nuclei usually contain between 5 and 30 PML NBs. The datasets that we have analysed contain at most 15. Indeed, this has also been a problem with regards to other analysis such as estimating summary statistics using methods such as the border estimate. However, the Imperial College London centre for structural biology has recently been producing more data that provides details of the spatial locations of PML and another subnuclear body, inside the same cell. These other bodies, including RNA Polymerase II, sometimes have a much greater body count compared to PML. Although the $Q(r, \tilde{r})$-function of course can be applied in general spatial point pattern analysis, as more PML data of this type is made available, the utility of the $Q(r, \tilde{r})$-function, for this particular application may become more frequent.

We have attempted to formulate a model that best describes the spatial distribution of PML NBs. Several important topics with respect to the spatial distribution of PML NBs and the modeling process have been covered throughout the modeling

process. In particular, with regards to the modelling process that, when formulating a new model for spatial point patterns, there are several important issues that may need to be addressed, for example, whether the model captures the real world phenomena affecting the distribution of the events. In this particular case, this means does the model support the cell biology?

During the modeling process we saw that there appears to be a relationship between PML NBs and the nuclear boundary. More, precisely, the PML tend to be located further away from the boundary compared to if they were placed uniformly in regions unoccupied by nucleoli. This, to a large extent would be in line with some of the cell biology literature such as McManus et al (2006) and recent unpublished work by Imperial College London centre for structural biology.

We have put forward a quadrature technique for fitting an inhomogeneous Poisson process to data that is similar to PPDS2 and PPDS3, and have subsequently fit (to PPDS2) an inhomogeneous Poisson process that has an intensity function capable of capturing the relationship between PML NBs and the nuclear boundary. Detailed diagnostic tests suggested an incorrect specification of the intensity function for some of the cells, and possible evidence for interaction amongst the PML NBs (which is in line with the initial CSR test results). In order to capture inter-PML interaction, we subsequently considered a spatial Markov point process with a spatially varying intensity, as an alternative model to the inhomogeneous Poisson. We showed through a pilot study that it was relatively impractical to fit the model to the replicated PML data, without having good prior knowledge of the model parameters. Nevertheless, we have demonstrated that, as a result of a lack in available closed form expressions for Markov models, difficult to implement optimisation routines that may require approximating partition functions using methods such as thermodynamic integration, may need to be adopted when fitting these models. We have also explored the possibility of fitting a log-Gaussian Cox process to PPDS2 by using minimum contrast. The minimum contrast approach required us to simulate a log-Gaussian Cox process inside a three-dimensional convex hull. We are not aware of any other authors that have explored this complicated task of simulating a log-Gaussian Cox process in three-dimensions.

In Chapter 6 we discussed the importance of marked spatial point process analysis and its application to the datasets PPDS2 and PPDS3. Early marked spatial point pattern applications have provided some noteworthy results. Analysis using

the $K_{ij}(r)$ function suggested that, generally fewer RNA Polymerase II exist within a particular distance of PML bodies compared to if the RNA Polymerase II exhibited CSR. Calculations showed that the results are, to some extent, in agreement with the findings of Xie & Pombo (2006). Following further marked spatial point pattern analysis, we did not find evidence to support a relationship between PML NB size in terms of length relative to the nuclear length, with spatial location in terms of relative distance to the nuclear boundary. In fact, through a straightforward correlation calculation, we found that the two were uncorrelated.

The conclusions we have presented here are based on the statistical analysis of point patterns that have been generated via a very complicated series of steps. We are aware that there are some issues with regards to generating the data. These issues, along with some of the research that we have presented, give rise to some potentially important future work.

During the data capture process, in order to obtain a three-dimensional representation of the cell nuclei, one of the coordinates is to some extent artificially extracted by 'stacking' a series of two-dimensional slices. Throughout this research, we have used replicated data in order to reduce the impact of the various sources of variability in the data. With this in mind, we would suggest that some of the experiments that have been carried out earlier, such as those regarding whether the spatial distribution of PML NBs is CSR, should be repeated for several other cells as more data is made available.

We have been cautious in comparing cells that are as 'similar' as possible. More specifically, when constructing the datasets, we have received guidance from Imperial College London CSB cell biologists on which cells can be compared. Following their advice, we have put together datasets (PPDS1, PPDS2 and PPDS3) that consists of cells which are of the same type, and in PPDS2, are also at the same stage of the cell cycle. We would recommend that a similar approach is adopted in future work.

During the early stages of this research, some cell biologists suggested that nucleoli are exclusion regions for PML. Throughout the modelling process we generally assumed that this is indeed the case. However, recent evidence suggests that this may not be the case. To adjust our models for this alternative idea, we could introduce the concept of a stochastic nucleoli exclusion zone. In future work, adapting our

models and simulation procedures to allow for PML bodies to exist within nucleoli is straightforward. We simply allow the points that are labeled nucleoli to be candidate empty space points within the nuclear interior.

Using the evidence we have obtained in carrying out this research, we believe that an appropriate model for the spatial distribution of PML should be one that correctly captures the interaction between PML and the boundary, captures inter-PML interaction, and is conditional on other bodies within the nucleus. Furthermore, we believe that a good candidate model is perhaps a multivariate, generalised Shot Noise Cox process (Møller & Torrisi, 2005). Such a model would be able to capture an appropriate intensity surface, inter-PML interaction (for example clustering), interaction between PML and other bodies, and any random variation between cells. However, from our experience, we believe that there will be difficulties in fitting such a complicated model, partly due to the low PML count and partly due to the complexity of the model.

The modeling process is a natural next step in point pattern analysis. However, obtaining a model for the spatial distribution of PML NBs goes beyond being an interesting statistical exercise, as it has apparent importance in the fields of cell biology and medicine. As an example, successfully fitting an inhomogeneous Poisson process to replicated PML NB data, suggests a form of a spatial preference for PML within the cell nucleus. Also, very importantly, the formulation of an appropriate model can have potential applications in spotting certain illnesses by comparing the distributon of PML NBs from cells that have been taken from the 'patient' being diagnosed, with the distribution of PML NBs as suggested by the model. At present, this is of course rather ambitious, given the technological limitations.

Appendix A

Basic overview on measure theory, topological and metric spaces

Definition 1 *A metric on some set S is a function $d : S \times S \to \mathbb{R}$ that satisfies*
i) $d(x, y) \geq 0$
ii) $d(x, y) = 0$ *if and only if* $x{=}y$
iii) $d(x, y) = d(y, x)$
iv) $d(x, z) \leq d(x, y) + d(y, z)$
for all x,y, *and* z *in* S.

Definition 2 *Let S be a set and d a metric on S. A metric space is the pair (S,d). S itself is called a metric space.*

Definition 3 *Let (S,d) be a metric space. If $x \in S$ and $r \in \mathbb{R}^+$, then the sets $b(x,r)$ defined by*
i) $\{y \in S : d(x, y) \leq r\}$
ii) $\{y \in S : d(x, y) < r\}$
are called a closed ball and open ball respectively.

Definition 4 *Let S be a metric space. The diameter of $A \subset S$, denoted $diam(S)$ is defined by*

$$diam(A) = sup\{d(x, y) : x, y \in A\}$$

Definition 5 *The set A is said to be bounded if $diam(A) \neq +\infty$.*

Definition 6 *Let S be a metric space and A a non-empty subset of S. The distance between the point x and A is defined by*

$$d(x, A) = inf\{d(x, y) : y \in A\}$$

Definition 7 *Let Ω be an arbitrary set. A collection $\mathcal{S}$ of subsets of Ω is an algebra on Ω if*
a)$\Omega \in \mathcal{S}$,
b) for every set S that belongs to $\mathcal{S}$ the set S^c belongs to $\mathcal{S}$.
c) for every finite sequence of sets, $S_1, \ldots, S_n$, that belong to $\mathcal{S}$, the set $\cup_{i=1}^{n} S_i$ belongs to $\mathcal{S}$, and
d) for each finite sequence of sets $S_1, \ldots, S_n$ that belong to $\mathcal{S}$ the set $\cap_{i=1}^{n} S_i$ belongs to $\mathcal{S}$.

Definition 8 *Let Ω be an arbitrary set. A collection $\mathcal{S}$ of subsets of Ω is a σ-algebra on Ω if*
a) $\Omega \in \mathcal{S}$,
b) for every set S that belongs to $\mathcal{S}$ the set S^c belongs to $\mathcal{S}$
c) for every infinite sequence of sets, $S_1, \ldots$ that belong to $\mathcal{S}$, the set $\cup_{i=1}^{\infty} S_i$ belongs to $\mathcal{S}$, and
d) for each infinite sequence of sets $S_1, \ldots$ that belong to $\mathcal{S}$ the set $\cap_{i=1}^{\infty} S_i$ belongs to $\mathcal{S}$.

Definition 9 *The Borel σ-algebra on $\mathbb{R}^d$ is the σ-algebra on $\mathbb{R}^d$ generated by the set of open subsets of $\mathbb{R}^d$.*

A Borel set is an element of a Borel σ-algebra. The set of rational numbers is an example of a Borel set.

Definition 10 *A measurable space is an ordered pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a σ-algebra on Ω.*

Definition 11 *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. A map $f : \Omega_1 \to \Omega_2$ is said to be measurable with respect to $\mathcal{F}_1$ and $\mathcal{F}_2$ if and only if for all $B \in \Omega_2$, we have that $f^{-1}(B) \in \Omega_1$*

Definition 12 *Let Ω be a set. A topology on Ω is a collection, $\mathcal{T}$ of subsets of Ω such that*
a) $\Omega \in \mathcal{T}$
b) $\emptyset \in \mathcal{T}$
c) if $\mathcal{S}$ is an arbitrary collection of sets that belong to $\mathcal{T}$, then $\cup \mathcal{S} \in \mathcal{T}$, and
d) if $\mathcal{S}$ is a finite collection of sets that belong to $\mathcal{T}$, then $\cap \mathcal{S} \in \mathcal{T}$

Definition 13 *A topological space is an ordered pair $(\Omega, \mathcal{T})$, where Ω is a set and $\mathcal{T}$ is a topology on Ω*

Definition 14 *A topological space,$T = (\Omega, \mathcal{T})$ is a Hausdorff space if and only if for all $\omega_1, \omega_2 \in \Omega$ with $\omega_1 \neq \omega_2$, there exists open sets U and V in Ω, such that*

$$\omega_1 \in U, \quad \omega_2 \in V, \quad U \cap V = \emptyset$$

Definition 15 *Let $T = (\Omega, \mathcal{T})$ be a topological space. The set A is open if $A \in \mathcal{T}$. Otherwise the set A is said to be closed.*

Definition 16 *A collection $\mathcal{S}$ of open subsets of a topological space.$T = (\Omega, \mathcal{T})$ is a base for T if for each $V \in \mathcal{T}$ and each $x \in V$ there is a set U that belongs to $\mathcal{S}$ and satisfies $x \in U \subset V$. Furthermore, T is said to have a countable base if it has a basis that contains only countably many sets.*

Definition 17 *Let T be a topological space and $t \in T$. A neighbourhood of t is a set V that contains an open set U containing t, $p \in U \subset V$*

Definition 18 *A set $S \subset \mathbb{R}^d$ is said to be compact if it is both closed and bounded. Examples of compact sets include the closed balls and the closed hypercubes.*

Definition 19 *The closure of the set A, is the intersection of all closed sets containing A.*

Definition 20 *A topological space is said to be locally compact if each of its points has an open neighbourhood whose closure is compact.*

Examples of locally compact Hausdorff spaces include the Euclidean spaces $\mathbb{R}^d$

Definition 21 *Let $(\Omega, \mathcal{F})$ be a measurable space. A measure on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0, \infty]$ with the properties*
i) $\mu(\emptyset) = 0$
ii) $\mu(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mu(A_k)$
for all $A_1, A_2 \ldots \in \mathcal{F}$ with $A_i \cap A_j = \emptyset$

Definition 22 *Let X be a metric space, $d \in \mathbb{R}^+$, $E \subset X$, $\Gamma(.)$ the gamma function. and*

$$\gamma_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(d/2+1)},$$

then the d-dimensional Hausdorff measure of E, $\mathcal{H}_d(E)$ is the number that is defined as follows:

$$\mathcal{H}_d(E) = \lim_{\epsilon \to 0^+} \mathcal{H}_d^{\epsilon}(E)$$

where

$$\mathcal{H}_d^{\epsilon}(E) = inf\left\{\sum_{i=0}^{\infty}\gamma_d\left(\frac{diam(E_i)}{2}\right)^d : E_i \subset X, \bigcup_{i=0}^{\infty} E_i \supset E. diam(E_i) \leq \epsilon \forall i = 0, 1, \ldots\right\}$$

for all $\epsilon > 0$

Until stated otherwise, let E be a locally compact Hausdorff space with Borel σ-algebra $\mathcal{F}$. Let $\mathcal{B}$ be the collection of bounded sets in $\mathcal{F}$

Definition 23 *A measure* μ *on* E *is a Radon measure if* $\mu(B) < \infty$ *for each* $B \in \mathcal{B}$.

Definition 24 *A Radon measure* μ *is said to be a point (or counting) measure if* $\mu(A) \in \mathbb{N}$ *for every* $A \in \mathcal{B}$.

Definition 25 *A Radon measure* μ *is a simple point measure if* $\mu(\{x\}) \leq 1$ *for all* $x \in E$.

Definition 26 *A Radon measure* μ *is said to be diffuse if* $\mu(\{x\}) = 0$ *for every* $x \in E$.

Theorem 8 (Radon-Nikodym theorem) *Let* $(\Omega, \mathcal{F})$ *be a measurable space and* μ_1 *and* μ_2 *be* σ*-finite positive measures on* $(\Omega, \mathcal{F})$ *with* μ_2 *absolutely continuous with respect to* μ_1. *Then there exists an* $\mathcal{F}$*-measurable function* $f : \Omega \to [0, \infty)$ *such that* $\nu_2(F) = \int_F f d\mu_1$ *holds for all* $F \in \mathcal{F}$. *The function* f *is unique up to* μ_1*-almost everywhere equality.*

Appendix B

Datasets PPDS1, PPDS2 and PPDS3

All of the datasets used (PPDS1, PPDS2 and PPDS3) were provided by the Imperial College London centre for structural biology. The cells used are MRC-5 cell nuclei (see for example Jacobs et al, 1970). Furthermore, the cells used for PPDS2 are all in the G_0 phase of the cell cycle.

B.1 Distance units for the data

Throughout the thesis we refer to "units" for reporting measured distances in the cell nuclei. r units corresponds to r pixels. There are 12 pixels in $1\mu m$. Hence 1 unit ≈ 0.083 micrometers (μm) or 83.3 nanometers (nm).

B.2 PPDS1

The dataset PPDS1 consists of 5 cells. We are provided with point coordinates that are the centroids of PML NBs inside the cell nuclei. The sampling region is obtained by calculating the smallest ellipsoid that contains all of the PML points (see Chapter 1).

B.3 PPDS2

Figure B.3 demonstrates some of the stages for obtaining the data that make up PPDS2. By calculating the maximum distance between two points on the convex hull, that represents the boundary of a cell, we obtain a nuclear 'length'.

Cell	PML Count
1	9
2	14
3	13
4	12
5	15

Figure B.1: The details for PPDS1

PPDS2 Details

Cell	PML Count	Cell Volume	U
1	12	1,260,497	49,942
2	11	1,062,381	50,189
3	9	1,256,605	49,975
4	7	1,185,842	49,969
5	9	1,248,903	49,900
6	11	1,204,560	49,974
7	14	913,805	50,091
8	6	937,887	50,078

PPDS2(b) Details

Cell	PML Count	U
1	14	81,965
2	14	87,461
3	12	79,623
4	15	82,288
5	14	106,571
6	18	83,021

Figure B.2: PPDS2 details. U is the number of interior points that are used for simulating inside the nucleus and estimating integrals, as done in Chapter 5. This is a subset of the number of interior points provided by the CSB, which is generally in excess of 100,000 points. We take a random subset so that computational burden is reduced and we are not faced with computer memory issues when running programs.

B.4 PPDS3

The dataset that makes up PPDS3 is formed in the same way as those that make up PPDS2.

B.5 Data simulated using perfect simulation

Figure B.5 shows simulated realisations from Model 3 (see Chapter 5) using perfect simulation. The data displayed in Figure B.5 is intended to mimic a set of data that

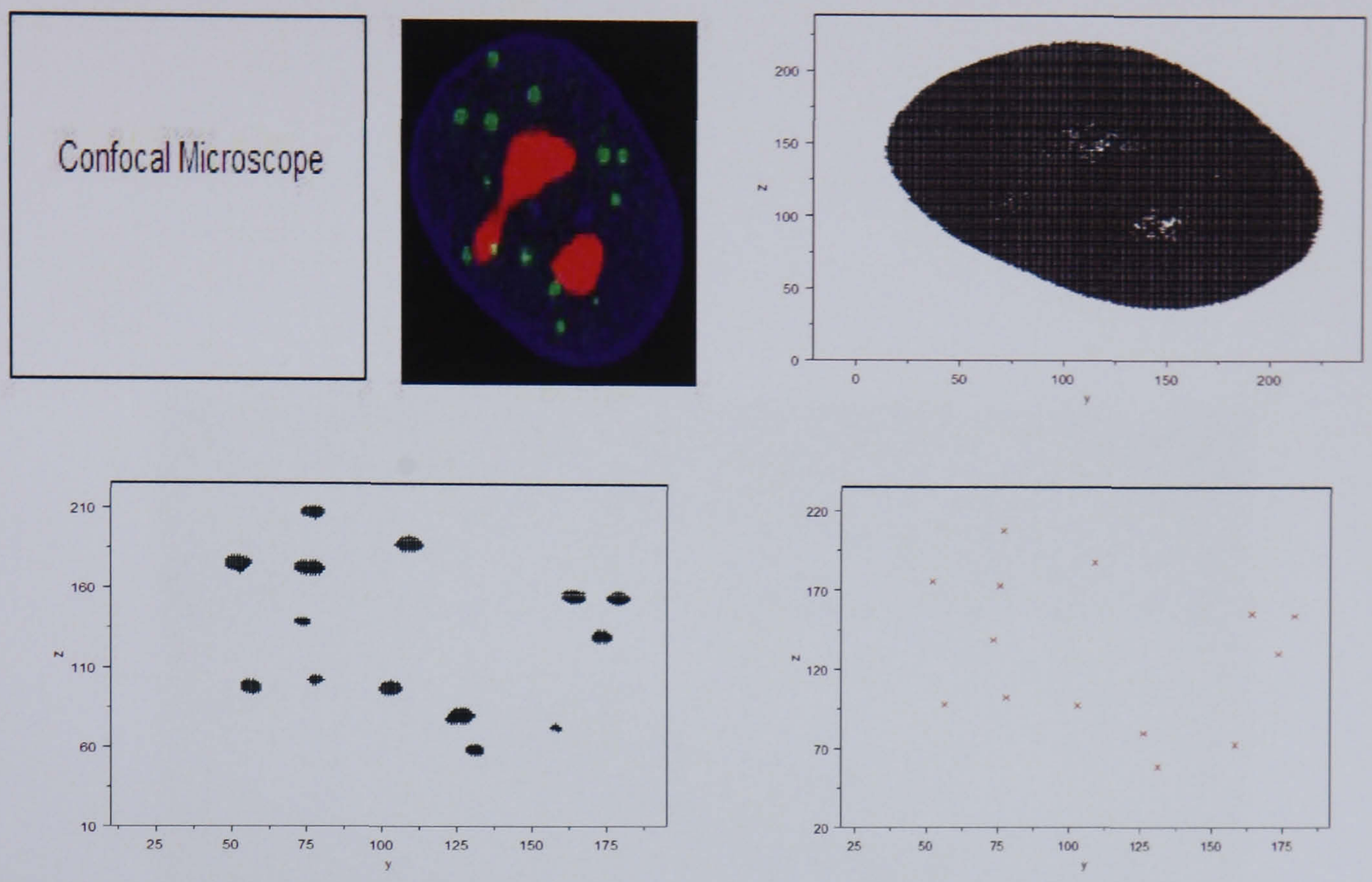


Figure B.3: The different processes for obtaining the data that make up PPDS2 (cell 7 is used as an example). First the confocal image is produced (top left). The image data is analysed by the CSB to form datasets consisting of points that are labeled PML (bottom left), nucleoli, nuclear boundary, or empty space (top right) inside the nucleus. We then run this large dataset through a program that forms a point pattern by converting the PML bodies into PML points, by calculating centroids. Note that the nuclear boundary is obtained by calculating a convex hull for the points that have been classified from the image data (see Russell, 2006 for details)

is similar to PPDS2 and PPDS3.

Cell	PML Count	RNA Pol II Count	U
1	6	349	50,216
2	8	269	50,253
3	13	296	50,266
4	12	72	50,266
5	10	226	50,250
6	14	125	49,982

Figure B.4: PPDS3 details. U is the number of interior points that are used for simulating inside the nucleus and estimating integrals, as done in Chapter 5. This is a subset of the number of interior points provided by the CSB, which is generally in excess of 100,000 points. We take a random subset so that computational burden is reduced and we are not faced with computer memory issues when running programs.

Figure B.5: Two-dimensional projections of simulated realisations of Model 3 (using CFTP) with parameters $\alpha = 100, \kappa = 0.004$ and $R = 1.5$ inside the sphere of radius 10. The horizontal and vertical axis are respectively the $\eta^{(2)}$ and $\eta^{(2)}$ coordinates

Appendix C

Look up Tables

Table C.1: Lookup table for values of σ for 100 simulations of a homogeneous Poisson process in the unit square

Intensity	mean σ	standard deviation σ	σ_{MAX}	σ_{MIN}
10	464	2760	27623	72
20	10395	96155	961096	51
30	1518	10986	107780	28
40	343	1971	19150	24
42	297	1472	14116	22
50	148	515	4869	16
60	125	422	3617	17
62	108	349	2956	16
65	89	278	2473	16
70	105	309	3530	16
80	75	178	1563	14
90	55	99	712	12
100	60	194	1885	8
150	27	52	382	3
200	14	18	98	2

Appendix D

Derivation of the Q-function for a Gauss-Poisson process

In this Appendix we provide further discussion on how we derive the Q-function of a Gauss-Poisson process. We require to evaluate the expression

$$\int \sum_{\eta,\zeta \in x \cap b^d(o,r)\setminus\{o\}} d_{\tilde{r}}(\eta,\zeta) P(dx) + 2 \int \int \sum_{\substack{\eta \in x \cap b^d(o,r)\setminus\{o\} \\ \zeta \in y \cap b^d(o,r)\setminus\{o\}}} d_{\tilde{r}}(\eta,\zeta) P(dx) c_o(dy)$$

$$+ \int \sum_{\eta,\zeta \in y \cap b^d(o,r)\setminus\{o\}} d_{\tilde{r}}(\eta,\zeta) c_o(dy) \tag{D.1}$$

First, we have

$$\int \sum_{\eta,\zeta \in x \cap b^d(o,r)\setminus\{o\}} d_{\tilde{r}}(\eta,\zeta) P(dx) = 2\lambda^2 Q_{\mathrm{poi}}(r,\tilde{r}).$$

We also have that

$$\int \sum_{\eta,\zeta \in y \cap b^d(o,r)\setminus\{o\}} d_{\tilde{r}}(\eta,\zeta) c_o(dy) = 0,$$

since the integral represents the number of candidate triples from the same cluster. Now, we have that for every measurable function f

$$\int f(y) c_o(dy) = \overline{c}^{-1} E \sum_{z \in X_0} f(X_o - z), \tag{D.2}$$

where $\overline{c}$ is the expected number of events for the representative cluster. Clearly

$$\overline{c} = 0 \cdot q_0 + 1 \cdot q_1 + 2 \cdot q_2.$$

Also, by stationarity of the Gauss-Poisson process

$$\int\int \sum_{\substack{\eta\in x\cap b^d(o,r)\setminus\{o\}\\ \zeta\in y\cap b^d(o,r)\setminus\{o\}}} d_{\tilde{r}}(\eta,\zeta)P(dx)c_o(dy) =$$

$$\lambda\int \sum_{\zeta\in y\cap b^d(o,r)} \nu(b^d(o,r)\cap b^d(\eta,r))c_o(dy) \tag{D.3}$$

For a Gauss-Poisson process, the distribution of the second cluster point, given the first, η, is the uniform distribution on $\partial b(\eta,s)$. Inserting this into (D.3) and combining with (D.2) yields the assertion.

Appendix E

$BIAS$ curves and simulation envelopes

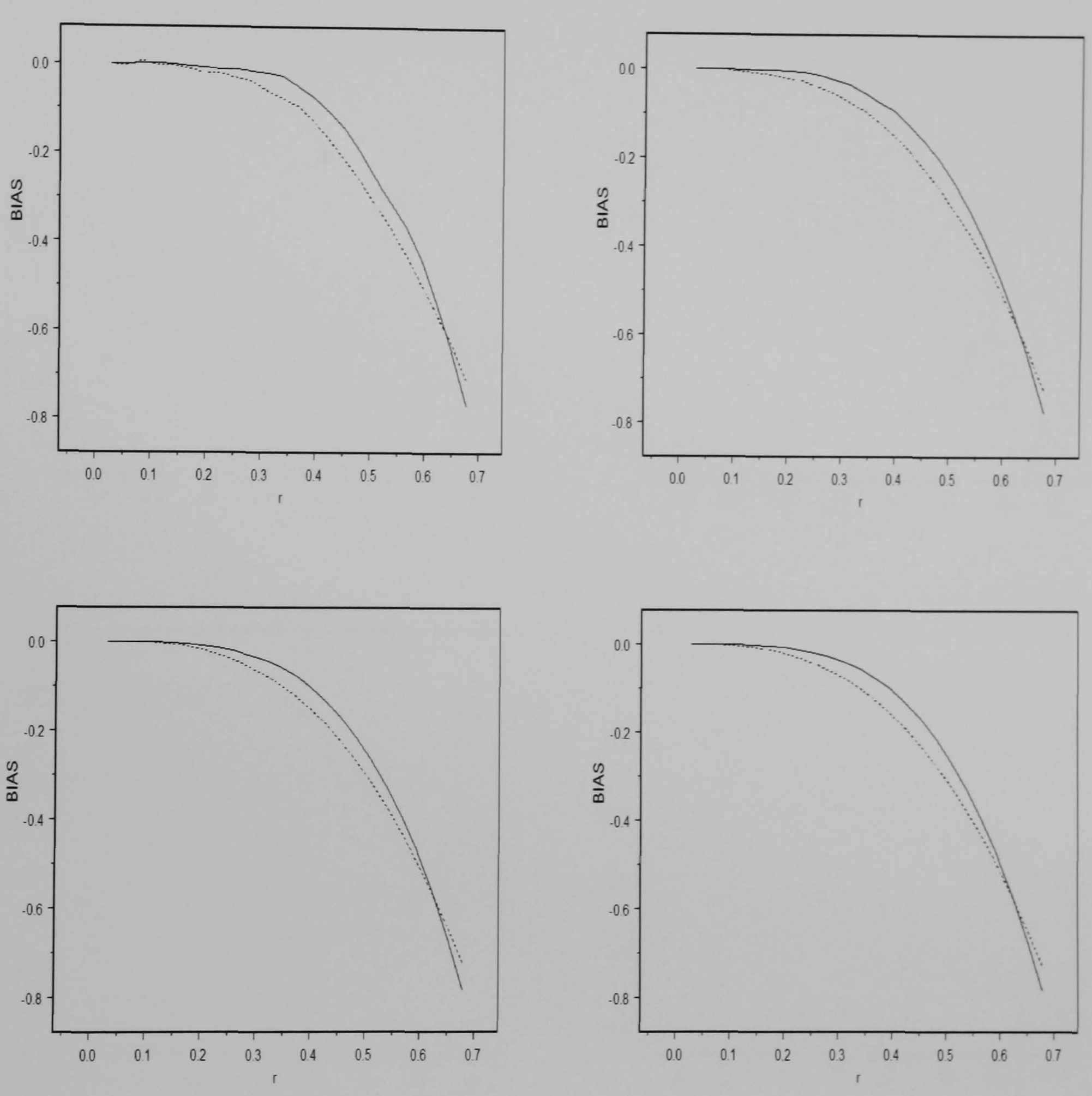


Figure E.1: $BIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 5$ (top left), 10 (top right), 30(bottom left), 100 (bottom right) conditional number of events) inside a cube (black line) and square (dashed line) of length 1

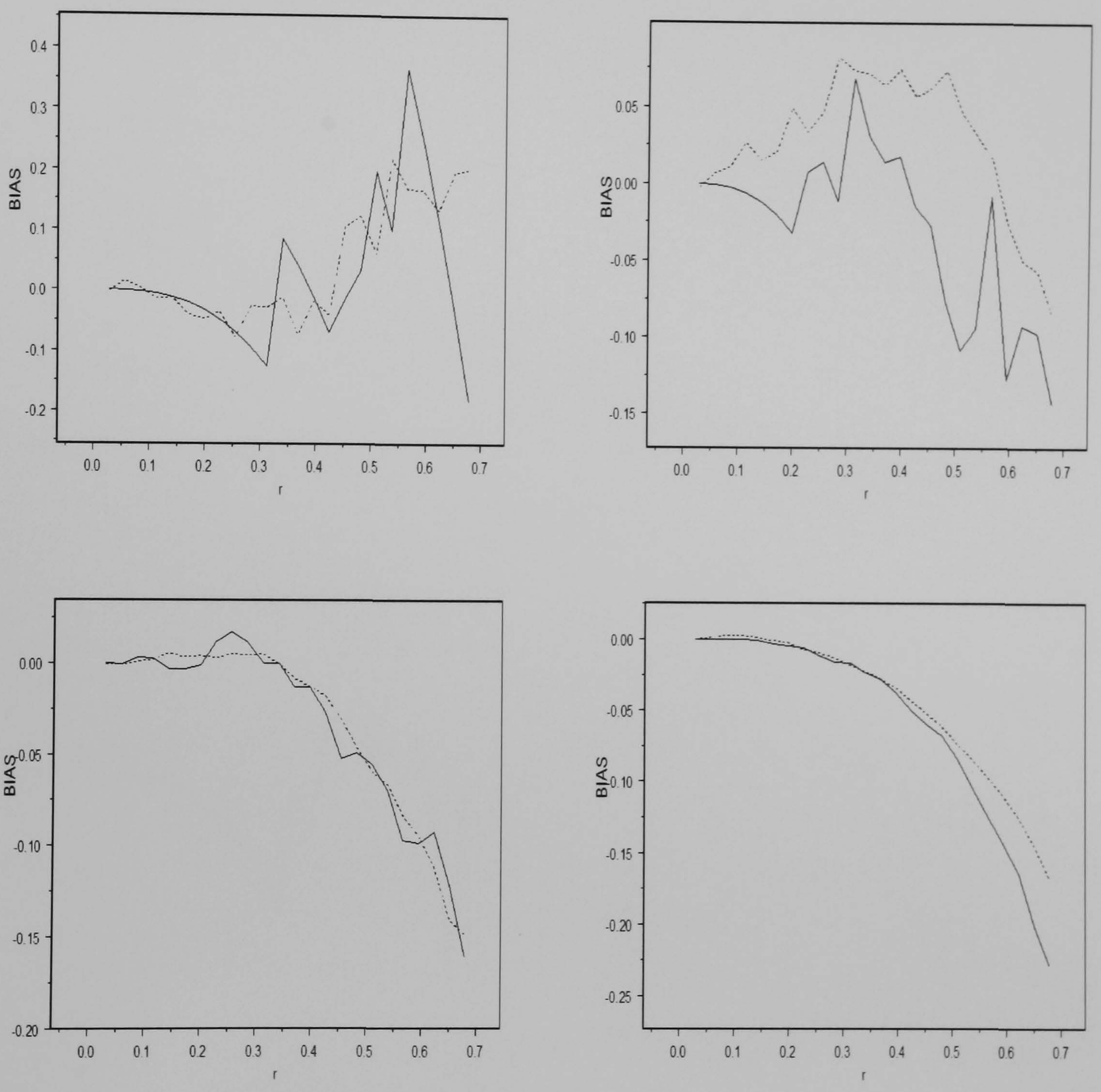
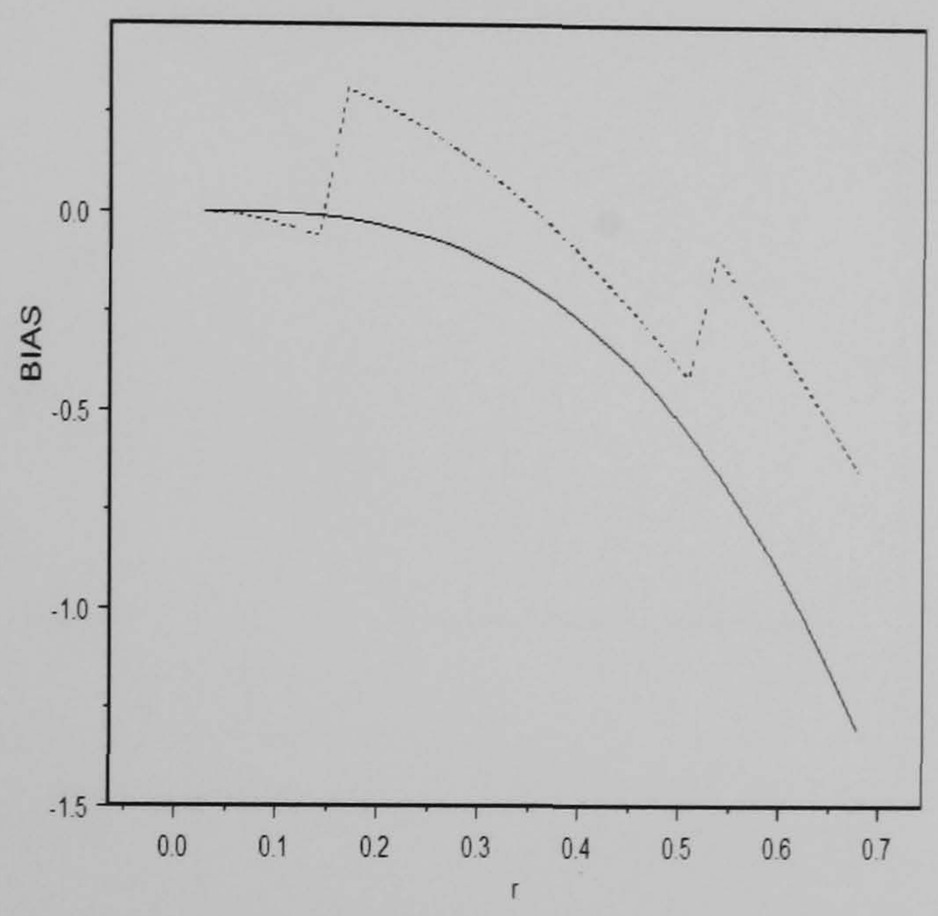
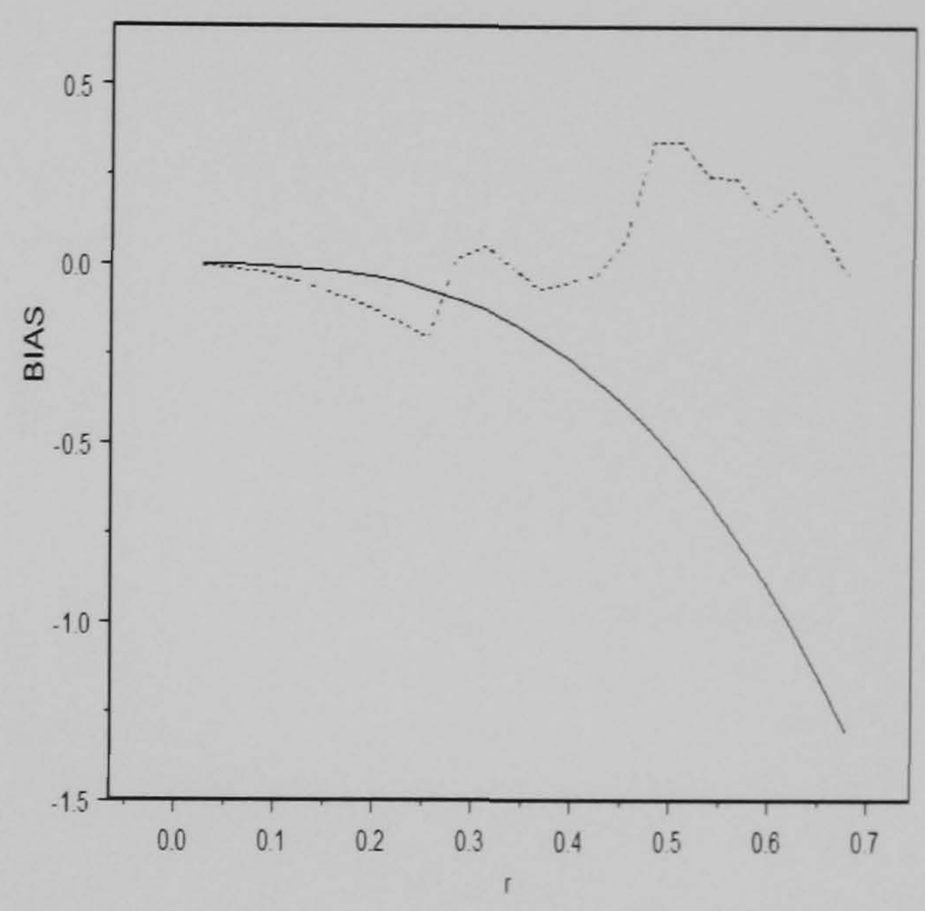
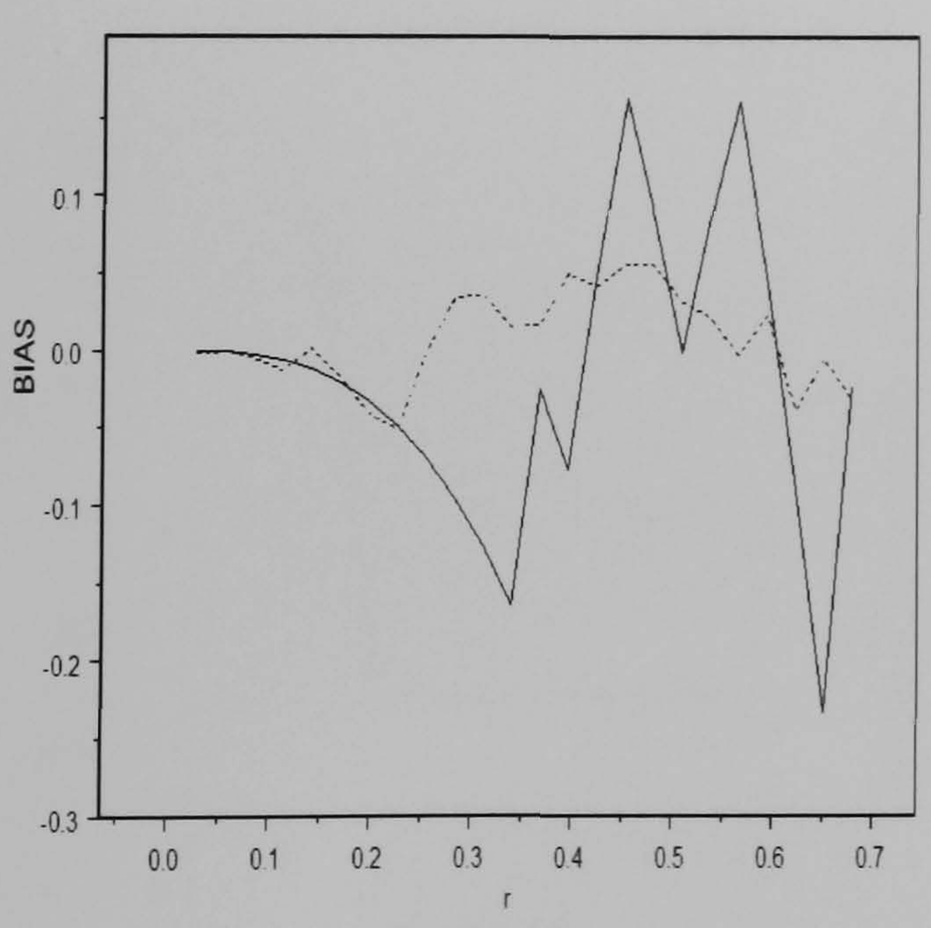


Figure E.2: $BIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 5$, 10, 30, 100 conditional number of events) inside a cube (black line) and square (dashed line) of length 5







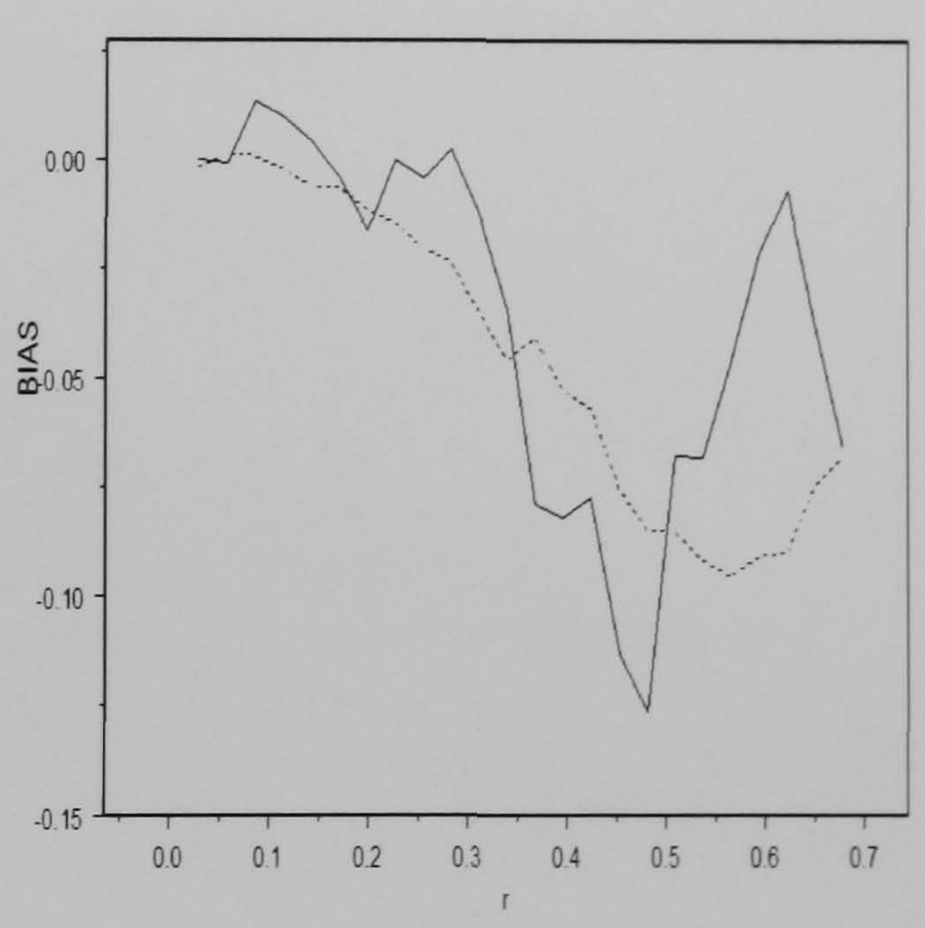
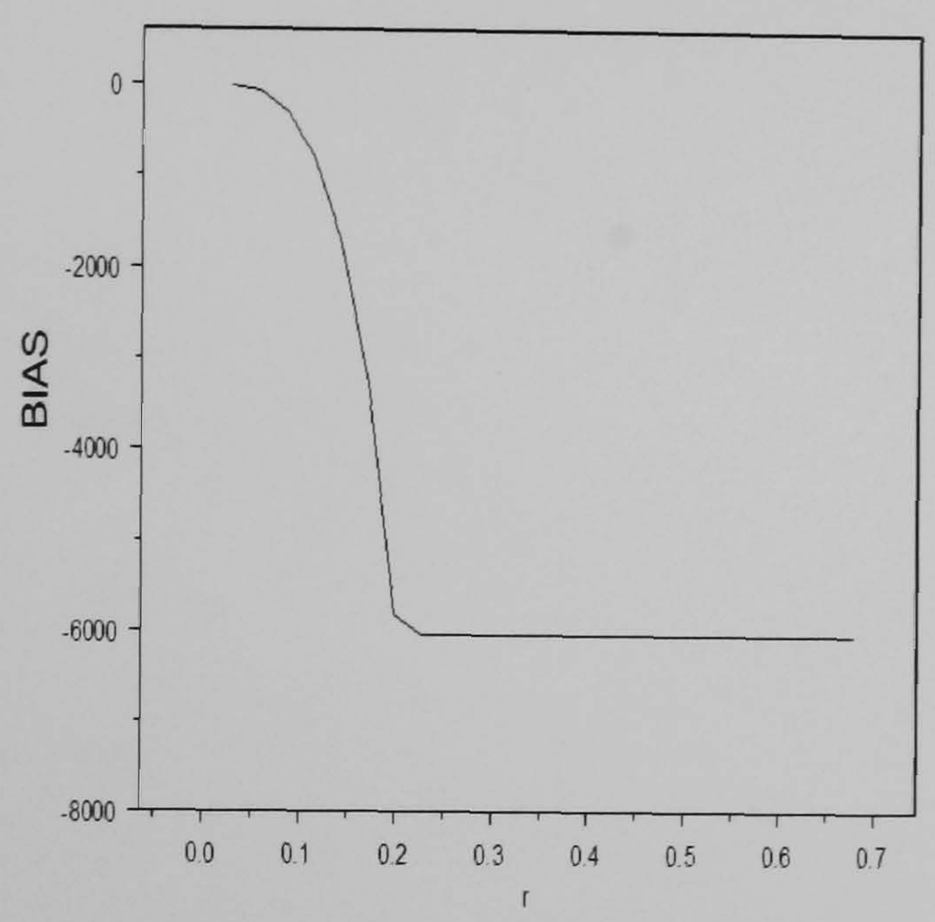
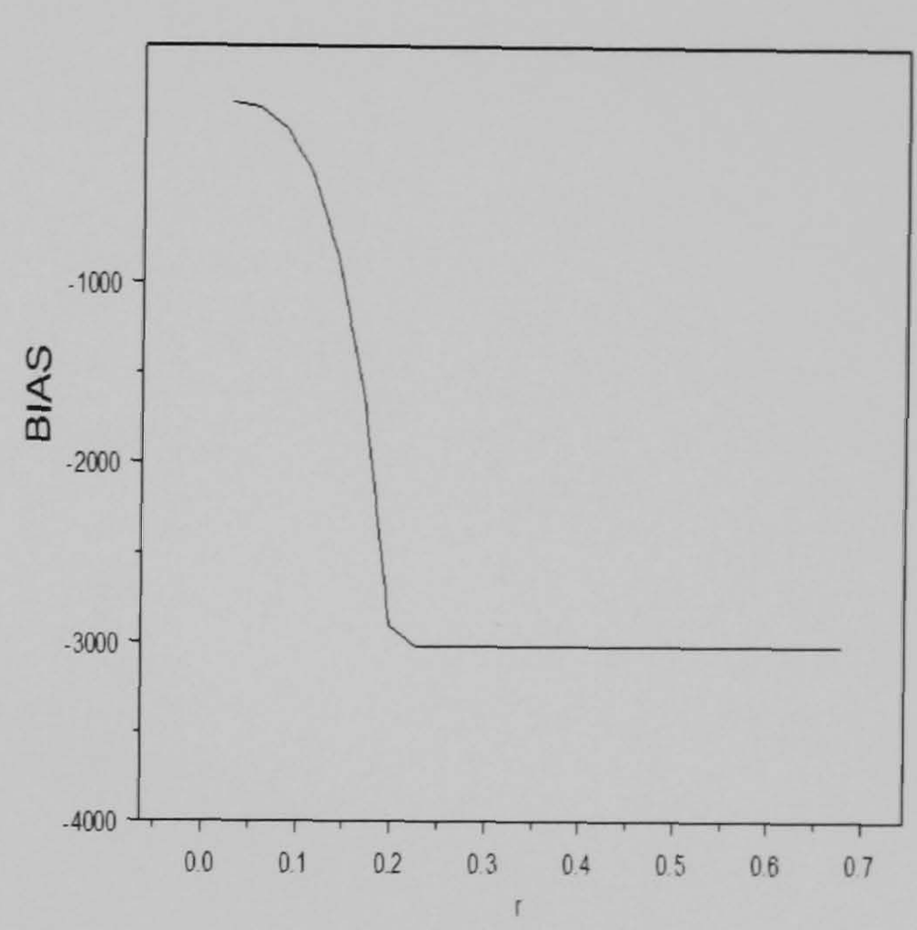
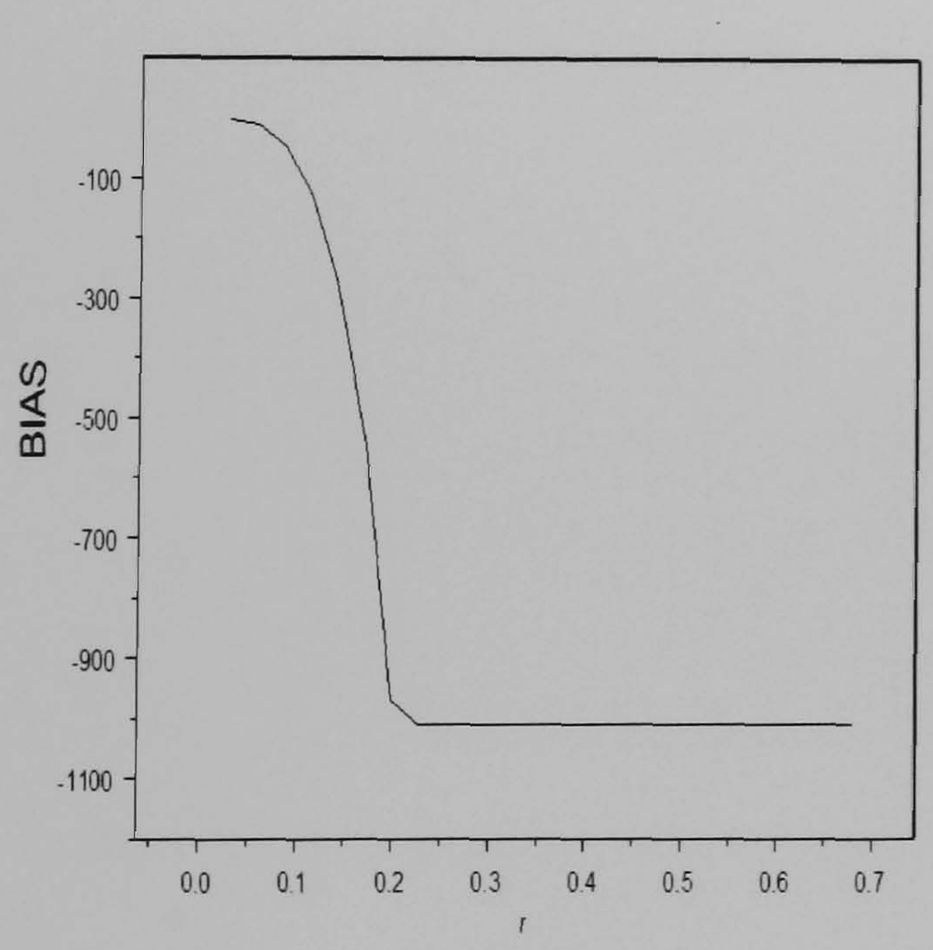


Figure E.3: $BIAS$ of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 5$, 10, 30, 100 conditional number of events) inside a cube (black line) and square (dashed line) of length 20







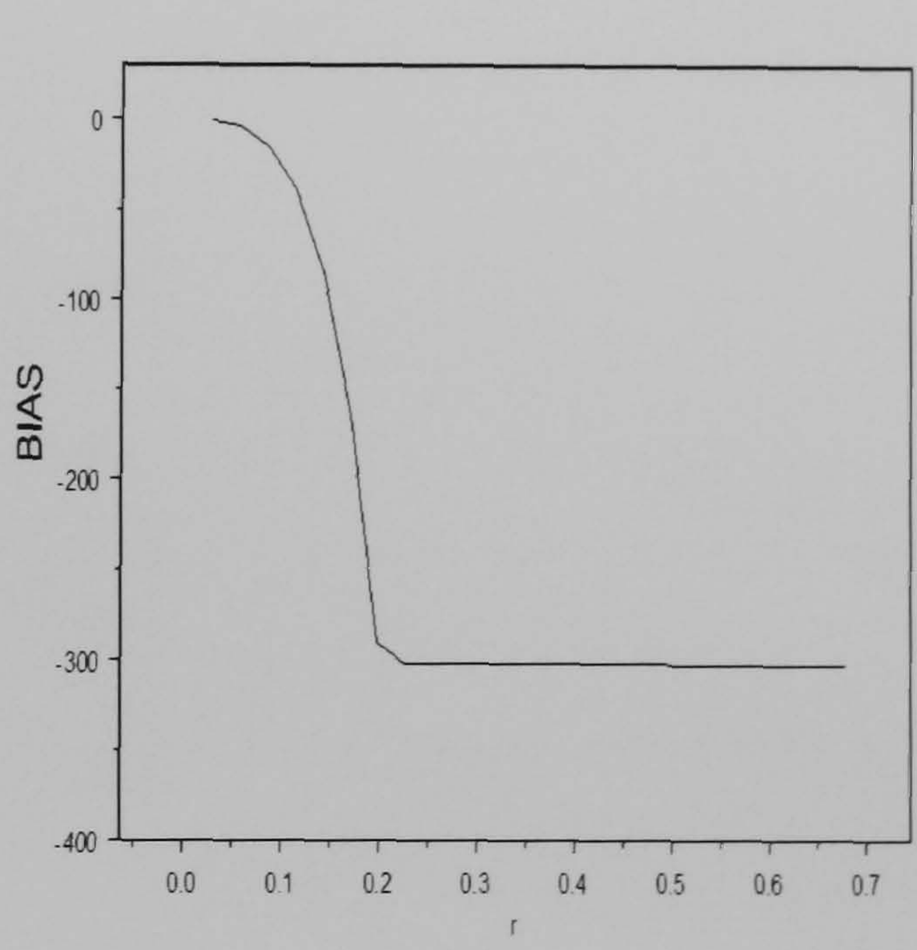
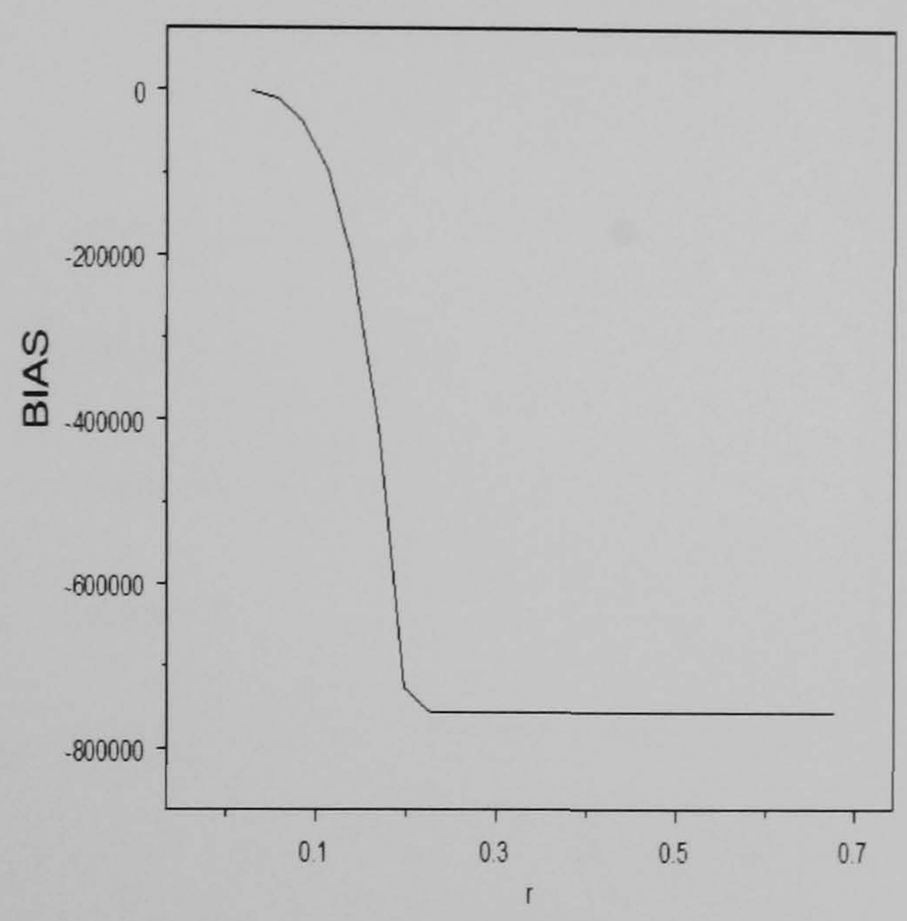
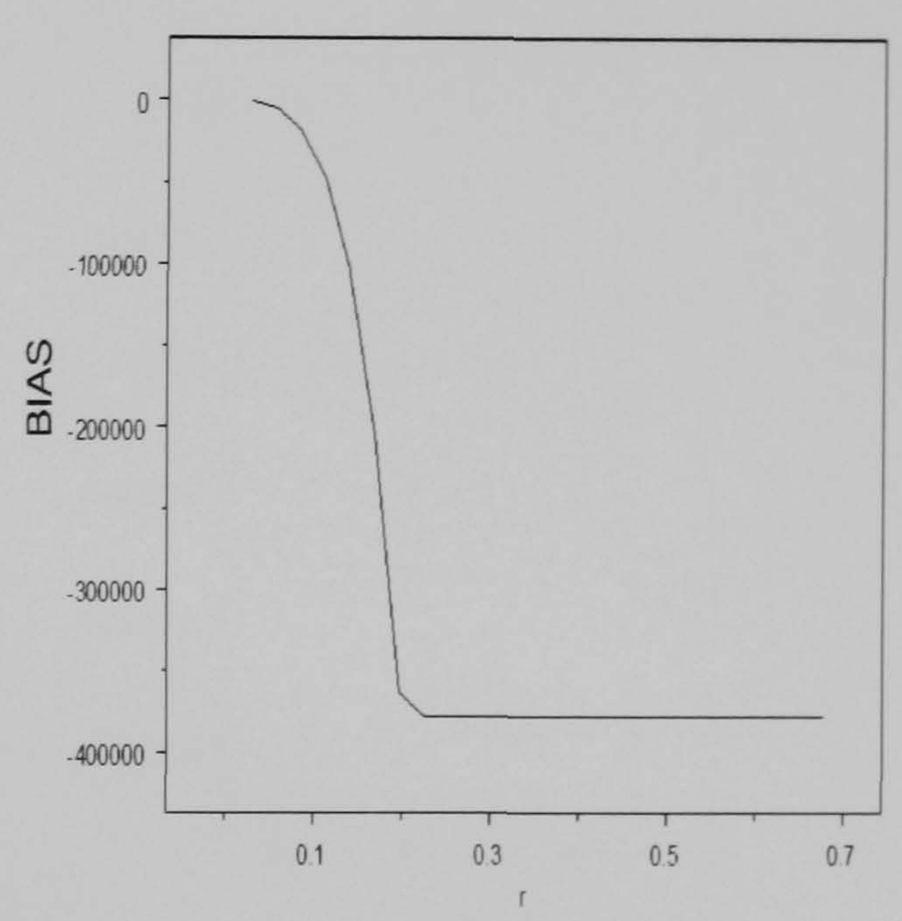
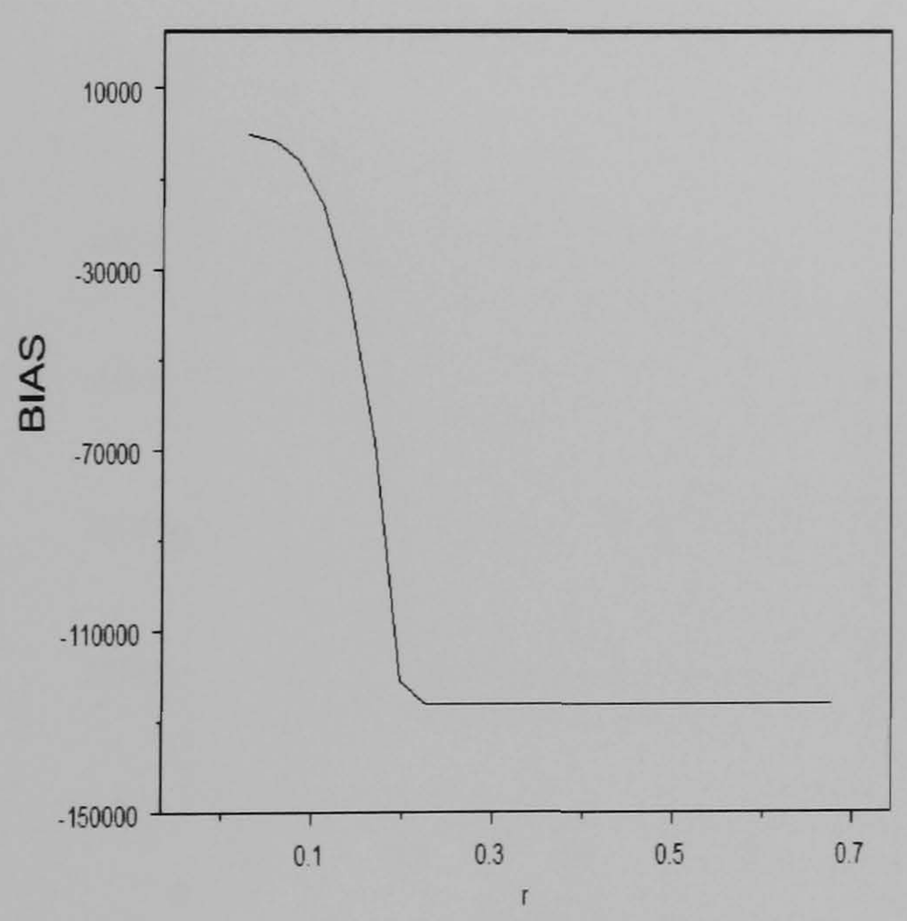


Figure E.4: $BIAS$ of the unedge-corrected K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) inside a cube of length 1







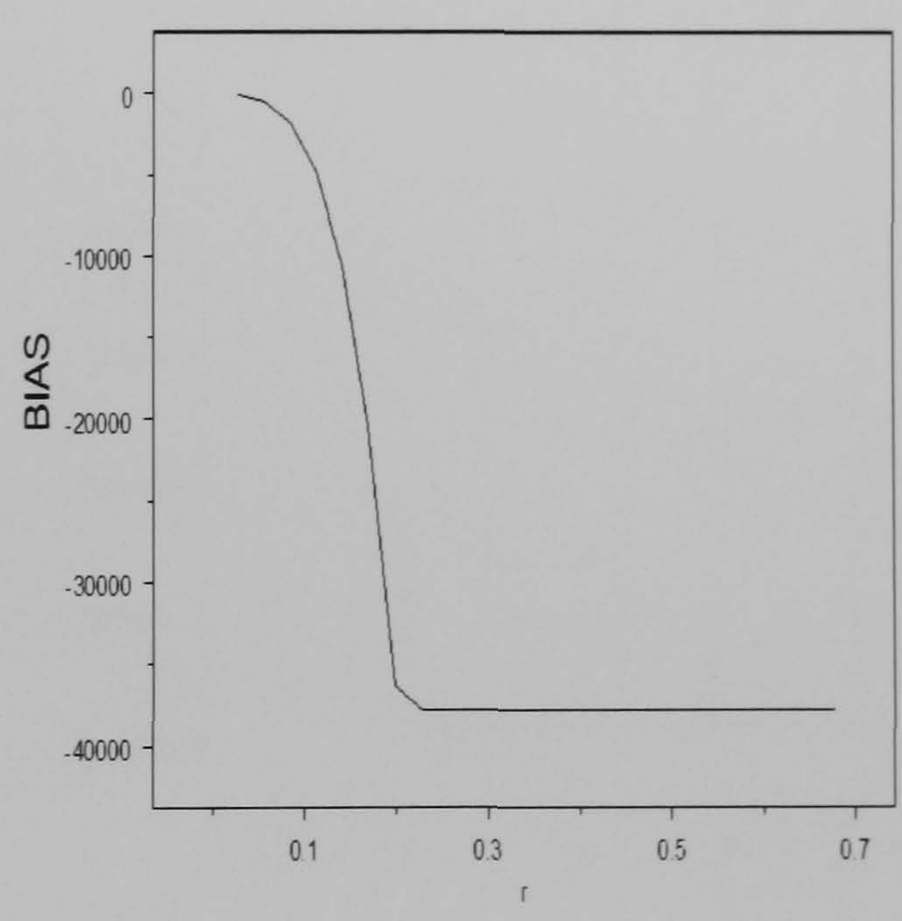


Figure E.5: $BIAS$ of the unedge-corrected K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) inside a cube of length 5

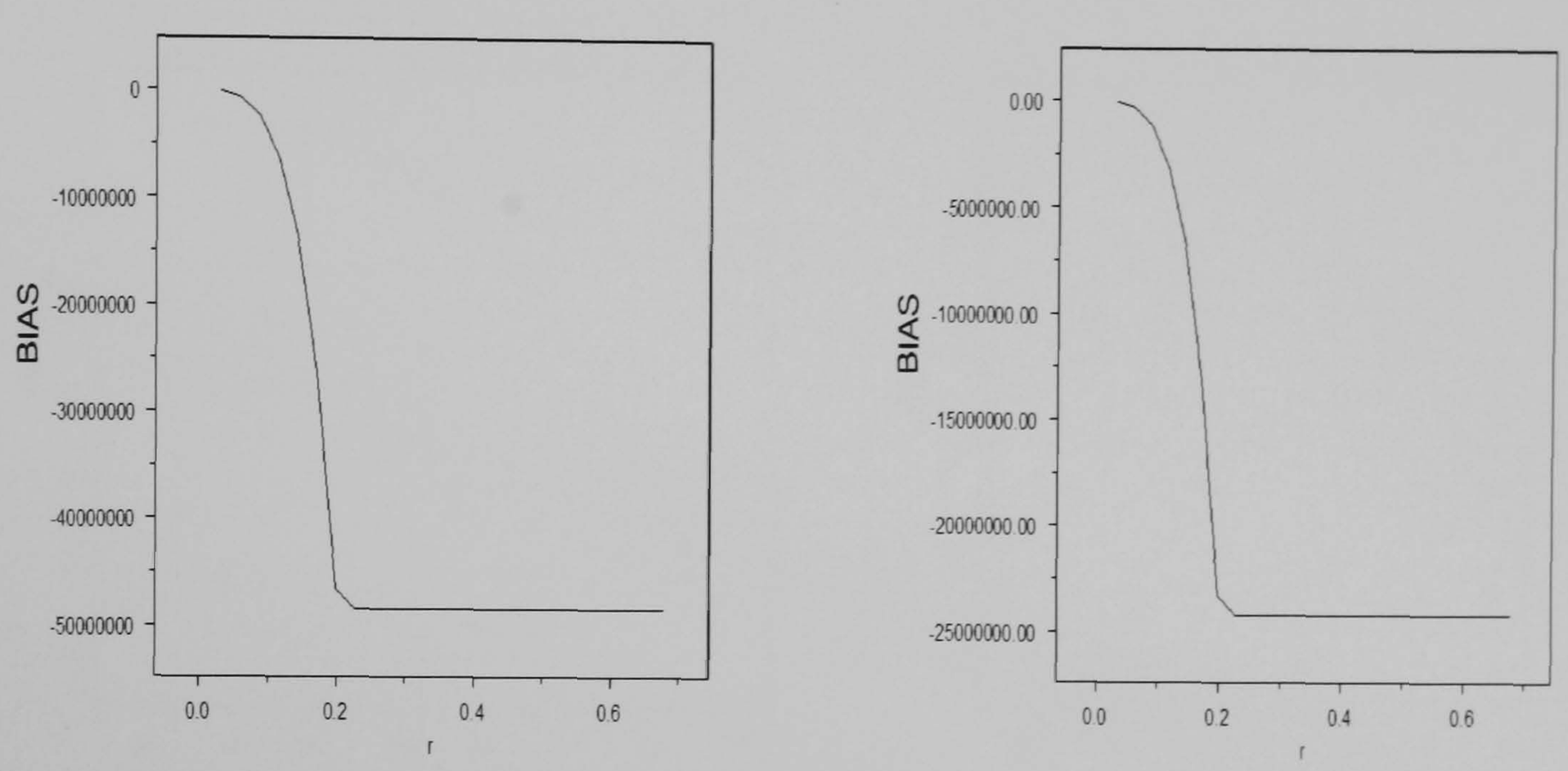
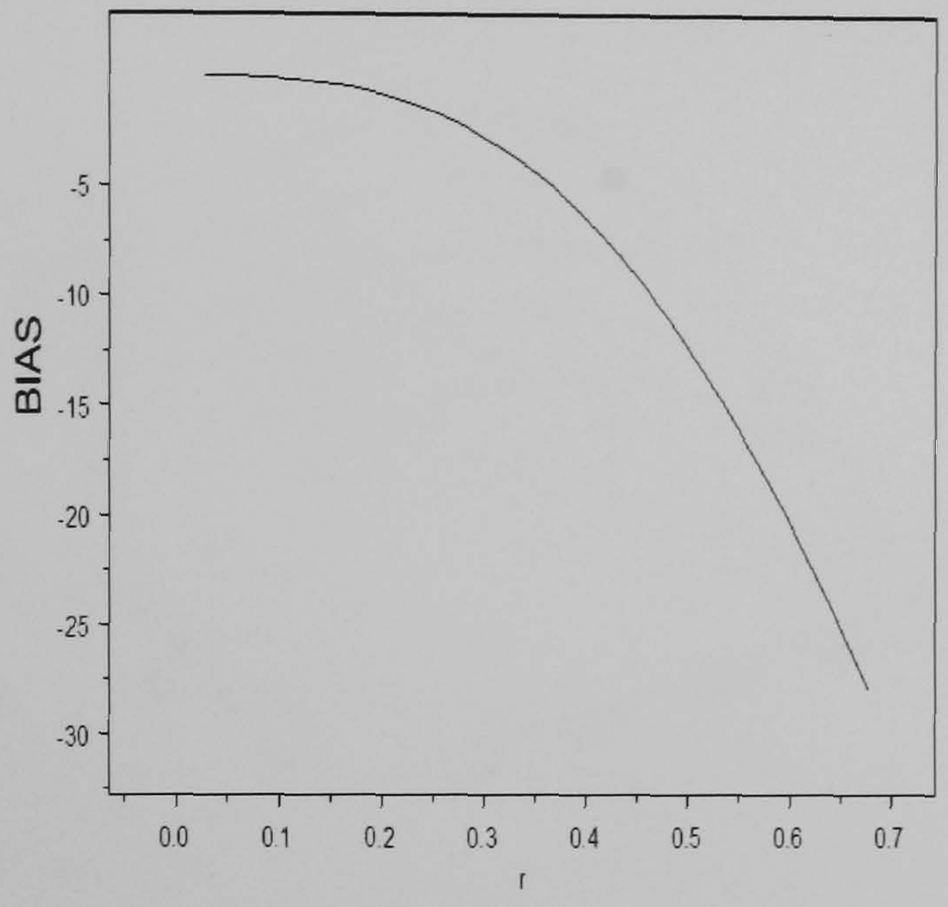
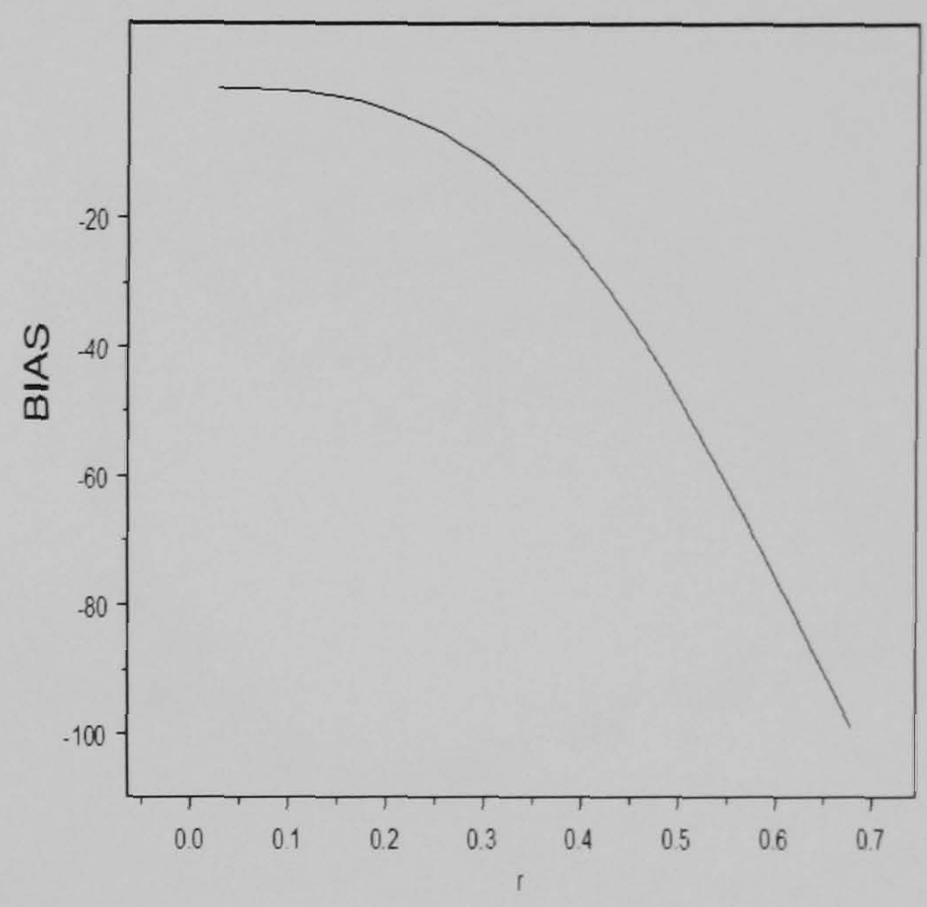
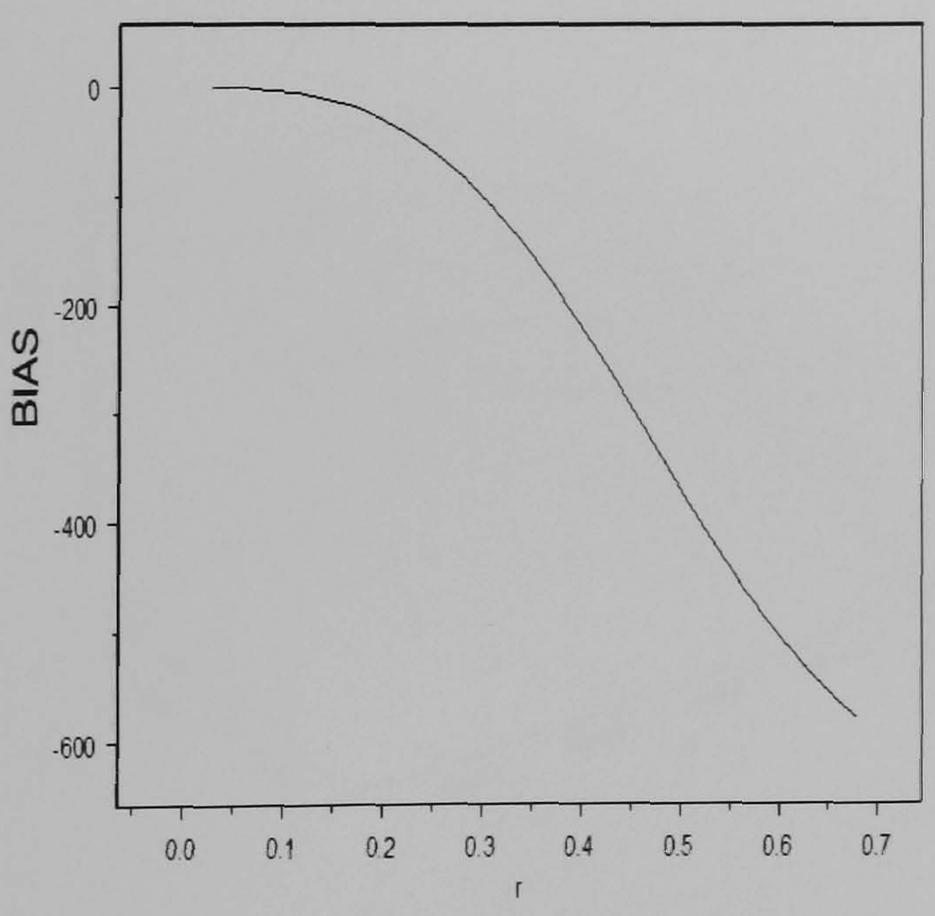


Figure E.6: $BIAS$ of the unedge-corrected K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) inside a cube of length 20







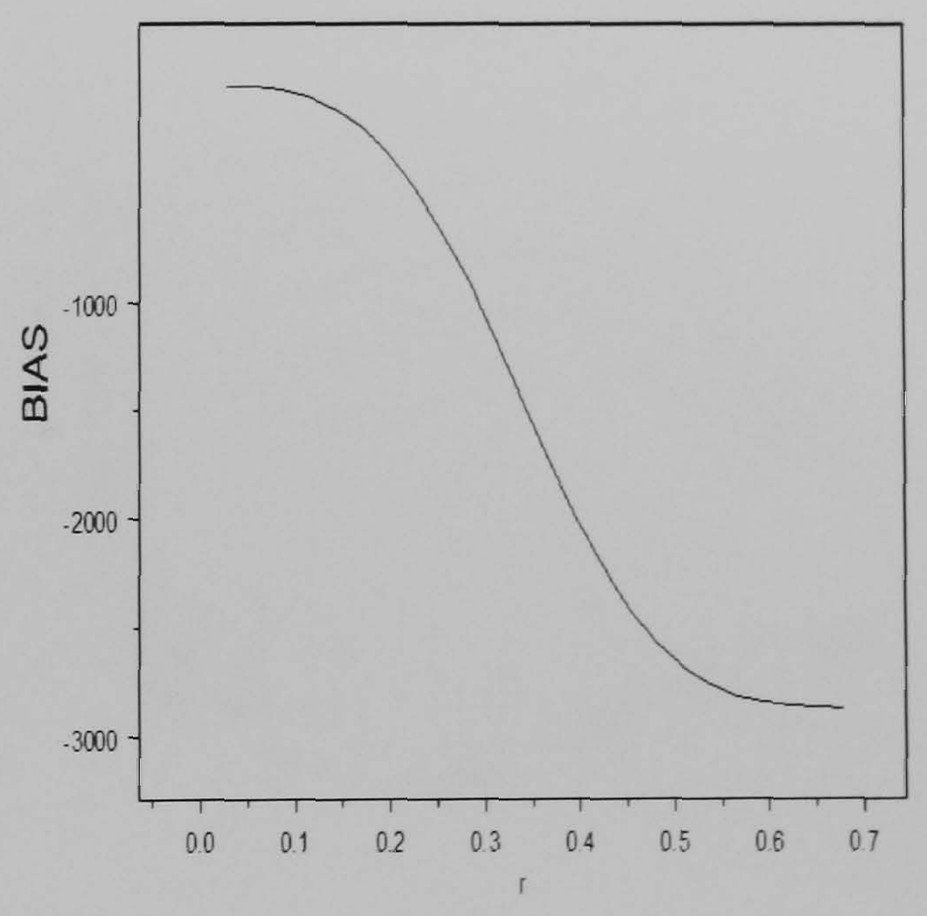


Figure E.7: $BIAS$ of the unedge-corrected K-function of a simple inhibition process (with underlying intensity = 5, 10, 30, 100 and $\delta = 0.1$) inside a cube of length 1

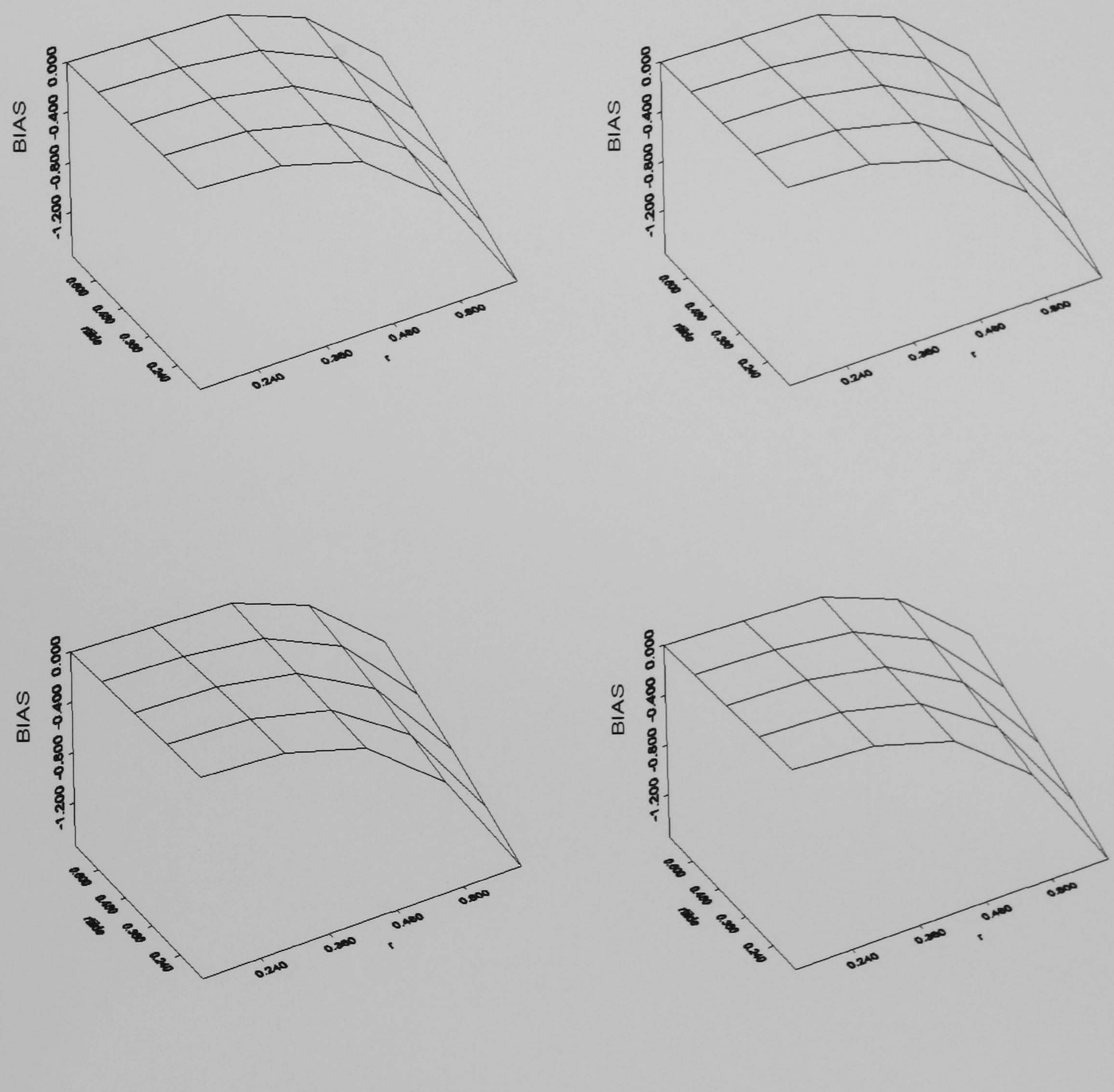


Figure E.8: $BIAS$ of the unedge-corrected Q-function of a homogeneous Poisson process inside a cube of length 1

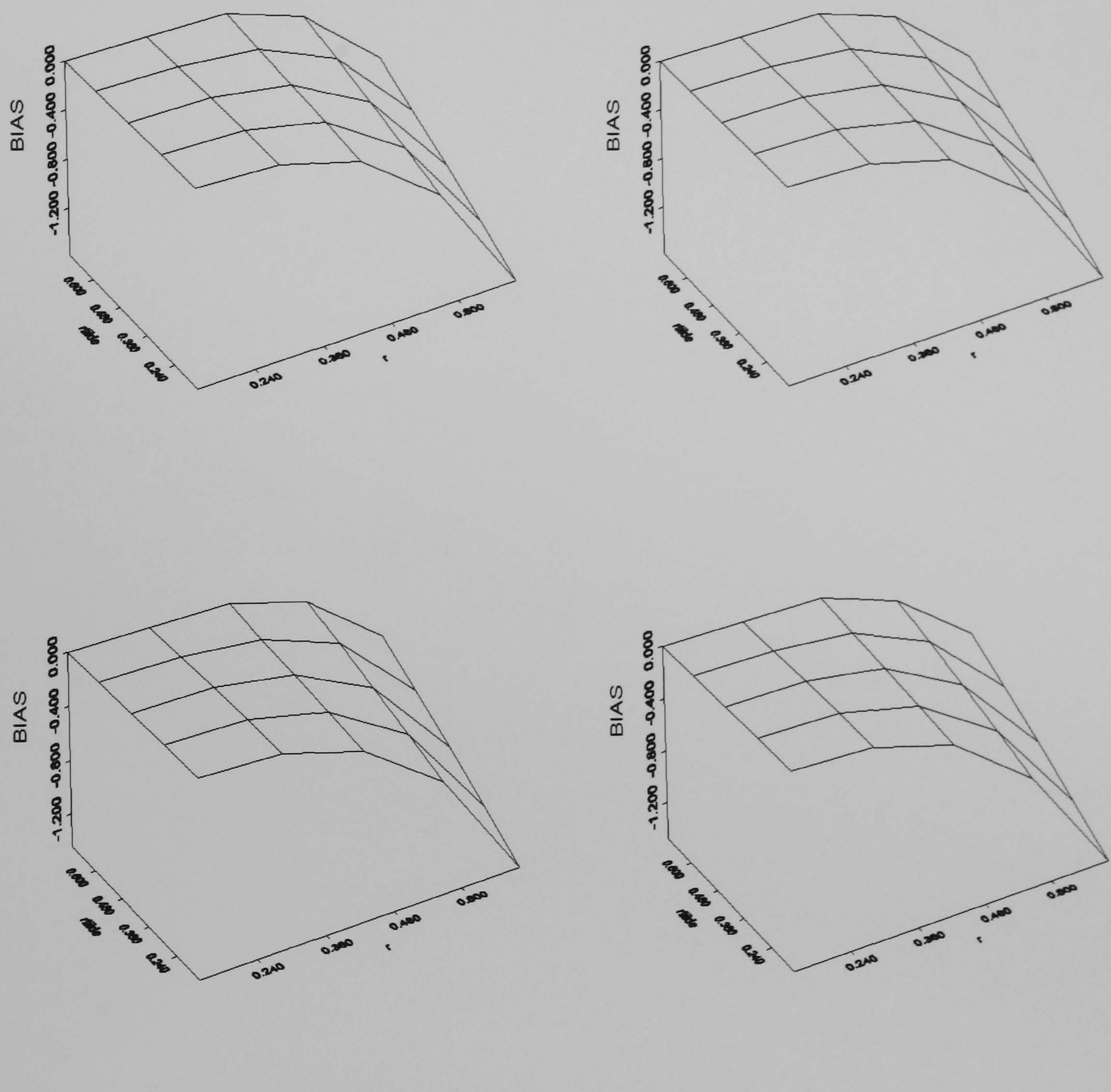


Figure E.9: $BIAS$ of the unedge-corrected Q-function of a homogeneous Poisson process inside a cube of length 5

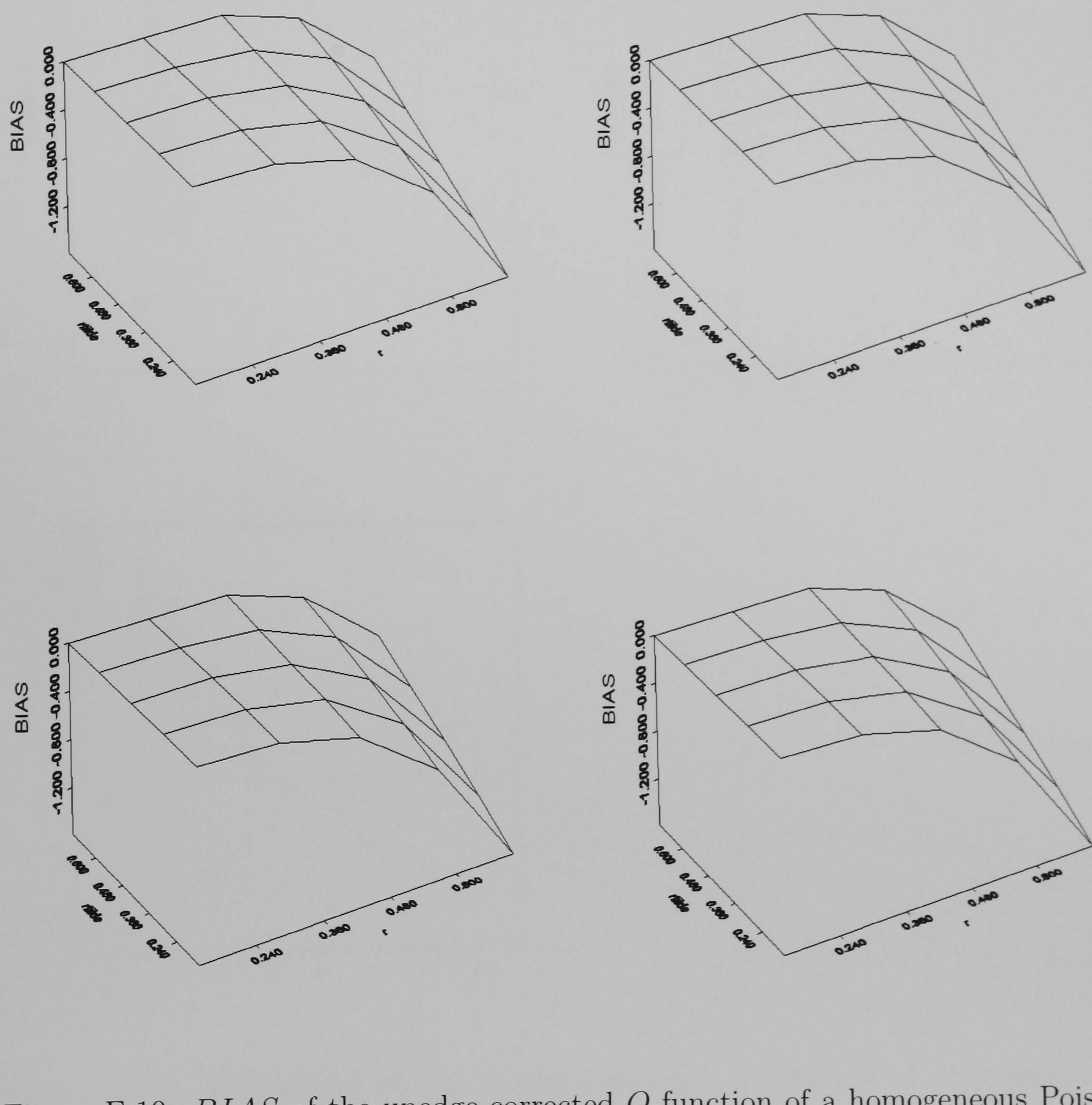
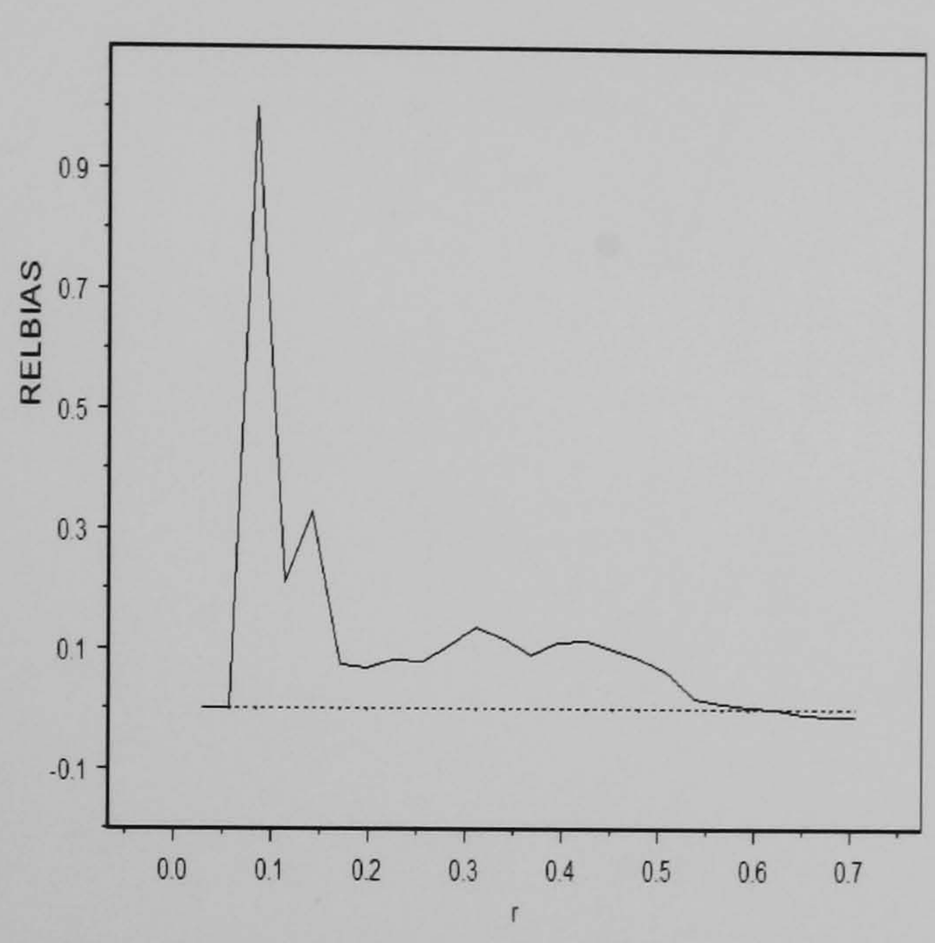
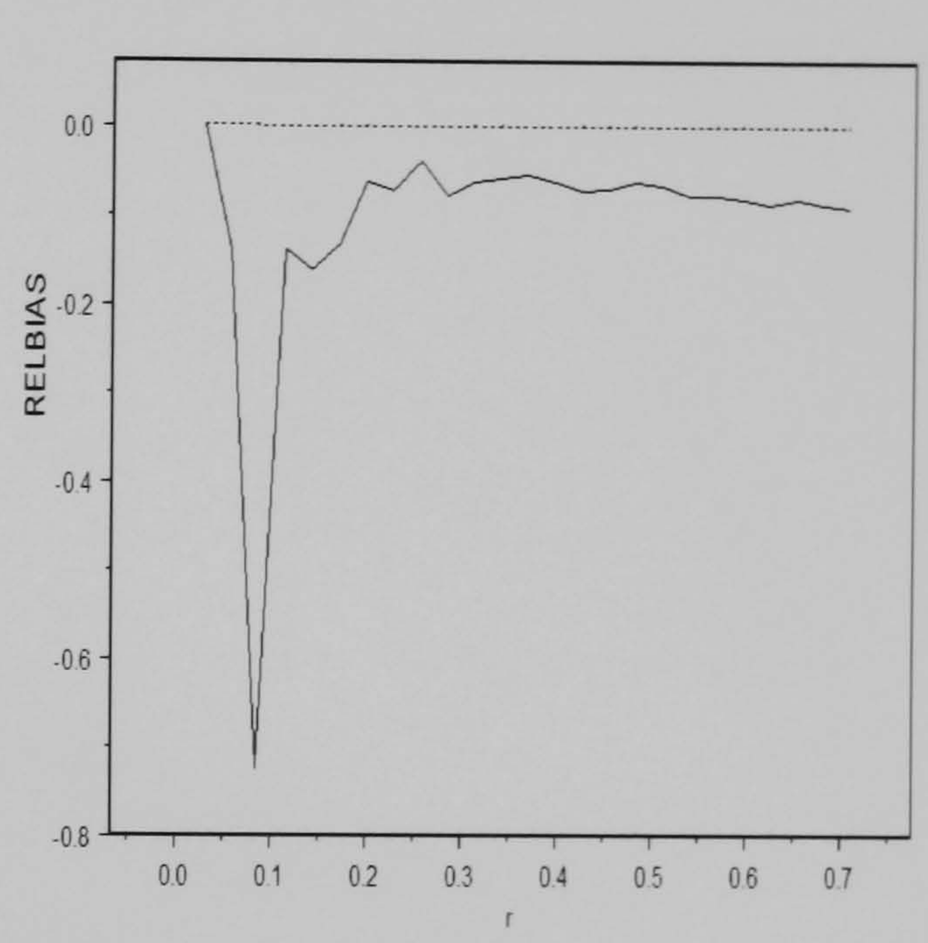


Figure E.10: $BIAS$ of the unedge-corrected Q-function of a homogeneous Poisson process inside a cube of length 20





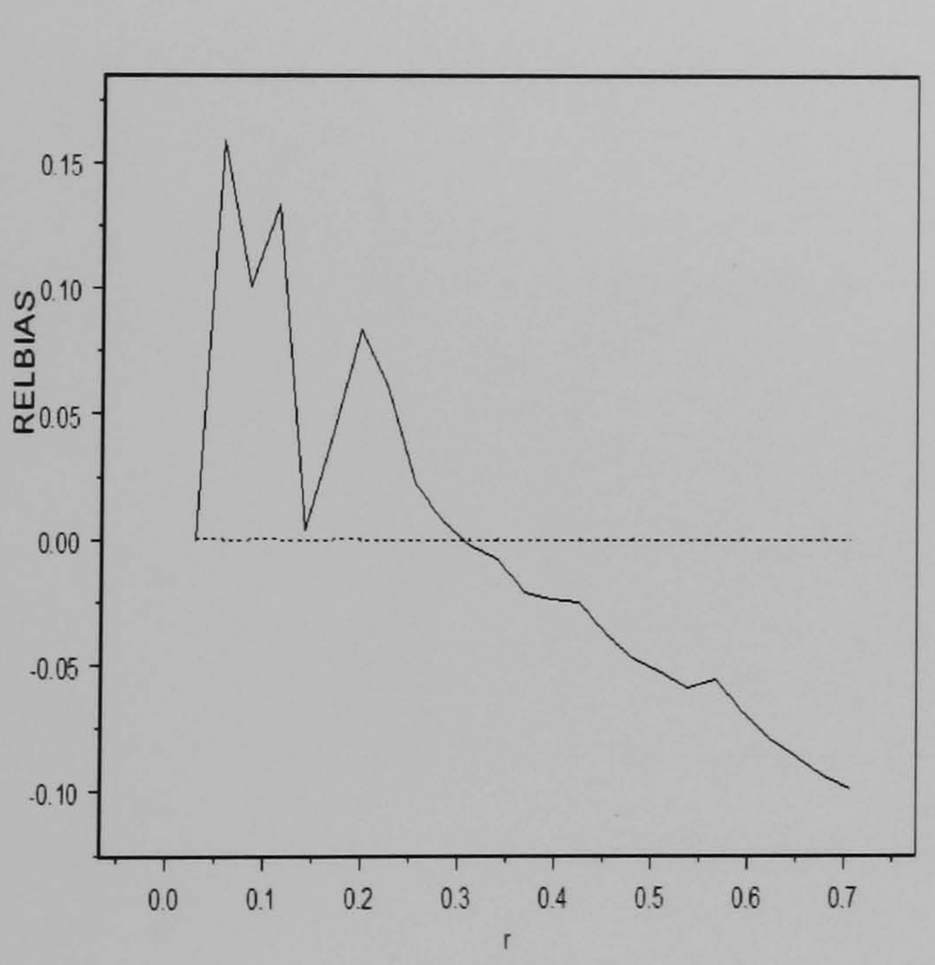
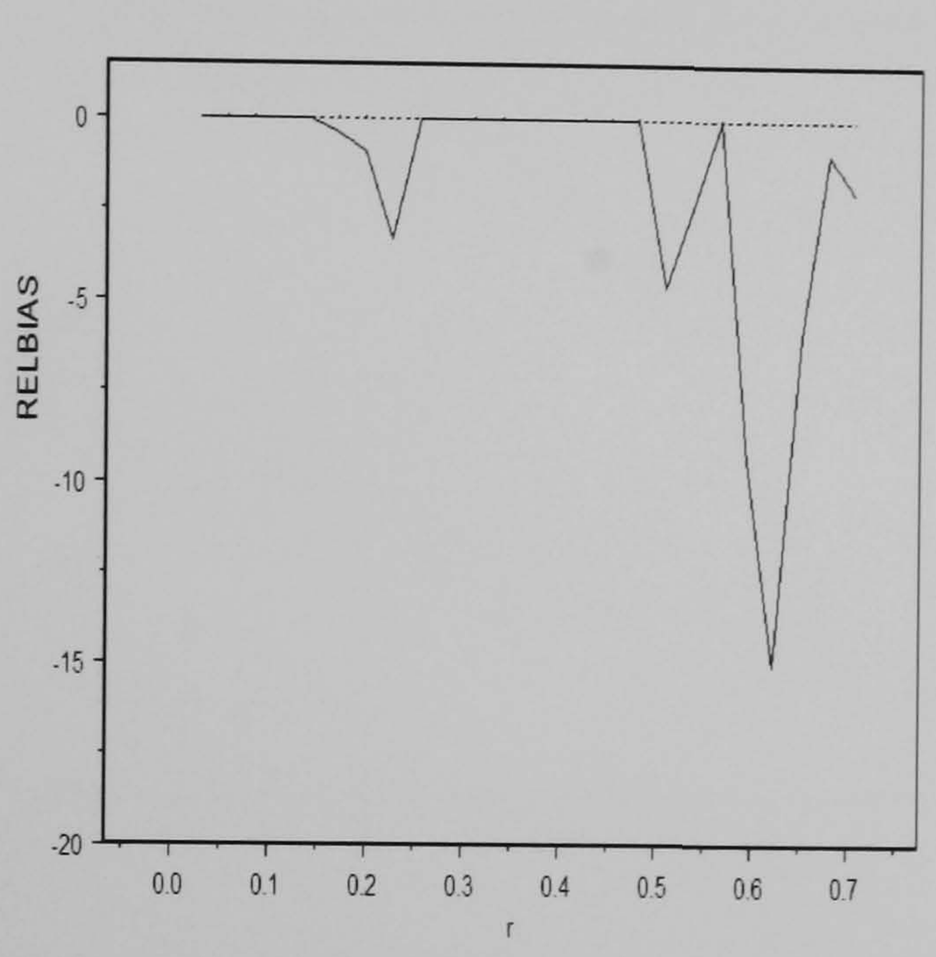
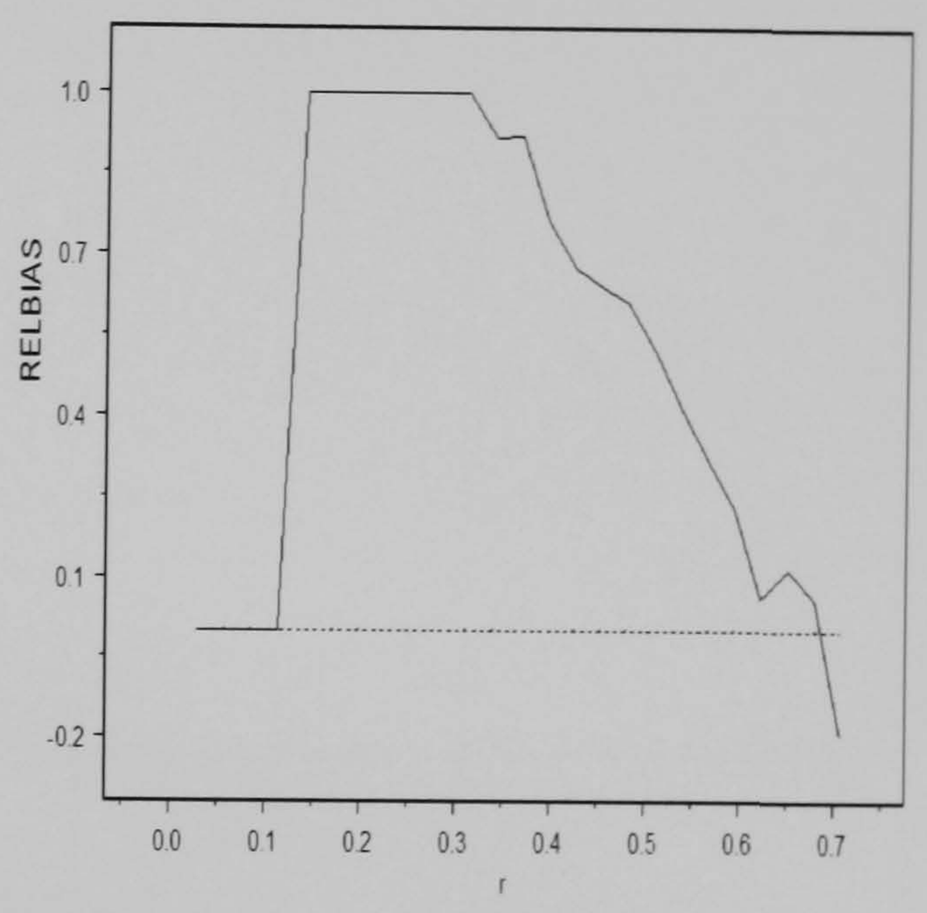


Figure E.11: *RELBIAS* of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 90, 150, 300$ conditional number of events) inside a cube (black line) and square (dashed line) of length 5





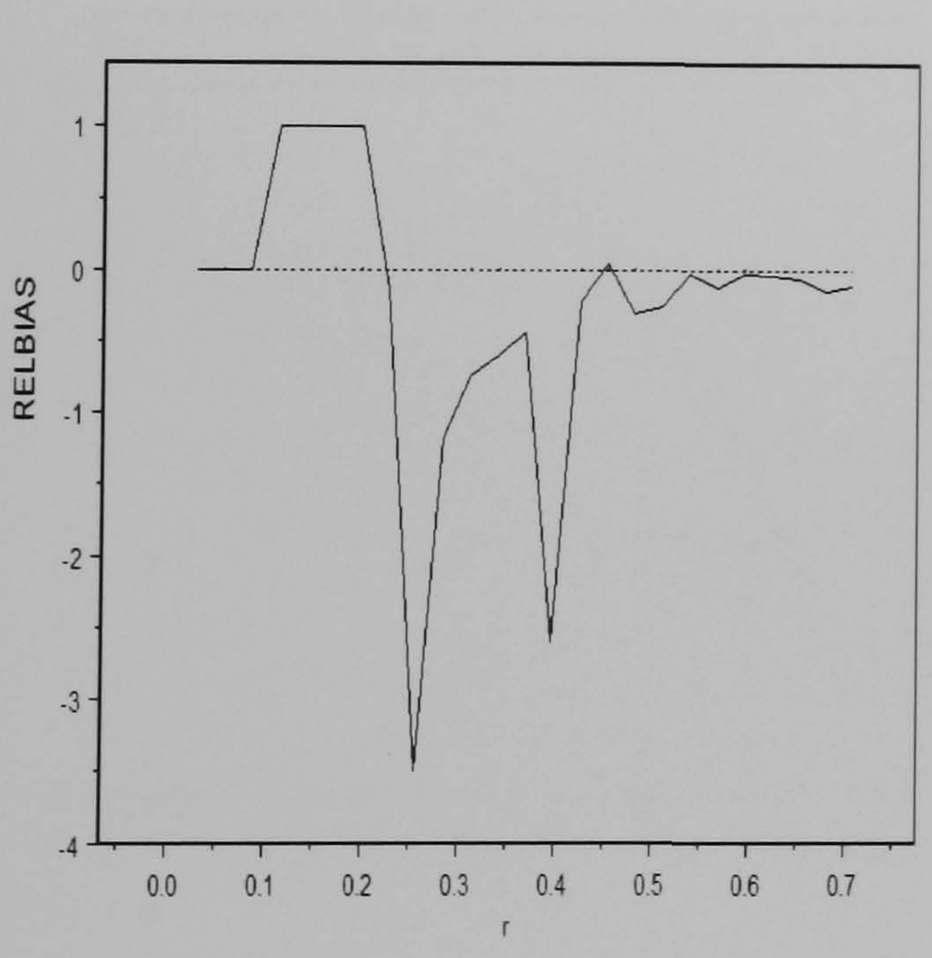
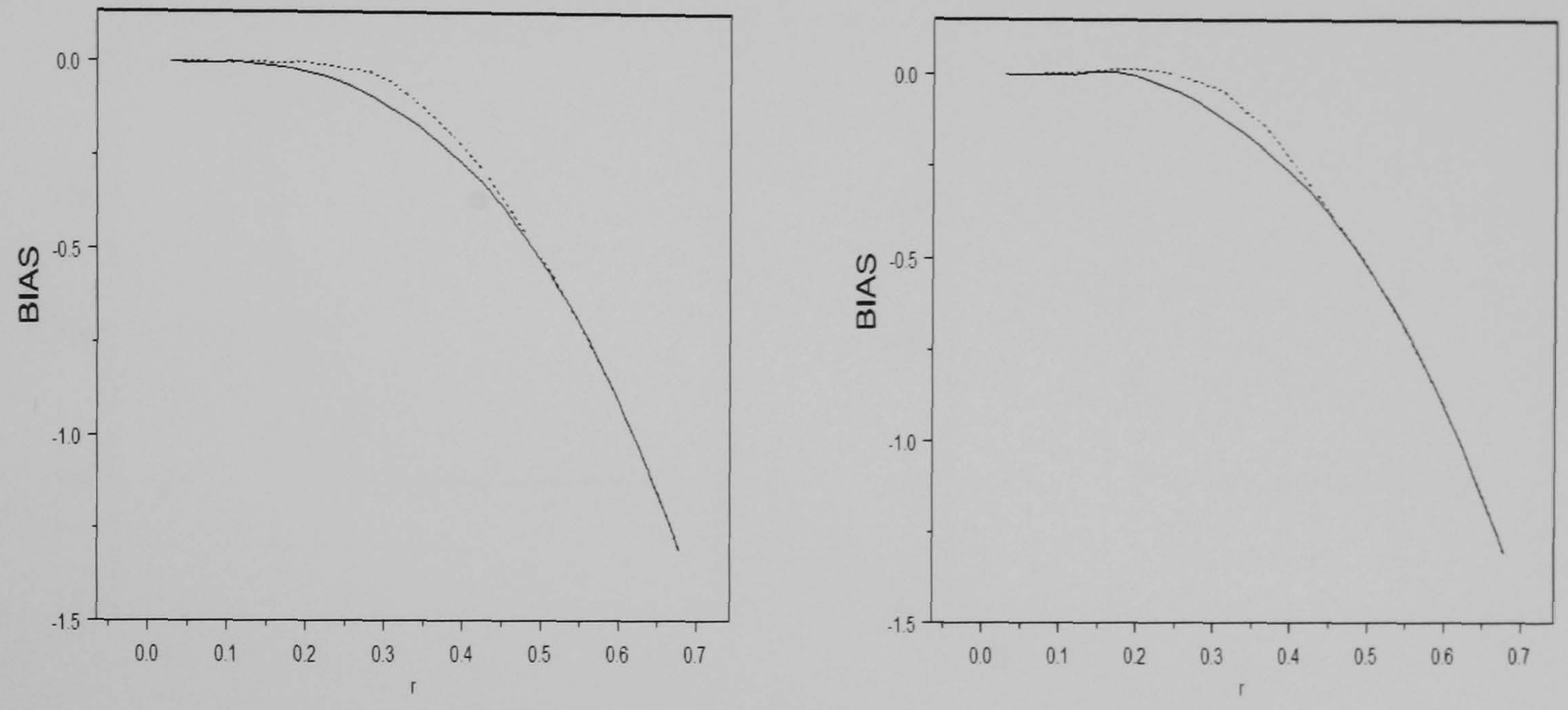


Figure E.12: *RELBIAS* of the unedge-corrected K-function of a homogeneous Poisson process (with $n = 90$, 150, 300 conditional number of events) inside a cube (black line) and square (dashed line) of length 20



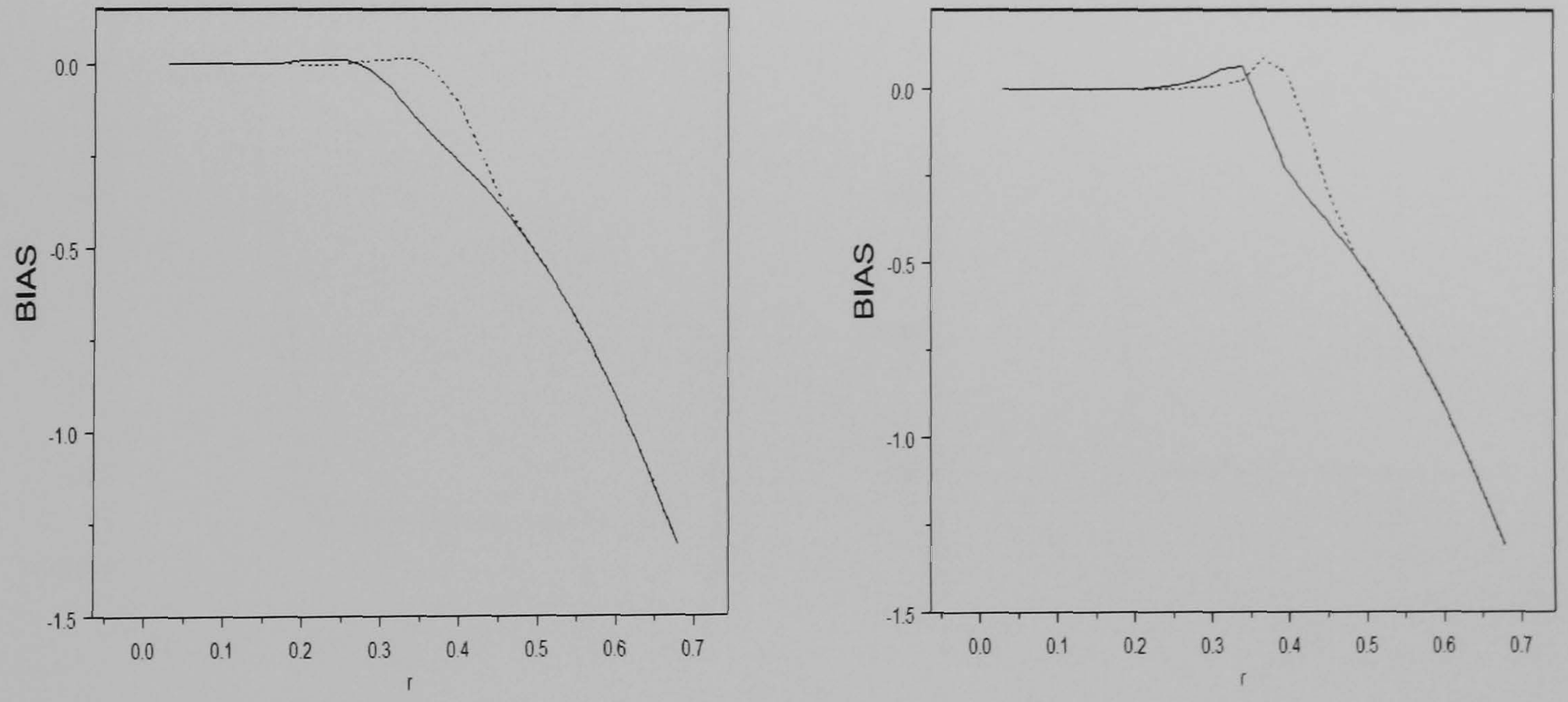
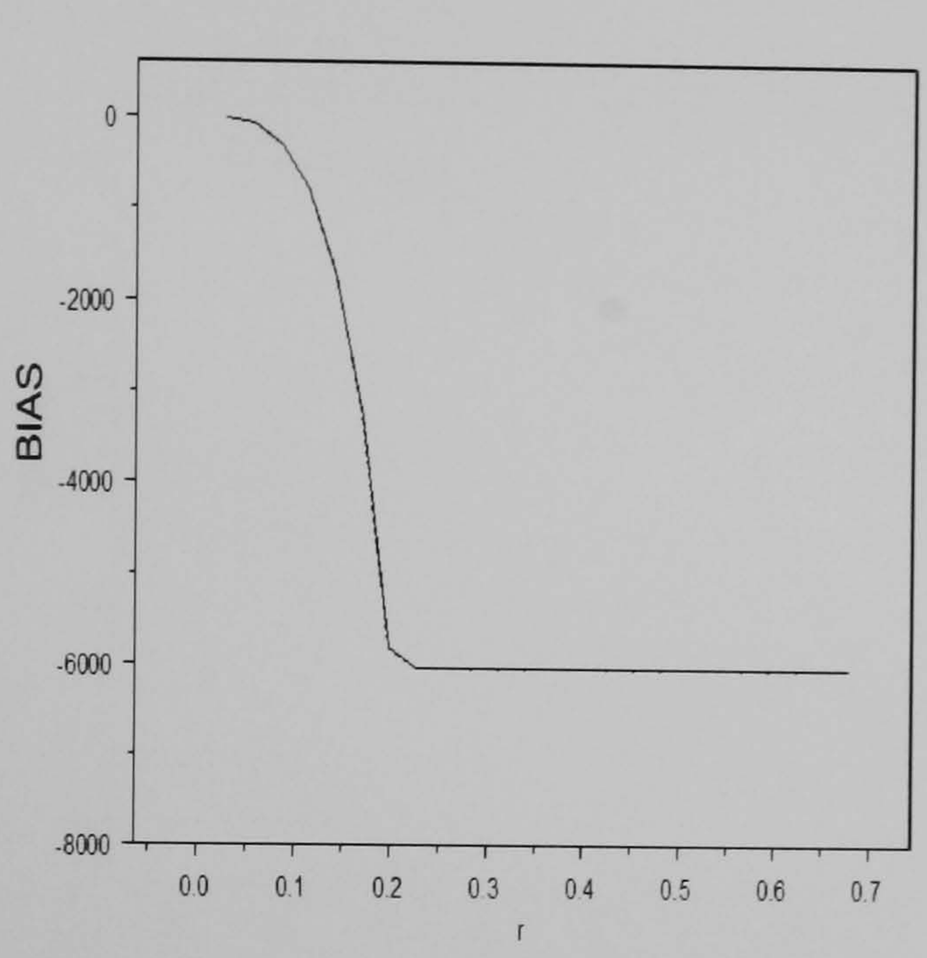
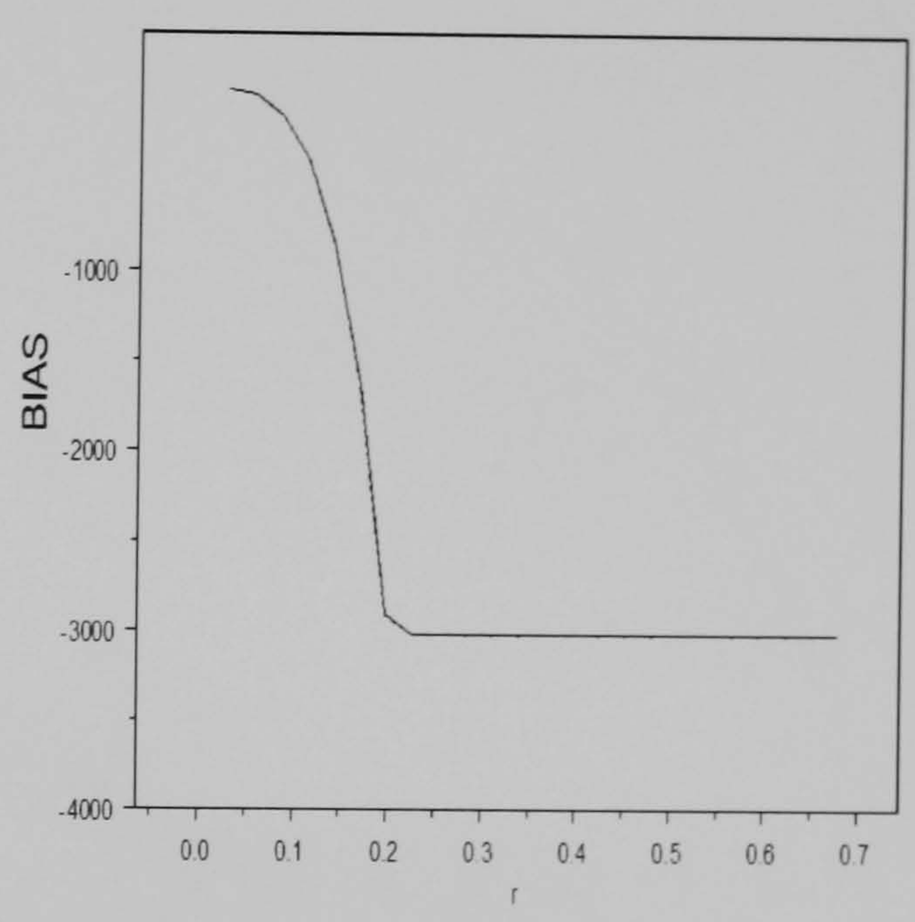
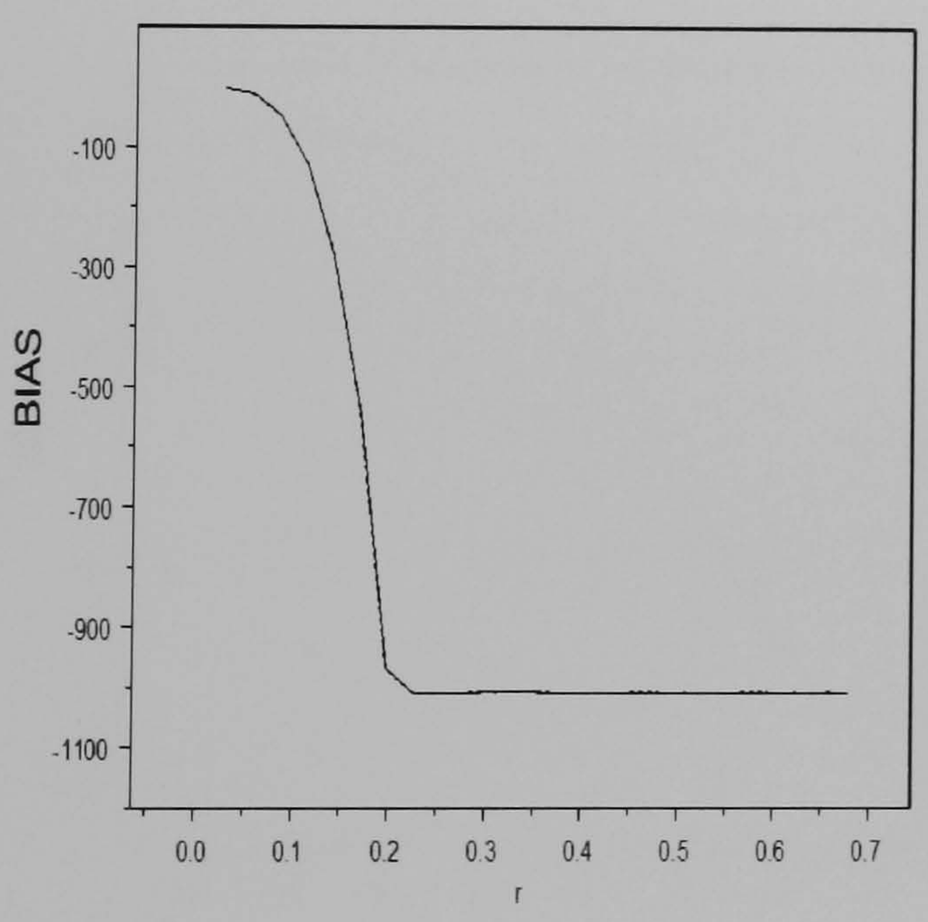


Figure E.13: $BIAS$ of the K-function of a homogeneous Poisson process (with n = 5, 10, 30, 100 conditional number of events) estimated using the border method (black line) and probabilistic border method (dashed curve) inside a cube of length 1







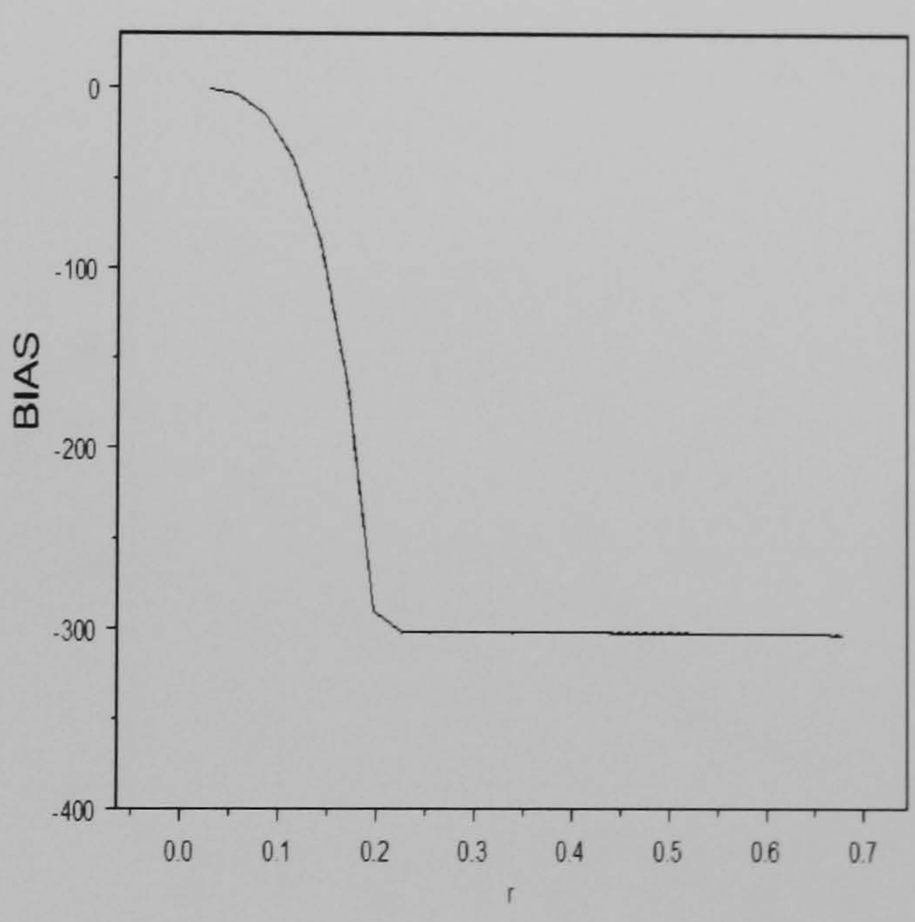
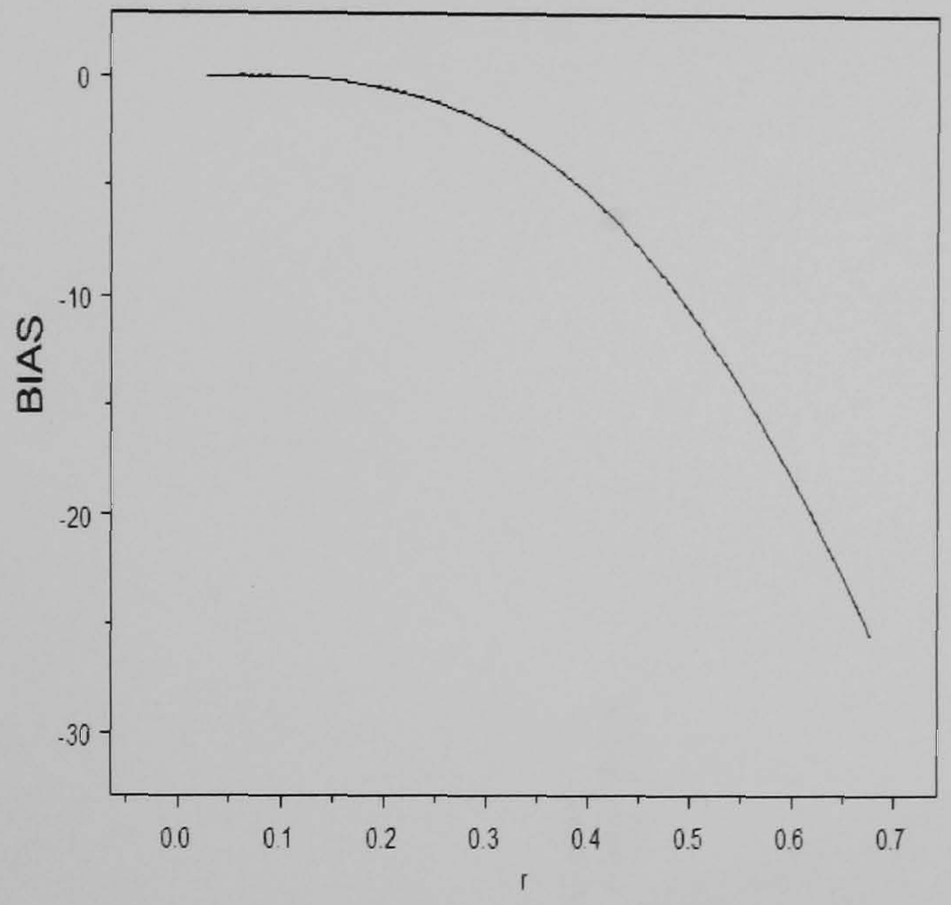
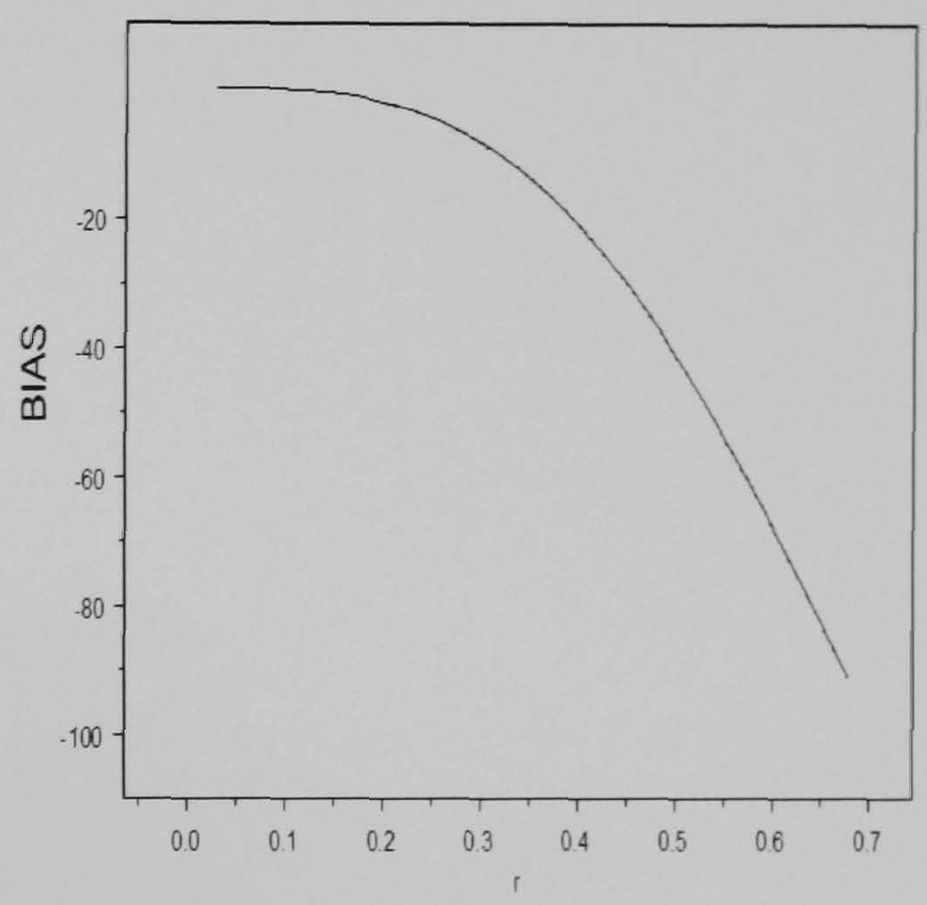
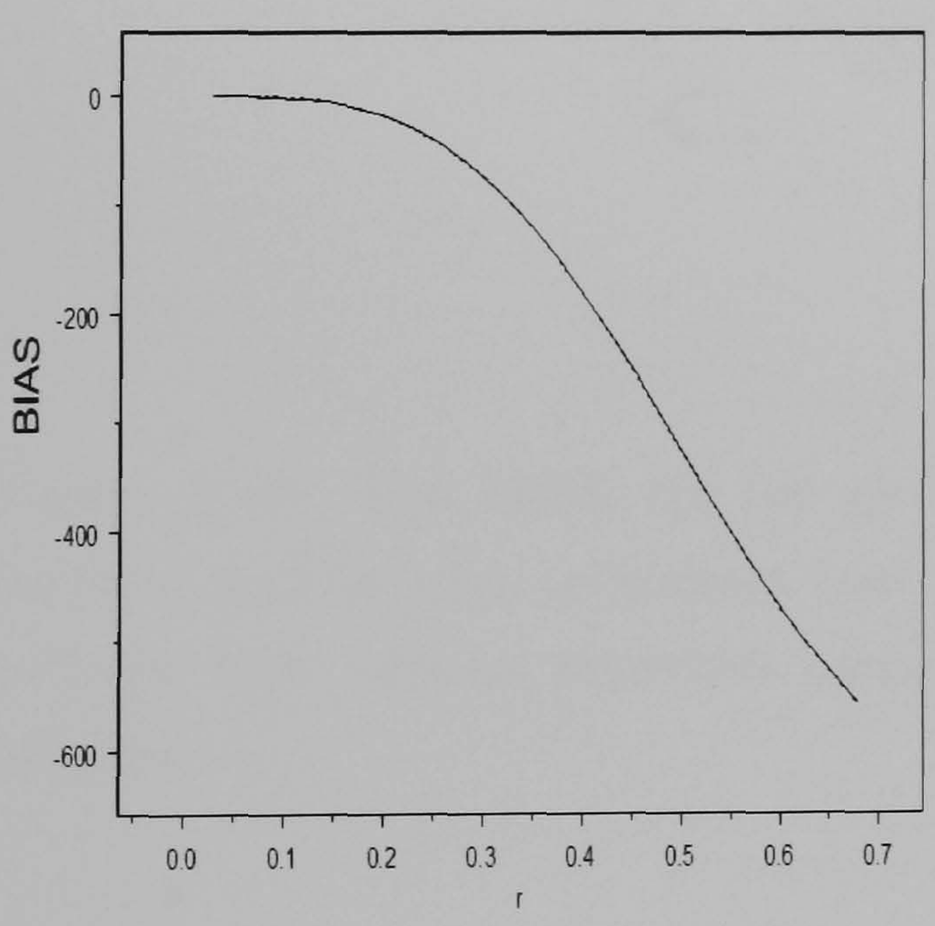


Figure E.14: $BIAS$ of the K-function of a Matérn cluster process (with $n = 5$, 10, 30, 100 conditional number of parent events and $R = 0.1$) estimated using the border method (black line) and probabilistic border method (dashed curve) inside a cube of length 1







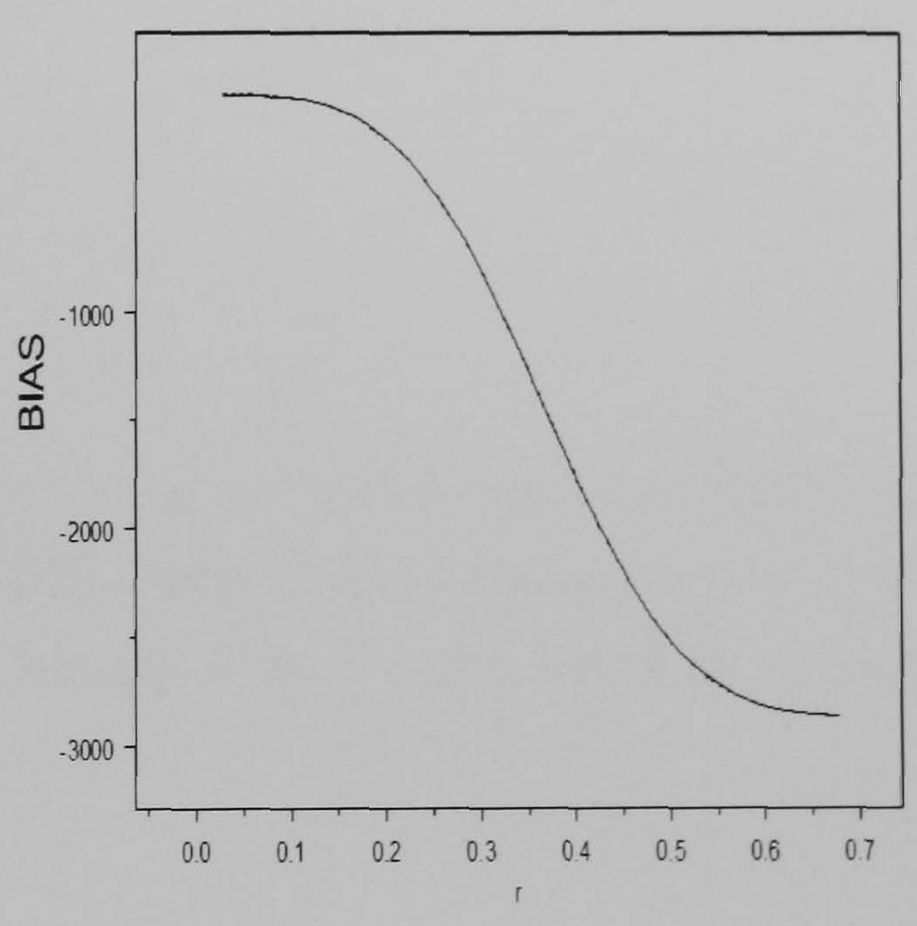


Figure E.15: $BIAS$ of the K-function of a simple inhibition process (with underlying intensity = 5, 10, 30, 100 and $\delta = 0.1$) estimated using the border method (black line) and probabilistic border method (dashed curve) inside a cube of length 1

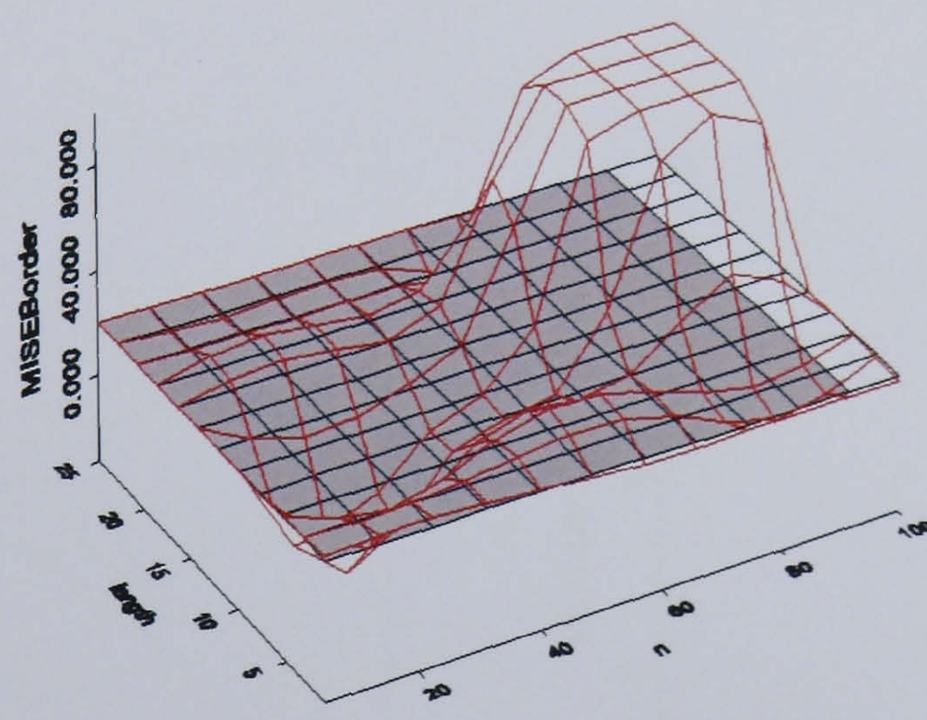
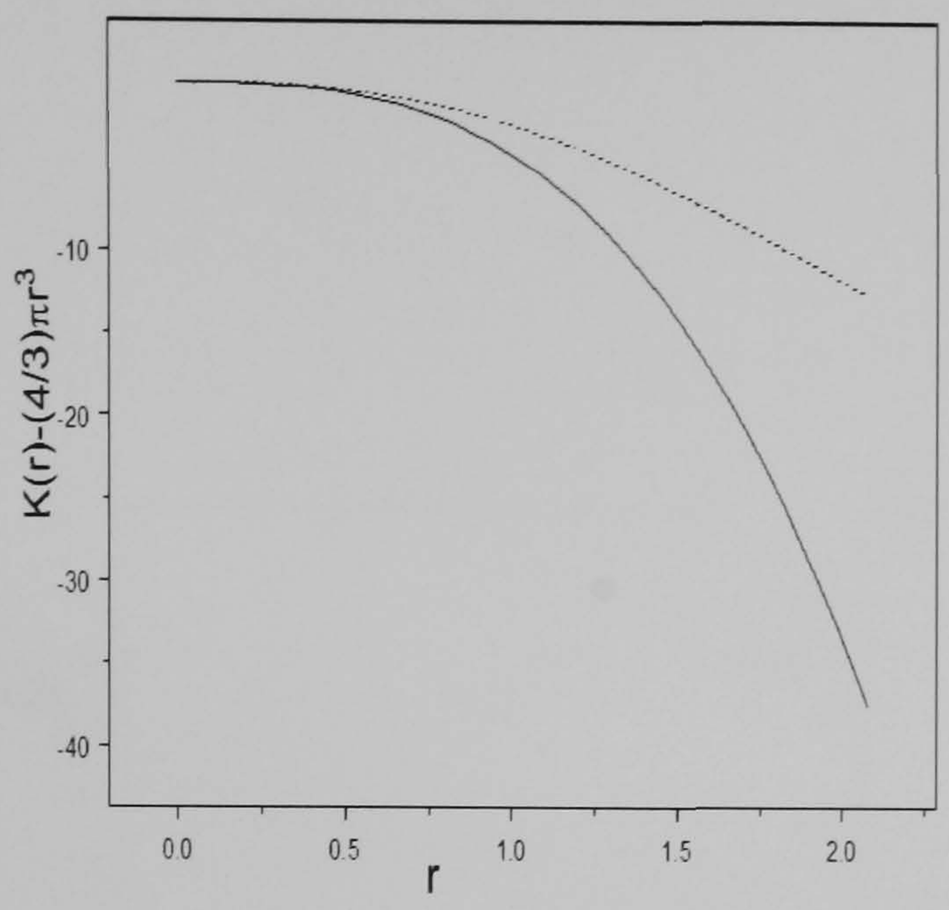
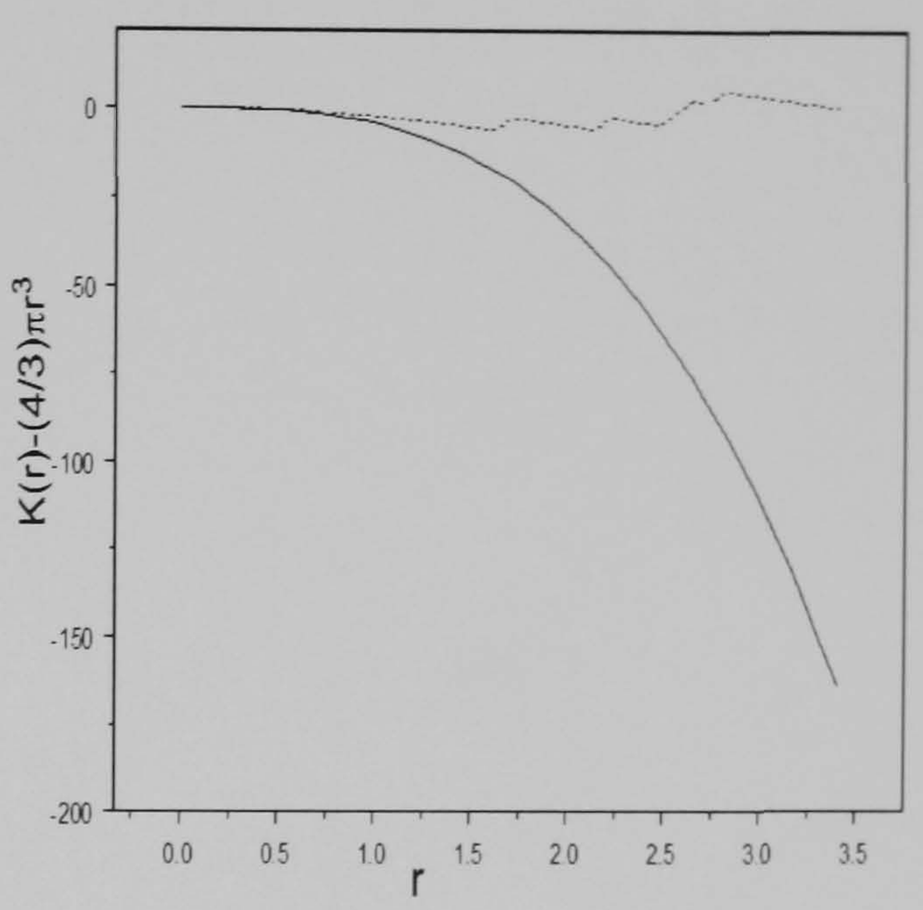
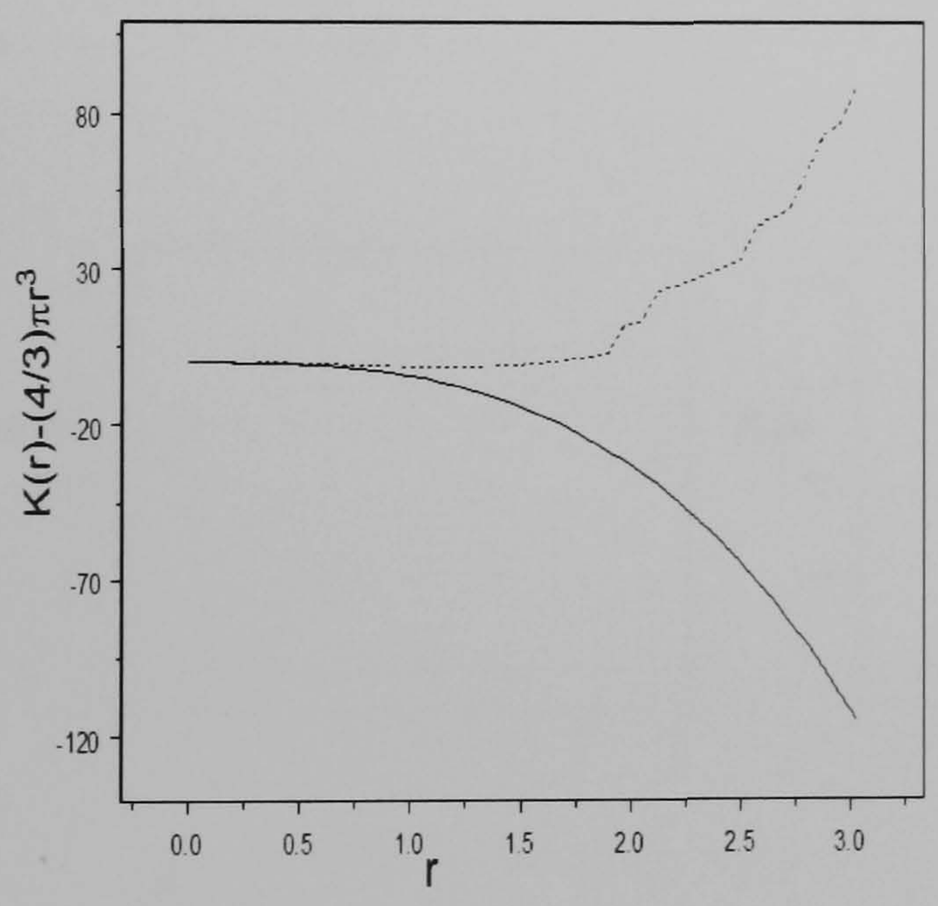
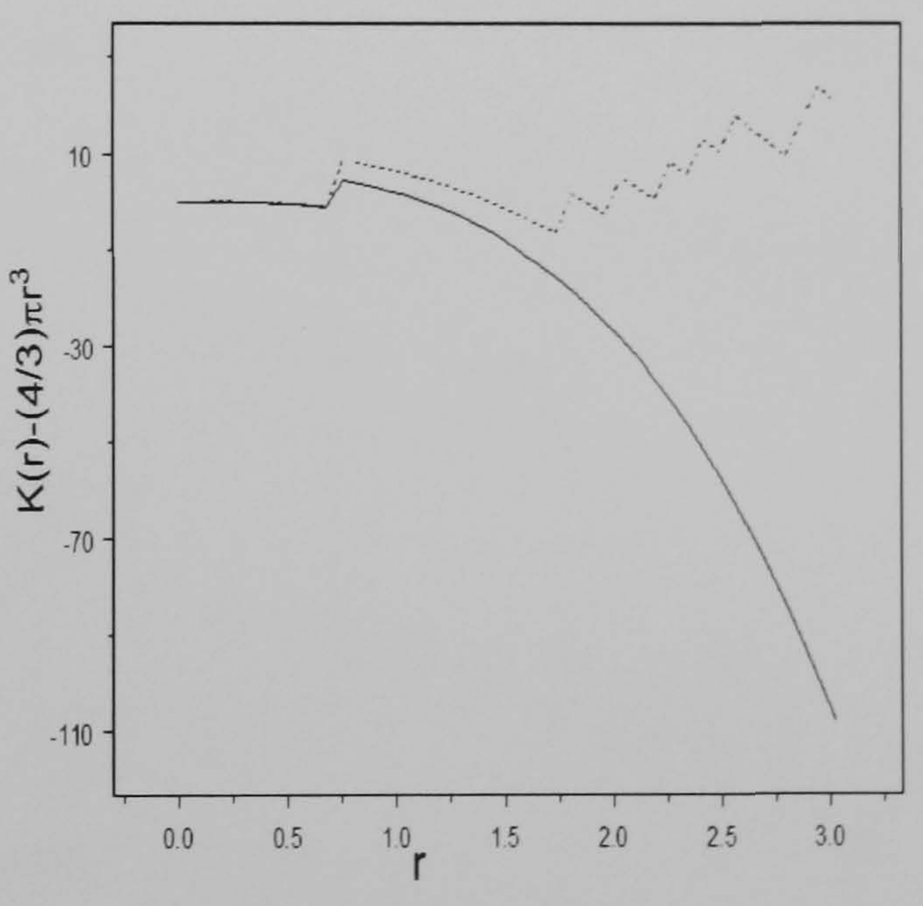


Figure E.16: The MISE for the Q-function using a border edge correction (grey surface) and no edge correction (red non-filled surface) for a homogeneous Poisson process with various expected number of events n in a cube window of various lengths.









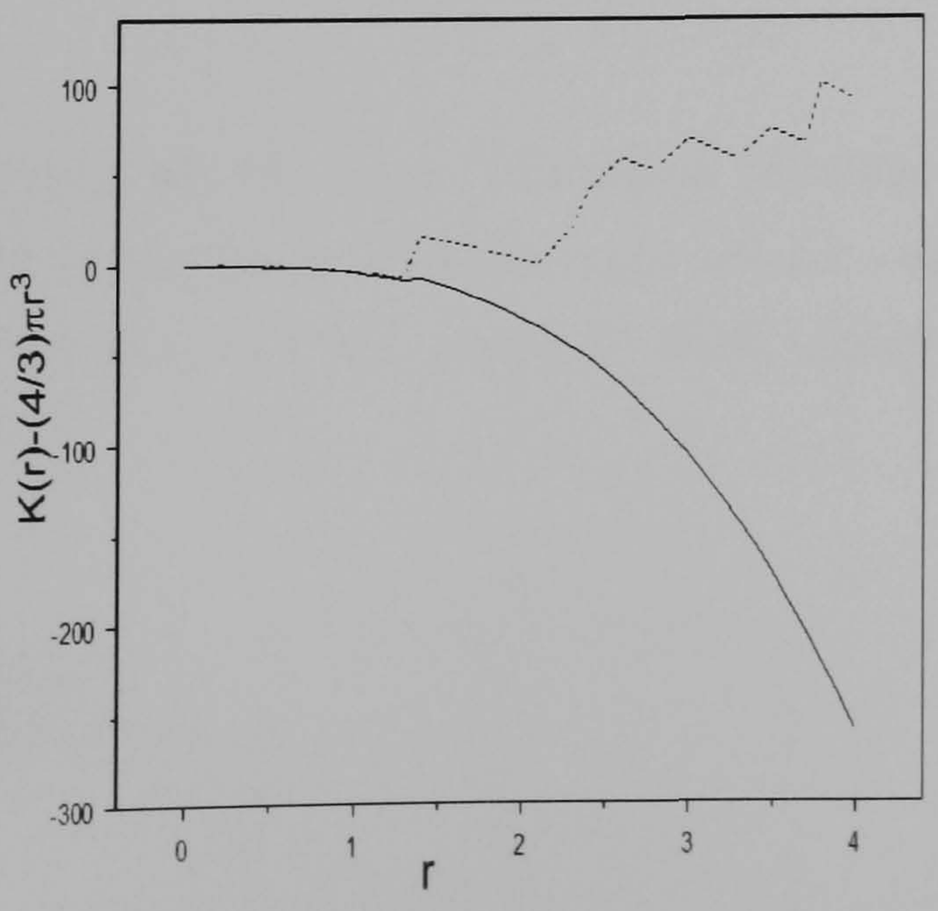


Figure E.17: Plots of $K(r) - \frac{4}{3}\pi r^3$ for the PML NBs in the cells that make up PPDS1, [illegible] ighting edge correction (black line) and probabilistic [illegible] urve).

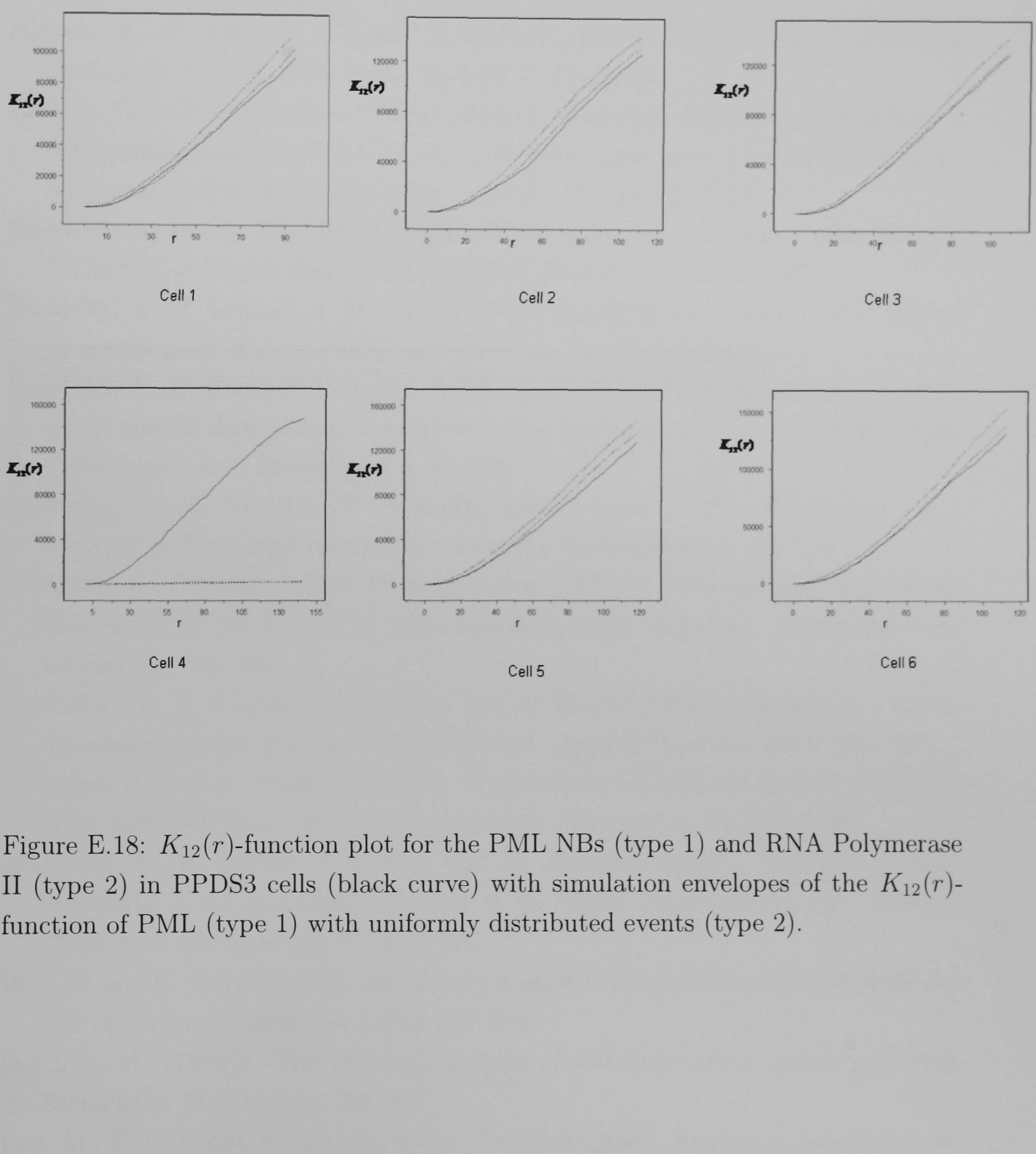


Figure E.18: $K_{12}(r)$-function plot for the PML NBs (type 1) and RNA Polymerase II (type 2) in PPDS3 cells (black curve) with simulation envelopes of the $K_{12}(r)$-function of PML (type 1) with uniformly distributed events (type 2).

Bibliography

Agard, D. and J. Sedat (1983). Three-dimensional architecture of a polytene nucleus. *Nature 302*(21), 676–681.

Alberts, B., A. Johnson, J. Lewis, K. Roberts, and P. Walter (2002). *Molecular Biology of the Cell.* New York: Garland Publications.

Ambler, G. and B. Silverman (2004). Perfect simulation of spatial point processes using dominated coupling from the past with application to a multiscale area-interaction point process. Preprint. *Applied Probability Trust.*

Baddeley, A. and R. Gill (1997). Kaplan-Meier estimators of distance distributions for spatial point processes. *The Annals of Statistics 25*(1), 263–292.

Baddeley, A., P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (2006). *Case studies in spatial point process modeling.* New York: Springer Science.

Baddeley, A., C. Howard, A. Boyde, and S. Reid (1987). Three-dimensional analysis of the spatial distribution of particles using the tandem-scanning reflected light microscope. *Acta Stereologica 6*, 87–100.

Baddeley, A., M. Kerscher, K. Schladitz, and B. Scott (2000). Estimating the J function without edge correction. *Statistica Neerlandica 54*, 315–328.

Baddeley, A., J. Møller, and R. Waagepetersen (2000(b)). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica 54*, 329–350.

Baddeley, A., R. Moyeed, C. Howard, and A. Boyde (1993). Analysis of a three-dimensional point pattern with replication. *Applied Statistics 42*(4), 641–668.

Baddeley, A. and B. Silverman (1984). A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics 40*, 1089–1093.

Baddeley, A., R. Turner, J. Møller, and M. Hazelton (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(5), 1–35.

Barr, D. and E. Sherrill (1999). Mean and variance of truncated normal distributions. *The American Statistician 53*(4), 357–361.

Bartlett, M. (1964). The spectral analysis of two-dimensional point processes. *Biometrika 51*(3 and 4), 299–311.

Beil, M., F. Fleischer, s. Paschke, and V. Schmidt (2005). Statistical analysis of the three-dimensional structure of centromeric heterochromatin in interphase nuclei.

Journal of Microscopy 217(1), 60–68.

Benes, V., K. Bodlák, J. Møller, and R. Waagepetersen (2002). Bayesian analysis of log Gaussian Cox process models for disease mapping. *Technical Report R-02-2001, Department of Mathematical Sciences. Aalborg University.*

Bernardi, R. and P. Pandolfi (2003). Role of PML and the PML-nuclear body in the control of programmed cell death. *Nature 22*. 9048–9057.

Berthelsen, K. and J. Møller (2002). Spatial jump processes and perfect simulation. In K. Mecke and D. Stoyan (Eds.), *Morthology of condensed Matter*. Volume 600 of *Lecture Notes in Physics*, pp. 391–412. Springer-Verlag.

Berthelsen, K. and J. Moller (2002(b)). A primer on perfect simulation for spatial point processes. *Bulletin of the Brazilian Mathematical Society 33*(33), 351–367.

Berthelsen, K. and J. Møller (2003). Likelihood and non-parametric Bayesian MCMC inference for spatial point processes based on perfect simulation and path sampling. *Scandinavian Journal of Statistics 30*, 549–564.

Berthelsen, K. and J. Møller (2006). Bayesian analysis of Markov point processes. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.). *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics. pp. 85. New York: Springer-Verlag.

Besag, J. and C. P. (1989). Generalised Monte Carlo significance tests. *Biometrika 76*(4), 633–642.

Besag, J. and D. P.J. (1977). Simple Monte Carlo tests for spatial data. *Applied Statistics 26*(3), 327–333.

Bolzer, A., G. Kreth, I. Solovei, D. Koehie, C. Fauth, S. Muller, R. Eils, C. Cremer, M. Speicher, and T. Cremer (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLOS Biology 3*(5), 826–842.

Borden, K. (2002). Pondering the promyelocytic leukemia protein (PML) puzzle: Possible functions for PML nuclear bodies. *Molecular and Cellular Biology 22*(15), 5259–5269.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability 31*(4), 929–953.

Brix, A. and W. Kendall (2002). Simulation of cluster point processes without edge effects. *Advances in Applied Probability 34*(2), 267–280.

C., L., T. Cheutin, M. Cremer, G. Cavalli, and T. Cremer (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature 8*, 104–115.

Ching, R., G. Dellaire, C. Eskiw, and D. Bazett-Jones (2005). PML bodies: a

meeting place for genomic loci? *Journal of Cell Science 118*(5), 847-854.

Cohn, D. (1980). *Measure Theory*. Boston: Birkhäuser.

Cowling, A., P. Hall, and M. Philips (1996). Bootstrap confidence regions for the intensity of a Poisson point process. *Journal of the American Statistical Association 91*(436), 1516–1524.

Cox, D. and V. Isham (1980). *Point processes*. New York: Chapman and Hall.

Cremer, M., J. Hase, T. Volm, A. Brero, K. G., J. Walter, C. Fischer, I. Solovei, C. Cremer, and T. Cremer (2001). Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Research 9*(7), 541-567.

Cressie, N. (1991). *Statistics for spatial data*. New York: John Wiley & Son.

Crick, F. and P. Lawrence (1975). Compartmens and polyclones in insect development. *Science 189*, 340–47.

Dellaire, G. and D. Bazett-Jones (2004). PML nuclear bodies: dynamic sensors of DNA damage and cellular stress. *BioEssays 26*(9), 963–977.

Diggle, P. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics 35*(1), 87–101.

Diggle, P. (1985). Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate point pattern. *Journal of Neuroscience Methods 18*, 115–125.

Diggle, P. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society: Series A (Statistics in society) 153*, 349–362.

Diggle, P. (2003). *Statistical analysis of spatial point patterns*. London: Arnold.

Diggle, P. and T. Cox (1983). Some distance-based tests of independence for sparsely sampled multivariate spatial point patterns. *International Statistical Review 51*. 11–23.

Diggle, P., S. Eglen, and J. Troy (2006). Modelling the bivariate spatial distribution of amacrine cells. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics, pp. 215–232. New York: Springer-Verlag.

Diggle, P., N. Lange, and F. Benes (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association 86*(415), 618–625.

Doguwa, S. and G. Upton (1990). On the estimation of the nearest neighbour distribution, g(t), for spatial processes. *Biometrika 32*, 863–876.

Dučinskas, K. and Šaltytė (2001). Quadratic discriminant analysis of spatially correlated data. *Nonlinear Analysis: Modelling and Control 6*(2), 15–28.

Dundr, M. and T. Misteli (2001). Functional architecture in the cell nucleus. *Journal of Biochemistry 356*, 297–310.

Fishman, G. (1996). *Monte Carlo.* New York: Springer-Verlag.

Fishman, P. and D. Snyder (1976). The statistical analysis of space-time point processes. *IEEE Transactions of Information Theory IT-22*(3), 257-274.

Fleischer, F., M. Beil, M. Kazda, and V. Schmidt (2006). Analysis of spatial point patterns in microscopic and macroscopic biological image data. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.). *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics, pp. 235. New York: Springer-Verlag.

Floresroux, E. and M. Stein (1996). A new method of edge correction for estimating the nearest neighbour distribution. *Journal of Statistical Planning and Inference 50*(3), 353–371.

Fox, A., Y. Lam, A. Leung, C. Lyon, J. Andersen, M. Mann, and A. Lamond (2002). Paraspeckls: A novel nuclear domain. *Current Biology 12*, 13–25.

Gelman, A. and X. Meng (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science 13*, 163–185.

Gignoux, J., C. Duby, and S. Barot (1999). Comparing the performance of diggle's test of spatial randomness for small samples with and without edge-effect correction: Application to ecological data. *Biometrics 55*(1), 156–164.

Gill, R. (1994). Lectures on survival analysis. In D. Bakry, R. Gill, and S. Molchanov (Eds.), *Lectures on probability theory*, Number 1581 in Lecture Notes in Mathematics, pp. 115. Berlin: Springer-Verlag.

Glasbey, C. and I. Roberts (1997). Statistical analysis of the distribution of gold particles over antigen sites after immunogold labelling. *Journal of Microscopy 186*(3), 258–262.

Grabarnik, P. and S. Chiu (2002). Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustered spatial point processes. *Biometrika 89*(2), 411–421.

Guan, Y. (2007). A composite likelihood cross-validation approach in selecting bandwidth for the estimation of the pair correlation function. *Scandinavian Journal of Statistics 34*(2), 336–346.

Haase, P. Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction.

Hahn, U., E. Jensen, M. Van Lieshout, and L. Nielsen (2003). Inhomogeneous spatial point processes by location-dependent scaling. *Advances in Applied Prob-*

ability 35, 319–336.

Halmos, P. (1974). *Measure Theory.* New York: Springer-Verlag.

Hanisch, K. and D. Stoyan (1979). Formulas for the second-order analysis of marked point processes. *Statistics A Journal of Theoretical and Applied Statistics 10*(4), 555–560.

Hodges, M., C. Tissot, K. Howe, D. Grimwade, and P. Freemont (1998). Structure, organization, and dynamics of promyelocytic leukemia protein nuclear bodies. *American Journal of Human Genetics* (63), 297–304.

Hogg, R. and A. Craig (1959). *Introduction to Mathematical Statistics.* New York: The Macmillan Company.

Hope, A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 30*(3), 582–598.

Jensen, E. and L. Nielsen (2000). Inhomogeneous Markov point processes by transformation. *Bernoulli 6*(5), 761–782.

Kallenberg, O. (1984). An informal guide to the theory of conditioning in point processes. *International Statistical Review 52*(2), 151–164.

Kapetanios, G. and M. Weeks. Non-nested models and the likelihood ratio statistic: A comparison of simulation and bootstrap based tests. *Working paper No. 490 (Unpublished Material).*

Karr, A. (1991). *Point processes and their statistical inference.* New York: Marcel Dekker.

Kelley, J. (1955). *General Topology.* New York: Litton Educational Publishing, Inc.

Kemp, C. (1988). Review. *The Statistician 37*(1), 84–85.

Kendall, W. and J. Møller (1999). Perfect Metropolis-Hastings simulation of locally stable point processes. *Research Report 348, Department of Statistics, University of Warwick.*

Kerscher, M. (1998). Regularity in the distribution of superclusters? *Astronomy and Astrophysics*, 29–34.

König, D., S. Carvajal-Gonalez, A. Downs, J. Vassy, and J. Rigaut (1991). Modelling and analysis of 3-d arrangements of particles by point processes with examples of application to biological data obtained by confocal scanning light microscopy. *Journal of Microscopy 161*(3), 405–433.

Kunsch, H. (1989). The jacknife and the bootstrap for general stationary observations. *The Annals of Statistics 17*(3), 1217–1241.

Kyan, M., G. L., M. Arnison, and C. Cogswell (2001). Feature extraction of chromosomes from 3-d confocal microscope images. *IEEE Transactions on Biomedical Engineering 48*(11), 1306–1318.

Lamond, A. and J. Sleeman (2003). Nuclear substructure and dynamics. *Current Biology 13*(21), R825–8.

Loh, J. and M. Stein (2004). Bootstrapping a spatial point process. *Statistica Sinica 14*, 69–101.

Loosmore, N. and E. Ford (2006). Statistical inference using the G or K point pattern spatial statistics. *Ecology 87*(8), 1925–1931.

Mahy, N., P. Perry, and W. Bickmore (2002). Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by fish. *Journal of Cell Biology 159*(5), 753–763.

McManus, K., D. Stephens, N. Adams, S. Islam, P. Freemont, and M. Hendzel (2006). The transcriptional regulator CBP has defined spatial association within interphase nuclei. *PLOS Computational Biology 2*(10), 1271–1283.

Melnick, A. and J. Licht (1999). Deconstructing a disease: Rarα, its fusion partners, and their roles in the pathogenesis of acute promyelocytic leukemia. *Blood 93*(10), 3167–3215.

Mikecz, A., S. Zhang, M. Montminy, E. M. Tan, and P. Hemmerich (2000). Creb-Binding Protein (Cbp/P300) and RNA Polymerase II Colocalize in Transcriptionally Active Domains in the Nucleus. *Journal of Cell Biology 150*(1), 265–274.

Milne, R. and M. Westcott (1972). Further results for Gauss-Poisson processes. *Advances in Applied Probability 4*(1), 151–176.

Møller, J. (2001). A review of perfect simulation in stochastic geometry. In I. Basawa, C. Heyde, and R. Taylor (Eds.), *Selected Proceedings of the Symposium on Inference for Stochastic Processes*, Volume 37 of *Lecture Notes & Monographs Series*, pp. 333. IMS.

Møller, J., A. Syversveen, and R. Waagepetersen (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics 25*, 451–482.

Møller, J. and G. Torrisi (2005). Generalised shot noise Cox processes. *Advances in Applied Probability 37*, 48–74.

Møller, J. and R. Waagepetersen. Statistical inference for Cox processes. In *Spatial Cluster Modelling*. Chapman and Hall.

Møller, J. and R. Waagepetersen (2003). An introduction to simulation-based inference for spatial point processes. In J. Møller (Ed.), *Spatial Statistics and Computational Methods*, Number 173 in Lecture Notes in Statistics, pp. 143. New York: Springer-Verlag.

Møller, J. and R. Waagepetersen (2004). *Statistical inference and simulation for spatial point processes*. Boca Raton: Chapman & Hall.

Naniche, D., S. Reed, and M. Oldstone (1999). Cell cycle arrest during measles

virus infection: a G_0-like block leads to suppression of retinoblastoma protein expression. *Journal of Virology*, 1894–1901.

Newman, D. (1970). A new family of point processes which are characterised by their second moment properties. *Journal of Applied Probability 7*(2), 338–358.

Numata, M. Forest vegetation in the vicinity of choshi. *Bulletin of Choshi Marine Laboratory*.

Ogata, Y. and M. Tanemura (1986). Likelihood estimation of interaction potentials and external fields of inhomogeneous spatial point patterns. In I. Francis, B. Manly, and F. Lam (Eds.), *Proceedings of the Pacific Statistical Congress*, Number 185, pp. 150. Amsterdam: Elsevier.

Paine, P., L. Moore, and S. Horowitz (1975). Nuclear envelope permeability. *Nature 254*, 109–114.

Perry, G. and B. Miller (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology 187*, 59–82.

Politis, D. and J. Romano (1992). A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *The Annals of Statistics 20*(4), 1985–2007.

Politis, D. and M. Sherman (2001). Moment estimation for statistics from marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 261–275(15).

Prasad, V., D. Semwogerere, and E. Weeks (2007). Confocal microscopy of colloids. *Journal of Physics: condensed matter 19*.

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.

Prokešovaá, M., U. Hahn, and E. Jensen (2006). Statistics for locally scaled point processes. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics, pp. 99. New York: Springer-Verlag.

Raghavan, N., P. Goel, and S. Ghosh (1998). Computer experiments for the classification of spatial point processes. *Technical Report No 600*.

Rathbun, S. and N. Cressie (1994). Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability 26*(1), 122–154.

Reed, M. and C. Howard (1997). Edge-corrected estimators of the nearest-neighbour distance distribution function for three-dimensional point patterns. *Journal of Microscopy 186*(2), 177–184.

Reid, M. and B. Szendrői (2005). *Geometry and Topology*. Cambridge: Cambridge

University Press.

Ripley, B. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society 39*(2), 172–212.

Ripley, B. (1979a). Algorithm as 137: Simulating spatial patterns: dependent samples from a multivariate density. *Applied statistics 28*(1). 109–112.

Ripley, B. (1979b). Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society: Series B (Methodological) 41*(3). 368–374.

Ripley, B. (1987). *Stochastic simulation.* New York: Wiley.

Russell, R. (2006). Automated methods for segmentation and quantitative analysis of nuclear compartment. *MSc Bioinformatics Individual Project, Imperial College London (unpublished material).*

Schladitz, K. and A. Baddeley (2000). A third order point process characteristic. *Scandinavian Journal of Statistics 27*, 657–671.

Semwogerere, D. and E. Weeks (2005). Confocal microscopy. *Encyclopedia of Biomaterials and Biomedical Engineering.*

Sherman, M. and E. Carlstein (1994). Nonparametric estimation of the moments of a general statistic computed from spatial data. *Journal of the American Statistical Association 89*(426), 496–500.

Shiels, C. (2005). *Personal communication.*

Shiels, C., N. Adams, S. Islam, D. Stephens, and P. Freemont (2007). Quantitative analysis of cell nucleus organisation. *PLoS Computational Biology 3*, 1161–1168.

Shiels, C., S. Islam, R. Vatcheva, P. Sasieni, M. Sternberg, P. Freemont, and D. Sheer (2001). PML bodies associate specifically with the mhc gene cluster in interphase nuclei. *Journal of Cell Science 114*, 3705–3715.

Shopland, L., C. Johnson, M. Byron, J. McNeil, and J. Lawrence (2003). Clustering of multiple specific genes and gene-rich r-bands around sc-35 domains: Evidence for local euchromatic neighbourhoods. *Journal of Cell Biology 162.*

Skellam, J. (1952). Studies in statistical ecology: I. spatial pattern. *Biometrika 39*(3/4), 346–362.

Snethlage, M. (1999). Is bootstrap really helpful in point process statistics? *Metrika 49*, 245.

Solow, A. (1989). Bootstrapping sparsely sampled spatial point patterns. *Ecology 70*(2), 379–382.

Spector, D. (2001). Nuclear domains. *Journal of Cell Science 114.*

Stoyan, D. (2006). Fundamentals of point process statistics. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics, pp. 3. New York:

Springer-Verlag.

Stoyan, D., W. Kendall, and J. Mecke (1995). *Stochastic Geometry and its applications (second edition)*. Chichester: John Wiley & son Ltd.

Stoyan, D. and H. Stoyan (1994). *Fractals, random shapes and point fields*. Chichester: John Wiley & son Ltd.

Stoyan, D. and H. Stoyan (1998). Non-homogeneous gibbs process models for forestry-a case study. *Biometrical Journal 40*(5), 521–531.

Stoyan, D. and H. Stoyan (2000). Improving ratio estimators of second order point process characteristics. *Scandinavian Journal of Statistics 27*, 641–656.

Strauss, D. (1975). A model for clustering. *Biometrika 62*(2), 467–475.

Taylor, C., I. Dryden, and R. Farnoosh (2001). The K-Function for nearly regular point processe. *Biometrics 57*(1), 224–231.

Thakar, R. and A. Csink (2005). Changing chromatin dynamics and nuclear organization during differentiation in drosophila larval tissue. *Journal of Cell Science 118*, 951–960.

Thompson, H. (1955). Spatial point processes, with application to ecology. *Biometrika 42*(1/2), 102–115.

Turlach, B. (1993). Bandwidth selection in kernel density estimation: A review. *Institut de Statistique*.

Van Lieshout, M. (2004). A J-function for marked point patterns. *The Institute of Statistical Mathematics*, 511–532.

Van Lieshout, M. and A. Baddeley (1999). Indeces of dependence between types in multivariate point patterns. *Scandinavian Journal of statistics 26*, 511–532.

Waagepetersen, R. (2007). An estimation function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics 63*.

Walsh, D. and A. Raftery (2005). Classification of mixtures of spatial point processes via partial Bayes factors. *Journal of Computational & Graphical Statistics 14*(1), 139–154(16).

Wand, M. and M. Jones (1995). *Kernel smoothing*. New York: Chapman and Hall.

Wang, J., C. Shiels, P. Sasieni, P. Wu, S. Islam, P. Freemont, and D. Sheer (2004). Promyelocytic leukemia nuclear bodies associate with transcriptionally active genomic regions. *The Journal of Cell Biology 164*(4), 515–526.

Webb, A. (1999). *Statistical Pattern Recognition*. London: Arnold.

Webster, S., P. Diggle, H. Clough, R. Green, and N. French (2006). Strain-typing transmissible spongiform encephalopathies using replicated spatial data. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan (Eds.), *Case Studies in Spatial Point Pattern Modelling*, Number 185 in Lecture Notes in Statistics.

pp. 197. New York: Springer-Verlag.

Weimer, R., T. Haaf, J. Kruger, M. Poot, and M. Schmid (1992). Characterization of centromere arrangements and test for random distribution in g0,g1,s.g2.g1. and early S' phase in human lymphocytes. *Human Genetics 88*(6). 673-682.

Wiblin, A., W. Cui, A. Clark, and W. Brickmore (2005). Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. *Journal of Cell Science 118*.

Xie, S. and A. Pombo (2006). Distribution of different phosphorylated forms of RNA polymerase II in relation to cajal and PML bodies in human cells: an ultrastructural study. *Histochemistry and Cell Biology 125*.

Zink, D., M. Amaral, A. Englmann, S. Lang, and L. Clarke (2004). Transcription-dependent spatial arrangements of cftr and adjacent genes in human cell nuclei. *Journal of Cell Biology 166*.

RHBL LONDON UNIV.