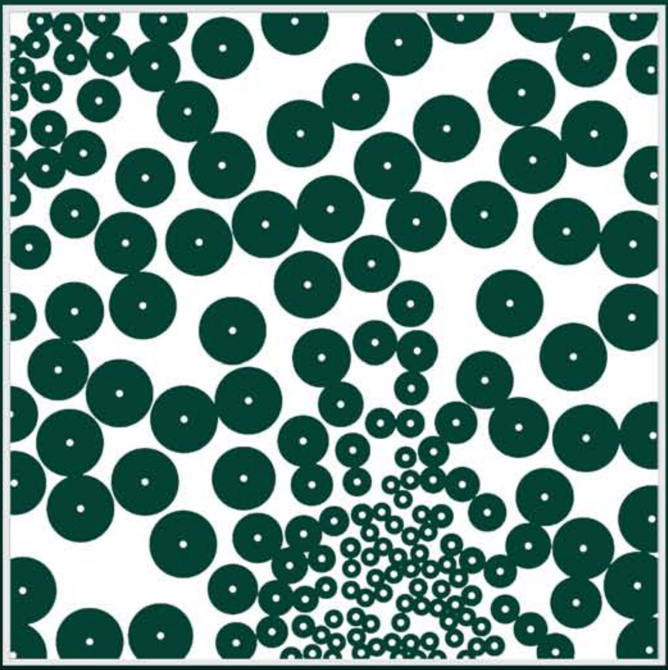
Monographs on Statistics and Applied Probability 107

Statistical Methods for Spatio-Temporal Systems



Edited by

Bärbel Finkenstädt

Leonhard Held

Valerie Isham

Chapman & Hall/CRC
Taylor & Francis Group

Monographs on Statistics and Applied Probability 107

Statistical Methods for Spatio-Temporal Systems

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (1960)
2 Queues *D.R. Cox and W.L. Smith* (1961)
3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5 Population Genetics *W.J. Ewens* (1969)
6 Probability, Statistics and Time *M.S. Barlett* (1975)
7 Statistical Inference *S.D. Silvey* (1975)
8 The Analysis of Contingency Tables *B.S. Everitt* (1977)
9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10 Stochastic Abundance Models *S. Engen* (1978)
11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12 Point Processes *D.R. Cox and V. Isham* (1980)
13 Identification of Outliers *D.M. Hawkins* (1980)
14 Optimal Design *S.D. Silvey* (1980)
15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16 Classification *A.D. Gordon* (1981)
17 Distribution-Free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19 Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen and E. Pesonen* (1984)
21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25 The Statistical Analysis of Composition Data *J. Aitchison* (1986)
26 Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27 Regression Analysis with Applications *G.B. Wetherill* (1986)
28 Sequential Methods in Statistics, 3rd edition
G.B. Wetherill and K.D. Glazebrook (1986)
29 Tensor Methods in Statistics *P. McCullagh* (1987)
30 Transformation and Weighting in Regression
R.J. Carroll and D. Ruppert (1988)
31 Asymptotic Techniques for Use in Statistics
O.E. Bandorff-Nielsen and D.R. Cox (1989)
32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33 Analysis of Infectious Disease Data *N.G. Becker* (1989)
34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36 Symmetric Multivariate and Related Distributions
K.T. Fang, S. Kotz and K.W. Ng (1990)
37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38 Cyclic and Computer Generated Designs, 2nd edition
J.A. John and E.R. Williams (1995)
39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40 Subset Selection in Regression *A.J. Miller* (1990)
41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)

44 Inspection Errors for Attributes in Quality Control
N.L. Johnson, S. Kotz and X. Wu (1991)

45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)

46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)

47 Longitudinal Data with Serial Correlation—A State-Space Approach
R.H. Jones (1993)

48 Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)

49 Markov Models and Optimization *M.H.A. Davis* (1993)

50 Networks and Chaos—Statistical and Probabilistic Aspects
O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall (1993)

51 Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)

52 Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)

53 Practical Risk Theory for Actuaries
C.D. Daykin, T. Pentikäinen and M. Pesonen (1994)

54 Biplots *J.C. Gower and D.J. Hand* (1996)

55 Predictive Inference—An Introduction *S. Geisser* (1993)

56 Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)

57 An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)

58 Nonparametric Regression and Generalized Linear Models
P.J. Green and B.W. Silverman (1994)

59 Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)

60 Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)

61 Statistics for Long Memory Processes *J. Beran* (1995)

62 Nonlinear Models for Repeated Measurement Data
M. Davidian and D.M. Giltinan (1995)

63 Measurement Error in Nonlinear Models
R.J. Carroll, D. Rupert and L.A. Stefanski (1995)

64 Analyzing and Modeling Rank Data *J.J. Marden* (1995)

65 Time Series Models—In Econometrics, Finance and Other Fields
D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (1996)

66 Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)

67 Multivariate Dependencies—Models, Analysis and Interpretation
D.R. Cox and N. Wermuth (1996)

68 Statistical Inference—Based on the Likelihood *A. Azzalini* (1996)

69 Bayes and Empirical Bayes Methods for Data Analysis
B.P. Carlin and T.A Louis (1996)

70 Hidden Markov and Other Models for Discrete-Valued Time Series
I.L. Macdonald and W. Zucchini (1997)

71 Statistical Evidence—A Likelihood Paradigm *R. Royall* (1997)

72 Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)

73 Multivariate Models and Dependence Concepts *H. Joe* (1997)

74 Theory of Sample Surveys *M.E. Thompson* (1997)

75 Retrial Queues *G. Falin and J.G.C. Templeton* (1997)

76 Theory of Dispersion Models *B. Jørgensen* (1997)

77 Mixed Poisson Processes *J. Grandell* (1997)

78 Variance Components Estimation—Mixed Models, Methodologies and Applications
P.S.R.S. Rao (1997)

79 Bayesian Methods for Finite Population Sampling
G. Meeden and M. Ghosh (1997)

80 Stochastic Geometry—Likelihood and computation
O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (1998)

81 Computer-Assisted Analysis of Mixtures and Applications—
Meta-analysis, Disease Mapping and Others *D. Böhning* (1999)

82 Classification, 2nd edition *A.D. Gordon* (1999)

83 Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)

84 Statistical Aspects of BSE and vCJD—Models for Epidemics
C.A. Donnelly and N.M. Ferguson (1999)

85 Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)

86 The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)

87 Complex Stochastic Systems
O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg (2001)

88 Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)

89 Algebraic Statistics—Computational Commutative Algebra in Statistics
G. Pistone, E. Riccomagno and H.P. Wynn (2001)

90 Analysis of Time Series Structure—SSA and Related Techniques
N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky (2001)

91 Subjective Probability Models for Lifetimes
Fabio Spizzichino (2001)

92 Empirical Likelihood *Art B. Owen* (2001)

93 Statistics in the 21st Century
Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells (2001)

94 Accelerated Life Models: Modeling and Statistical Analysis
Vilijandas Bagdonavicius and Mikhail Nikulin (2001)

95 Subset Selection in Regression, Second Edition *Alan Miller* (2002)

96 Topics in Modelling of Clustered Data
Marc Aerts, Helena Geys, Geert Molenberghs, and Louise M. Ryan (2002)

97 Components of Variance *D.R. Cox and P.J. Solomon* (2002)

98 Design and Analysis of Cross-Over Trials, 2nd Edition
Byron Jones and Michael G. Kenward (2003)

99 Extreme Values in Finance, Telecommunications, and the Environment
Bärbel Finkenstädt and Holger Rootzén (2003)

100 Statistical Inference and Simulation for Spatial Point Processes
Jesper Møller and Rasmus Plenge Waagepetersen (2004)

101 Hierarchical Modeling and Analysis for Spatial Data
Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand (2004)

102 Diagnostic Checks in Time Series *Wai Keung Li* (2004)

103 Stereology for Statisticians *Adrian Baddeley and Eva B. Vedel Jensen* (2004)

104 Gaussian Markov Random Fields: Theory and Applications
Håvard Rue and Leonhard Held (2005)

105 Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition
Raymond J. Carroll, David Ruppert, Leonard A. Stefanski,
and Ciprian M. Crainiceanu (2006)

106 Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood
Youngjo Lee, John A. Nelder, and Yudi Pawitan (2006)

107 Statistical Methods for Spatio-Temporal Systems
Bärbel Finkenstädt, Leonhard Held, and Valerie Isham (2007)

Monographs on Statistics and Applied Probability 107

Statistical Methods for Spatio-Temporal Systems

Edited by

Bärbel Finkenstädt
University of Warwick
Coventry, U.K.

Leonhard Held
Ludwig-Maximilians University
Munich, Germany

Valerie Isham
University College London
U.K.

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-593-9 (Hardcover)
International Standard Book Number-13: 978-1-58488-593-1 (Hardcover)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Preface

This volume contains a selection of invited papers presented by the contributors at the sixth Séminaire Européen de Statistique (SemStat) held as a summer school of the European Mathematical Society (EMS) at Castle Höhenried in Bernried near Munich in Germany on 12–18 December 2004. The aim of SemStat is to provide scientists who are at early stages of their careers with an opportunity to get quickly to the forefront of knowledge and research in areas of statistical science that are of current major interest. The chosen theme for this SemStat was “Statistical Methods for Spatio-Temporal Systems” and the invited papers were presented in either one or a series of three lectures by leading researchers and scientists in this field. These invited papers correspond to chapters in this book. Around 40 young researchers from various European countries participated in the 2004 SemStat summer school. They gave short presentations of their own research as listed on the following pages.

As with previous SemStat volumes the book concentrates on important statistical methods, theoretical aspects, and topical applications. The structure of the book is not strictly monographic in that each chapter is self-contained, consisting of a long expository article that starts by introducing the subject and progresses swiftly to incorporate recent research trends. The study of other chapters is beneficial but not vital to the understanding of the material in any single chapter. The order of the chapters is not important and readers may directly pick a chapter they are particularly interested in, study a collection of chapters, or read the whole book, ordering the chapters as they prefer. Applied scientists dealing with spatio-temporal data in a variety of research areas should benefit from this book, as should statistical researchers interested in modern statistical methodologies. Lecturers will find a variety of material suitable for graduate lecture courses in a statistical degree programme.

Spatio-temporal systems are systems that evolve over both space and time. The statistical viewpoint is to regard spatio-temporal data as realizations of random variables spread out in space and evolving in time. Statistical or stochastic models, such as point process models, for example, are used to provide the probabilistic backbone facilitating statistical inference from data. Such an approach is described in Chapter 1 where Diggle presents a review of statistical methods for spatio-temporal point process data. These methods are illustrated with specific examples of epidemic data on bovine tuberculosis, gastroenteric disease surveillance, and the foot and mouth outbreak in the U.K. Each of the applications gives rise to different statistical modelling approaches.

The second chapter by Vedel Jensen et al. is concerned with the important issue of modelling randomly growing objects as observed in diverse biological systems such as colonies of bacteria, tumours, and plant populations. The authors describe recent advances in stochastic growth models based on spatio-temporal point processes as well as growth models based on Lévy bases.

Readers who are interested in the use of data transformation tools such as power spectra, wavelets, and empirical orthogonal functions will appreciate the overview presented in Chapter 3 by Guttorp, Fuentes, and Sampson. The authors review methods, illustrating their use in a variety of applications in ecology and air quality. They also develop formal statistical tests verifying important assumptions such as stationarity and separability of a space-time covariance. The latter is also central to the following chapter. Geostatistical approaches to modelling spatio-temporal data rely on parametric covariance models and rather stringent assumptions, such as stationarity, separability, and full symmetry. Chapter 4 by Gneiting, Genton, and Guttorp reviews recent advances in the literature on space-time covariance functions, illustrated using wind data from Ireland.

Readers of the previous SemStat conference volume on *Extreme Values in Finance, Telecommunications, and the Environment* published in this monograph series (Vol. 99, 2004) will be familiar with statistical approaches to rainfall or windspeed data based on extreme value theory. In hydrological applications such as flood risk assessment, the simulation of rainfall data with high spatial-temporal resolution is required. In Chapter 5, Chandler et al. describe some stochastic and statistical models that can be used to provide simulated rainfall sequences for hydrological use. Model construction, inference, and diagnostics are all discussed. Many of the techniques described have applicability in more general space-time settings. This is also the case for the material introduced in Chapter 6 by Higdon, which provides a comprehensive primer on space-time modelling on the basis of Gaussian spatial and space-time models from a Bayesian perspective. Gaussian Markov random field specifications and Bayesian computational inference via Gibbs sampling and Markov chain Monte Carlo are central issues of this chapter. The methods are introduced and illustrated by a variety of examples using data on temperature surfaces, dioxin concentrations, ozone concentrations, and also simulated data from a well-established deterministic dynamical weather model.

It is not possible to cover all aspects of the conference theme in a single volume. Currently there are few direct links between the mathematical approaches to the mechanistic modelling of spatio-temporal systems using, for example, differential equations, pair approximations, or interactive particle systems and the stochastic and statistical modelling approaches as introduced here. A major reason for this is that the complexity and mechanistic realism that can be formulated mathematically is much larger than the model complexity that can be entertained regarding any available data from such systems. Statistical approaches generally start from the viewpoint of the data assuming stochastic or statistical modelling approaches as a vehicle for faciliating inference. It is our hope that the coverage provided by this volume will help readers acquaint themselves speedily with current statistical research issues in modelling spatio-temporal data and that it will enable further understanding and possible advances between the mechanistic and the statistical modelling communities.

The Séminaire Européen de Statistique is an activity of the European Regional Committee of the Bernoulli Society for Mathematical Statistics and Probability. The scientific programme was organised by the SemStat steering group, which, at the time of the conference, comprised Ole Barndorff-Nielsen (Aarhus), Bärbel Finkenstädt (Warwick), Leonhard Held (Munich), Alex Lindner (Munich), Enno Mammen (Mannheim), Gesine Reinert (Oxford), Michael Sørensen (Copenhagen), Ingrid Van Keilegom (Louvain-la-Neuve), and Aad van der Vaart (Amsterdam). The local organization of the meeting was in the hands of Leonhard Held (Munich) and the smooth running was in large part due to the enthusiastic help of Susanne Breitner, Michael Höhle, and Thomas Kneib (Munich). This SemStat was funded as an EMS summer school by the European Union under the sixth European framework (Marie Curie Actions) and some essential additional funding was provided by the Collaborative Research Centre Sonderforschungsbereich FB386, German research foundation (DFG). We are very grateful for this support and thanks are due to Luc Lemaire (Free University, Brussels) for coordinating the EMS summer school proposals. We would also like to thank all anonymous referees for reading the chapters and helping in improving their presentation.

On behalf of the steering group,
B. Finkenstädt (Warwick), L. Held (Munich), and V. Isham (London)

Participants

Sofia Åberg, Lund University (Sweden),
Forecasting radar precipitation using image warping.

Paul Anderson, University of Warwick (United Kingdom),
How tangled is nature?

John Aston, Academia Sinica (Taiwan),
Wavelet based spatial and temporal neuroimage analysis.

Enrica Bellone, University College London (United Kingdom),
Identification and modelling of rain event sequences.

Susanne Breitner, Ludwig-Maximilians-University Munich (Germany),
Modelling the association between air pollution and cardiorespiratory symptoms.

Petruta Caragea, Iowa State University (USA),
Alternative estimation of spatial parameters for large data sets.

Ignacio Cascos, Public University of Navarre (Spain),
Integral trimming.

Stella David, University of Augsburg (Germany),
Multivariate K-function.

Péter Elek, Eötvös Loránd University (Hungary),
Modelling extreme water discharges: comparison of a GARCH-type model with conventional hydrologic models.

Georgia Escaramis, University of Barcelona (Spain),
Techniques to estimate confidence intervals of risks in disease mapping.

Francesco Fedele, University of Vermont (USA),
Successive wave crests in Gaussian seas.

Laura Ferracuti, University of Perugia (Italy),
MCEM estimation for multivariate geostatistical non-Gaussian models.

Edith Gabriel, University of Montpellier II (France),
Estimating and testing zones of abrupt change for spatial data.

Axel Gandy, University of Ulm (Germany),
On goodness of fit tests for Aalen's additive risk model.

Susanne Gschlößl, Munich University of Technology (Germany),
Modelling spatial claim frequencies using a conditional autoregressive prior.

Elizabeth Heron, University of Warwick (United Kingdom),
Spatial modelling of cracks in bone cement.

Mathias Hofmann and Volker Schmid, Ludwig-Maximilians-University Munich (Germany),
A stochastic model for surveillance of infectious diseases.

Michael Höhle, Ludwig-Maximilians-University Munich (Germany),
Towards multivariate surveillance methods.

Rosaria Ignaccolo, University of Turin (Italy),
Model testing for spatial correlated data.

Janine Illian, University of Abertay at Dundee (United Kingdom),
Ecological communities – spatial point process modelling of highly complex systems.

José Ernesto Jardim, IPIMAR - National Institute for Agriculture and Fisheries Research (Portugal),
Designing bottom trawl surveys with geostatistical simulation.

Thomas Kneib, Ludwig-Maximilians-University Munich (Germany),
A general mixed model approach for spatio-temporal regression data.

Stephan Kötzer, University of Aarhus (Denmark),
Lower-tail behaviour in Wicksell's corpuscle problem.

Nadja Leith, University College London (United Kingdom),
Uncertainties in statistical downscaling.

Raquel Menezes, Lancaster University (United Kingdom),
Assessing spatial dependency under preferential sampling.

Tomasz Niedzielski, University of Wroclaw (Poland),
Testing for the Markov property within sedimentary rocks: case study from the Roztoka-Mokrzeszw Graben (SW Poland).

Irene Oliveira, University of Tràs-os-Montes e Alto Douro (Portugal),
Hilbert CCA — A new methodology to study time series processes in multivariate analysis.

Bianca Ortuani, University of Milan (Italy),
A geostatistical application in inverse problem for highly heterogeneous porous media.

Zbynek Pawlas, Charles University in Prague (Czech Republic),
Weak convergence of empirical distribution functions in germ-grain models.

Nathalie Peyrard, INRA — Institut National de la Recherche Agronomique (France),
The pair and Bethe approximations for the contact process on a graph.

Alexandre Pintore, University of Oxford (United Kingdom),
Rank-reduced methods for non-stationary covariance functions.

Philipp Pluch, University of Klagenfurt (Austria),
A Bayesian way out of the dilemma of underestimation of prediction errors in kriging.

Michalea Prokesova, Charles University in Prague (Czech Republic),
Statistics for locally scaled point processes.

Jakob Gulddahl Rasmussen, Aalborg University (Denmark),
Approximate simulation of Hawkes Processes.

Fabio Rigat, EURANDOM (The Netherlands),
Bayesian predictive survival trees.

Abel Rodriguez, Duke University (USA),
Bayesian structural alignment of proteins.

Ida Scheel, University of Oslo (Norway),
Spread of infectious agents in salmon farming.

Jana Sillmann, Max Planck Institute for Meteorology (Germany),
Weather extremes in climate change simulations.

Sibylle Sturtz, University of Dortmund (Germany),
Poisson-Gamma models as a tool for analysing different spatial patterns in the presence of covariates.

Thordis Linda Thorarinsdottir, University of Aarhus (Denmark),
Spatio-temporal model for fMRI data.

Ingunn Fride Tvete, University of Oslo (Norway),
Examining patterns in the spawning of Pollock by the Alaska West coast.

Jan van de Kassteele, Wageningen University (The Netherlands),
A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output.

Hanne Wist, Norwegian University of Science and Technology (Norway),
Speeding up block-MCMC algorithms for hierarcial GMRF models.

Contributors

Ole E. Barndorff-Nielsen
The T.N. Thiele Centre
Department of Mathematical Sciences
University of Aarhus
Aarhus, Denmark

Enrica Bellone
Department of Statistical Science
University College London
London, United Kingdom

Richard E. Chandler
Department of Statistical Science
University College London
London, United Kingdom

Peter J. Diggle
Department of Mathematics and Statistics
Lancaster University
Lancaster, United Kingdom

Montserrat Fuentes
Department of Statistics
North Carolina State University
Raleigh, North Carolina

Marc G. Genton
Department of Statistics
Texas A&M University
College Station, Texas

Tilmann Gneiting
Department of Statistics
University of Washington
Seattle, Washington

Peter Guttorp
Department of Statistics
University of Washington
Seattle, Washington

David Higdon
Los Alamos National Laboratory
Statistical Sciences Group
Los Alamos, New Mexico

Valerie Isham
Department of Statistical Science
University College London
London, United Kingdom

Eva B. Vedel Jensen
The T.N. Thiele Centre
Department of Mathematical Sciences
University of Aarhus
Aarhus, Denmark

Kristjana Ýr Jónsdóttir,
The T.N. Thiele Centre
Department of Mathematical Sciences
University of Aarhus
Aarhus, Denmark

Paul Northrop
Department of Statistical Science
University College London
London, United Kingdom

Paul D. Sampson
Department of Statistics
University of Washington
Seattle, Washington

Jürgen Schmiegel
The T.N. Thiele Centre
Department of Mathematical Sciences
University of Aarhus
Aarhus, Denmark

Chi Yang
Department of Statistical Science
University College London
London, United Kingdom

Contents

CHAPTER 1

Spatio-Temporal Point Processes: Methods and Applications

Peter J. Diggle

Contents

1.1 Introduction

This chapter is concerned with the analysis of data whose basic format is $(x_i, t_i) : i = 1, \ldots, n$, where each x_i denotes the location and t_i the corresponding time of occurrence of an *event* of interest. We shall assume that the data form a complete record of all events which occur within a pre-specified spatial region A and a prespecified time-interval, $(0, T)$. We call a data-set of this kind a *spatio-temporal point pattern*, and the underlying stochastic model for the data a *spatio-temporal point process*.

1.1.1 Motivating examples

1.1.1.1 Amacrine cells in the retina of a rabbit

One general approach to analysing spatio-temporal point process data is to extend existing methods for purely spatial data by considering the time of occurrence as a distinguishing feature, or *mark*, attached to each event. Before giving an example of this, we give an even simpler example of a marked spatial

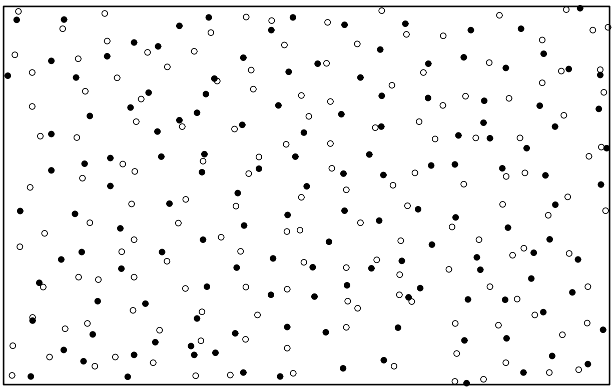


Figure 1.1 Amacrine cells in the retina of a rabbit. On and off cells are shown as open and closed circles, respectively. The rectangular region on which the cells are observed has dimension 1060 by 662 μm.

point pattern, in which the events are of just two qualitatively different types. Each event in Figure 1.1 represents the location of an amacrine cell in the retina of a rabbit. These cells play a fundamental role in mammalian vision. One type transmits information when a light goes *on*; the other type similarly transmits information when a light goes *off*. The data consist of the locations of 152 on cells and 142 off cells in a rectangular region of dimension 1060 by 662 μm.

The primary goal for the analysis of these data is to discriminate between two competing developmental hypotheses. The first hypothesis is that the pattern forms initially in two separate layers, corresponding to their pre-determined functionality; the second is that the pattern forms initially in a single, undifferentiated layer with function determined at a later developmental stage. One way to formalise this in statistical terms is to ask whether the two component patterns are statistically independent. Approximate independence would favour the first hypothesis. As we shall discuss in Section 1.2, this statement is a slight over-simplification but it provides a sensible starting point for an analysis of the data.

Our description and later analysis of these data are based on material in Diggle et al. (2005a). For a general discussion of the biological background, see Hughes (1985).

1.1.1.2 Bovine tuberculosis in Cornwall, U.K.

Our second example concerns the spatio-temporal distribution of reported cases of bovine tuberculosis (BTB) in the county of Cornwall, U.K., over the years 1991 to 2002. Individual cases are identified from annual inspections of farm herds; hence the effective time-resolution of the data is 1 year.

The prevalence of BTB has been increasing during the 12-year period covered by the data, but the observed annual counts exaggerate this effect because the scale of the annual inspection programme has also increased. Each recorded case is classified genetically, using the method of spoligotyping (Durr et al., 2000). The main scientific interest in these data lies not so much in the overall spatio-temporal distribution of the disease, but rather in the degree of spatial segregation amongst the different spoligotypes, and whether this spatial segregation is or is not stable over time. If the predominant mode of transmission is through local cross-infection, we might expect to find a stable pattern of spatial segregation, in which locally predominant spoligotypes persist over time; whereas if the disease is spread primarily by the introduction of animals from remote locations which are bought and sold at market, the resulting pattern of spatial segregation should be less stable over time (Diggle et al., 2005c).

Figure 1.2 shows the spatial distributions of cases corresponding to each of the four most common spoligotypes. The visual impression is one of strong spatial segregation, with each of the four types predominating in particular sub-regions.

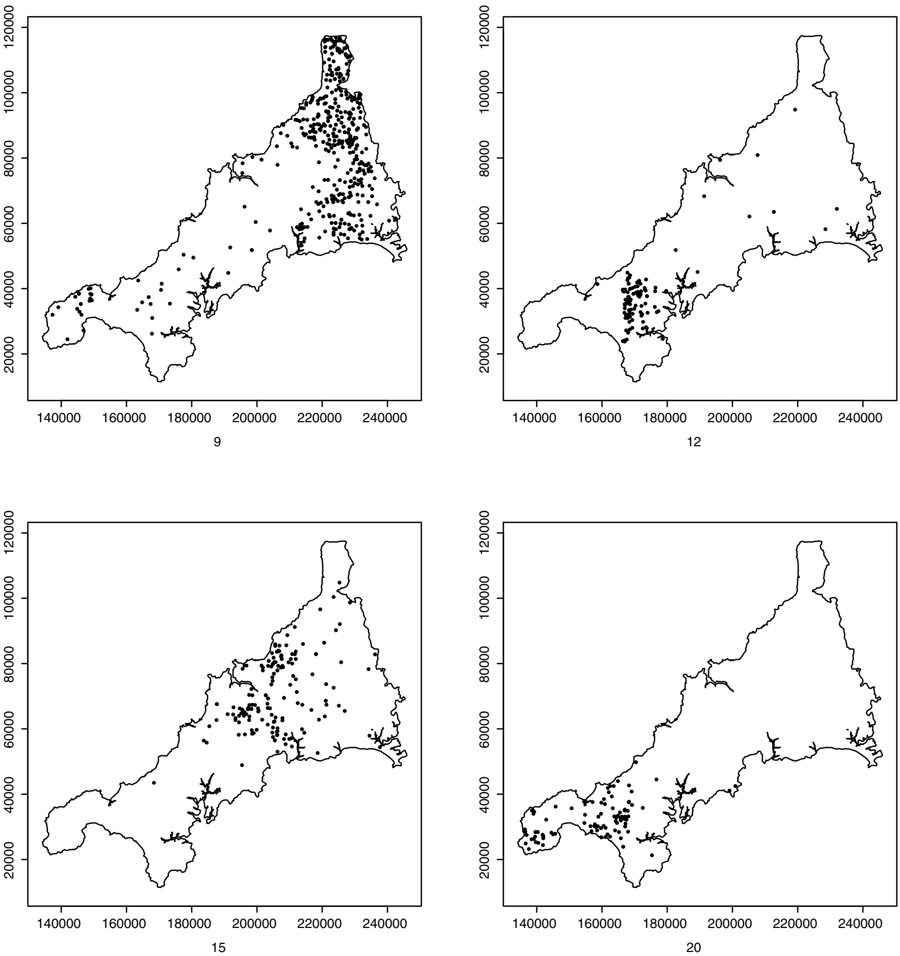


Figure 1.2 Spatial distributions of the four most common spoligotype data over the 14 years. Top row: spoligotype 9 (left) and spoligotype 12 (right). Bottom row: spoligotype 15 (left) and spoligotype 20 (right).

1.1.1.3 Gastroenteric disease in Hampshire, U.K.

Our third example concerns the spatio-temporal distribution of gastroenteric disease in the county of Hampshire, U.K., over the years 2001 and 2002. The data are derived from calls to NHS Direct, a 24-hour, 7-day phone-in service operating within the U.K. National Health Service. Each call to NHS Direct generates a data-record which includes the caller's post-code, the date of the call and a symptom code (Cooper et al., 2003). Figure 1.3 shows the locations of the 7167 calls from patients resident in Hampshire whose assigned symptom code corresponded to acute, non-specific gastroenteric disease. The spatial distribution of cases largely reflects that of the population of Hampshire, with strong concentrations in the large cities of Southampton and Portsmouth, and

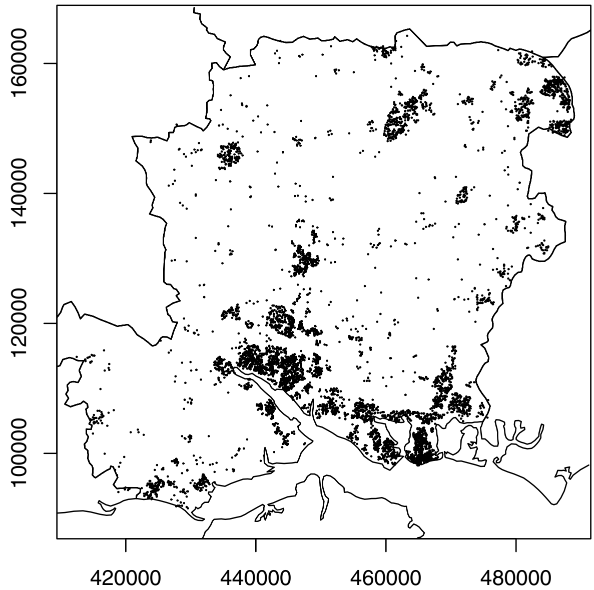


Figure 1.3 Locations of 7167 incident cases of non-specific gastroenteric disease in Hampshire, 1 January 2001 to 31 December 2002.

smaller concentrations in other towns and villages. Inspection of a dynamic display of the space-time coordinates of the cases suggests the kind of pattern typical of an endemic disease, in which cases can occur at any point in the study region at any time during the 2-year period. Occasional outbreaks of gastroenteric disease, which arise as a result of multiple infections from a common source, should result in anomalous spatially and temporally localised concentrations of cases.

The data were collected as part of the AEGISS project (Diggle et al., 2003), whose overall aim was to improve the timeliness of the disease surveillance systems currently used in the U.K. The specific statistical aims for the analysis of the data are to establish the normal pattern of spatial and temporal variation in the distribution of reported cases, and hence to develop a method of real-time surveillance to identify as quickly as possible any anomalous incidence patterns which might signal the onset of an outbreak requiring some form of public health intervention.

1.1.1.4 The U.K. 2001 epidemic of foot-and-mouth disease

Foot-and-mouth disease (FMD) is a highly infectious viral disease of farm livestock. The virus can be spread directly between animals over short distances in contaminated airborne droplets, and indirectly over longer distances, for example via the movement of contaminated material. The U.K. experienced a

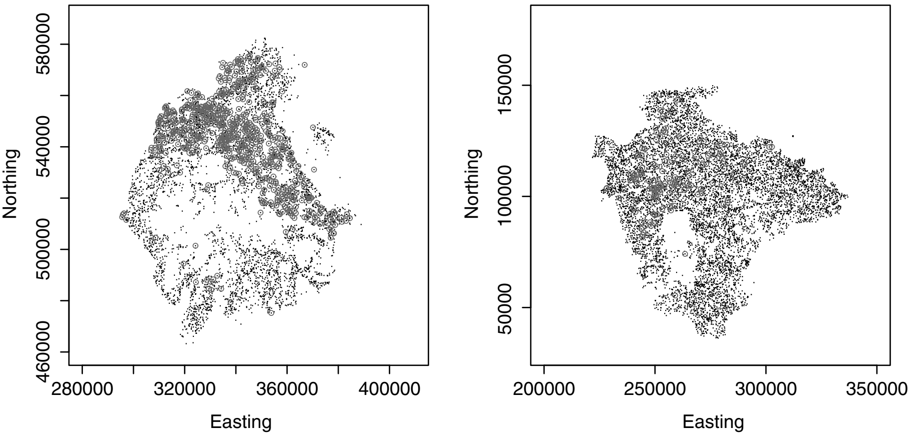


Figure 1.4 (SEE COLOR INSERT FOLLOWING PAGE 142) Locations of at-risk farms (black) and FMD case-farms (red) in Cumbria (left-hand panel) and in Devon (right-hand panel).

major FMD epidemic in 2001, which resulted in the slaughter of more than 6 million animals. Its estimated total cost to the U.K. economy was around £8 billion (U.K. National Audit Office, 2002). The epidemic affected 44 counties, and was particularly severe in the counties of Cumbria, in the north-west of England, and Devon, in the south-west. Figure 1.4 shows the spatial distributions of all farms in Cumbria and Devon which were at risk at the start of the epidemic, and of the farms which experienced the disease. In sharp contrast to the data on gastroenteric disease in Hampshire, the case-farms are strongly concentrated in sub-regions within each of the two counties. Dynamic plotting of the space-time locations of case-farms confirms the typical pattern of a highly infectious, epidemic disease. The predominant pattern is of transmission between near-neighbouring farms, but there are also a few, apparently spontaneous outbreaks of the disease far from any previously infected farms.

The main control strategies used during the epidemic involved the pre-emptive slaughter of animal holdings at farms thought to be at high risk of acquiring, and subsequently spreading, the disease. Factors which could affect whether a farm is at high risk include, most obviously, its proximity to infected farms, but also recorded characteristics such as the size and species composition of its holding. One objective in analysing these data is to formulate and fit a model for the dynamics of the disease which incorporates these effects. A model of this kind could then provide information on what forms of control strategy would be likely to prove effective in any future epidemic.

1.1.2 Chapter outline

In Section 1.2 we give a brief review of statistical methods for spatial point patterns, illustrated by an analysis of the amacrine cell data shown in Figure 1.1. We refer the reader to Diggle (2003) or Møller and Waagepetersen

(2003) for more detailed accounts of the methodology, and to Diggle et al. (2005a) for a full account of the data-analysis.

In Section 1.3 we discuss strategies for analysing spatio-temporal point process data. We argue that an important distinction in practice is between data for which the individual events (x_i, t_i) occur in a space-time continuum, and data for which the time-scale is either naturally discrete, or is made so by recording only the aggregate spatial pattern of events over a sequence of discrete time-periods. Our motivating examples include instances of each of these scenarios. Other scenarios which we do not consider further are when the locations are coarsely discretised by assigning each event to one of a number of sub-regions which form a partition of A. Methods for the analysis of spatially discrete data are typically based on Markov random field models. An early, classic reference is Besag (1974). Book-length treatments include Cressie (1991), Banerjee et al. (2003), and Rue and Held (2005).

In later sections we describe some of the available models and methods through their application to our motivating examples. This emphasis on specific examples is to some extent a reflection of the author's opinion that generic methods for analysing spatio-temporal data-sets have not yet become well established; certainly, they are less well established than is the case for purely spatial data. Nevertheless, in the final section of the chapter we will attempt to draw some general conclusions which go beyond the specific examples considered, and can in that sense be regarded as pointers towards an emerging general methodology.

1.2 Statistical methods for spatial point processes

1.2.1 Descriptors of pattern: spatial regularity, complete spatial randomness, and spatial aggregation

A convenient, and conventional, starting point for the analysis of a spatial point pattern is to apply one or more tests of the hypothesis of *complete spatial randomness* (CSR), under which the data are a realisation of a homogeneous Poisson process. A homogeneous Poisson process is a point process that satisfies two conditions: the number of events in any planar region A follows a Poisson distribution with mean $\lambda|A|$, where $|\cdot|$ denotes area and the constant λ is the *intensity*, or mean number of events per unit area; and the numbers of events in disjoint regions are independent. It follows that, conditional on the number of events in any region A, the locations of the events form an independent random sample from the uniform distribution on A (see, for example, Diggle, 2003, Section 4.4). Hence, CSR embraces two quite different properties: a uniform marginal distribution of events over the region A; and independence of events. We emphasise that this is only a starting point, and that the hypothesis of CSR is rarely of any scientific interest. Rather, CSR is a dividing hypothesis (Cox, 1977), a test of which leads to a qualitative classification of an observed pattern as regular, approximately random or aggregated.

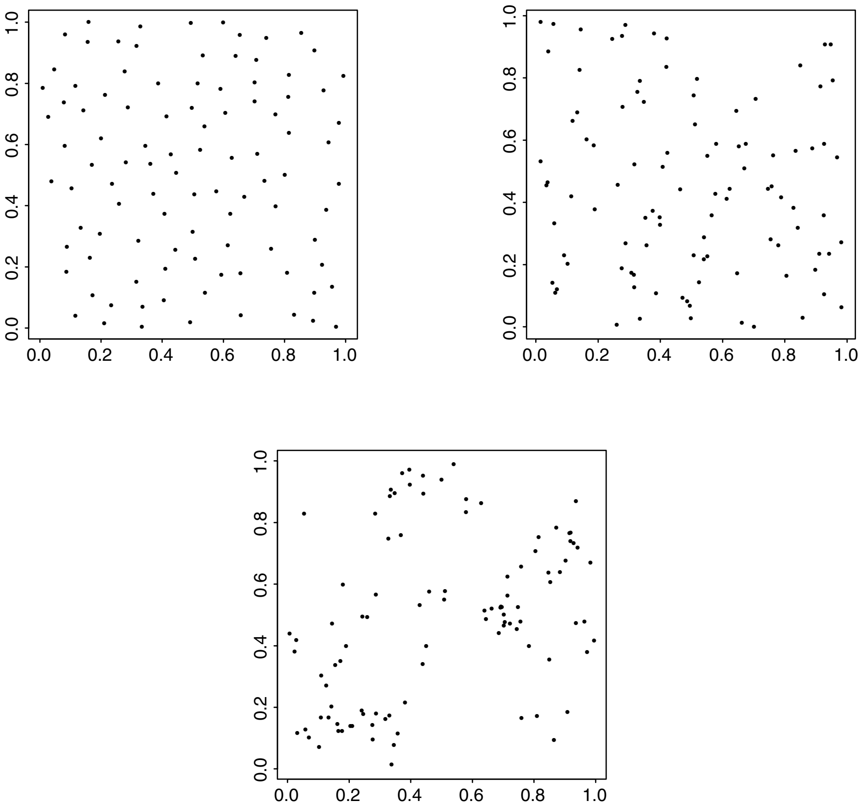


Figure 1.5 Examples of a regular (upper-left panel), a completely random (upper-right panel), and an aggregated (lower panel) spatial point pattern.

We do not attempt a precise mathematical definition of the descriptions "regular" and "aggregated." Roughly speaking, a regular pattern is one in which events are more evenly spaced throughout A than would be expected under CSR, and typically arises through some form of inhibitory dependence between events. Conversely, an aggregated pattern is one in which events tend to occur in closely spaced groups. Patterns of this type can arise as a consequence of marginal non-uniformity, or a form of attractive dependence, or both. In general, as shown by Bartlett (1964), it is not possible to distinguish empirically between underlying hypotheses of non-uniformity and dependence using the information presented by a single observed pattern. Figure 1.5 shows an example of a regular, a completely random, and an aggregated spatial point pattern. The contrasts amongst the three are clear.

1.2.2 Functional summary statistics

Tests of CSR which are constructed from functional summary statistics of an observed pattern are useful for two reasons: when CSR is conclusively rejected, their behaviour gives clues as to the kind of model which might provide a reasonable fit to the data; and they may suggest preliminary estimates of model parameters. Two widely used ways of constructing functional summaries are through nearest neighbour and second-moment properties. Third and higher-order moment summaries are easily defined, but appear to be rarely (possibly too rarely) used in data-analysis; an exception is Peebles and Groth (1975). They do feature, for example, in the theoretical analysis of ecological models, as discussed in Murrell, Dieckmann and Law (2004), and undoubtedly offer potential insights which are not captured by second-moment properties.

Two nearest neighbour summaries are the distribution functions of X, the distance from an arbitrary origin of measurement to the nearest event of the process, and of Y, the distance from an arbitrary event of the process to the nearest other event. We denote these by $F(x)$ and $G(y)$, respectively. The empirical counterpart of $F(x)$ typically uses the distances, d_i say, from each of m points in a regular lattice arrangement to the nearest event, leading to the estimate $\tilde{F}(x) = m^{-1} \sum I(d_i \leq x)$ where $I(\cdot)$ is the indicator function. Similarly, if e_i is the distance from each of n events to its nearest neighbour, then $\tilde{G}(y) = n^{-1} \sum I(e_i \leq y)$. Edge-corrected versions of these simple estimators are sometimes preferred, and are necessary if we wish to compare empirical estimates with the corresponding theoretical properties of a stationary point process.

Derivations, and further discussion, of results in the remainder of this section can be found, for example, in Diggle (2003, Chapter 4).

Under CSR, $F(x) = G(x) = 1 - \exp(-\lambda\pi x^2)$, where λ is the *intensity*, or mean number of events per unit area. Typically, in a regular pattern $G(x) < F(x)$, whereas in an aggregated pattern $G(x) > F(x)$.

To describe the second-moment properties of a spatial point process, we need some additional notation. Let dx denote an infinitesimal neighbourhood of the point x, and $N(dx)$ the number of events in dx. Then, the *intensity function* of the process is

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

Similarly, the *second-moment intensity function* is

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

and the *covariance density* is

$$\gamma(x, y) = \lambda_2(x, y) - \lambda(x)\lambda(y).$$

The process is stationary and isotropic if its statistical properties do not change under translation and rotation, respectively. If we now assume that the process is stationary and isotropic, the intensity function reduces to a constant, λ, equal to the expected number of events per unit area. Also, the second-moment intensity reduces to a function of distance, $\lambda_2(x, y) = \lambda_2(r)$ where $r = ||x - y||$ is the distance between x and y, and the covariance density is $\gamma(r) = \lambda_2(r) - \lambda^2$. In this case, the scaled quantity $\rho(r) = \lambda_2(r)/\lambda^2$ is called, somewhat misleadingly, the *pair correlation function*. For a homogeneous Poisson process, $g(r) = 1$ for all r.

A more tangible interpretation of the pair correlation function is obtained if we integrate over a disc of radius s. This gives the reduced second-moment measure, or *K-function*,

$$K(s) = 2\pi \int_0^s \rho(r) r \, dr. \tag{1.1}$$

Ripley (1976, 1977) introduced the K-function as a tool for data-analysis. One of its advantages over the pair correlation function is that it can be interpreted as a scaled expectation of an observable quantity. Specifically, let $E(s)$ denote the expected number of further events within distance s of an arbitrary event. Then,

$$K(s) = \lambda^{-1} E(s). \tag{1.2}$$

The result (1.2) leads to several useful insights. First, it suggests a method of estimating $K(s)$ directly by the method of moments, without the need for any smoothing; this is especially useful for relatively small data-sets. Second, it explains why $K(s)$ is a good descriptor of spatial pattern. For a completely random pattern, events are positioned independently; hence $E(s) = \lambda \pi s^2$ and $K(s) = \pi s^2$. This gives a benchmark against which to assess departures from CSR. For aggregated patterns, $K(s)$ is relatively large at small distances s because each event typically forms part of a "cluster" of mutually close events. Conversely, for regular patterns, $K(s)$ is relatively small at small distances s because each event tends to be surrounded by empty space. Another useful property is that $K(s)$ is invariant to random thinning, i.e., retention or deletion of events according to a series of independent Bernoulli trials. This follows immediately from (1.2), which expresses $K(s)$ as the ratio of two quantities, both of which vary by the same constant of proportionality under random thinning.

We use the following edge-corrected method of moments estimator proposed originally by Ripley (1976, 1977). For data $x_i \in A : i = 1, \ldots, n$, a natural estimator for $E(s)$ is

$$\tilde{E}(s) = n^{-1} \sum_{i=1}^{n} \sum_{j \neq i} I(r_{ij} \leq s), \tag{1.3}$$

where $r_{ij} = ||x_i - x_j||$. Except for very small values of s, this estimator suffers from substantial negative bias because events outside A are not recorded in the data. A remedy is to replace the simple count in (1.3) by a sum of weights w_{ij}, where w_{ij}^{-1} is the proportion of the circumference of the circle with centre

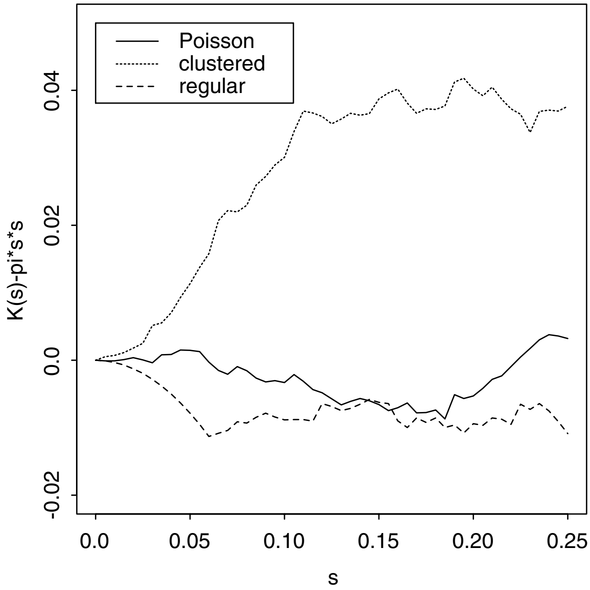


Figure 1.6 Estimates $\hat{K}(s) - \pi s^2$ for a regular (dashed line), a completely random (Poisson process, solid line), and an aggregated or clustered (dotted line) point pattern.

x_i and radius r_{ij} which lies within A. Finally, we estimate λ by $(n-1)/|A|$, where $|A|$ denotes the area of A, to give

$$\hat{K}(s) = |A|\{n(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(r_{ij} \leq s). \tag{1.4}$$

Ripley used $n/|A|$ to estimate λ. Our preference for $(n-1)/|A|$ has a slightly arcane theoretical justification which is discussed in Chetwynd and Diggle (1998) but this is clearly of no great consequence when n is large.

Figure 1.6 shows estimates $\hat{K}(s) - \pi s^2$ for each of the three point patterns shown in Figure 1.5. Subtraction of the CSR benchmark, $K(s) = \pi s^2$, emphasises departures from CSR, in effect acting as a magnifying glass applied to the estimate $\hat{K}(s)$.

Multivariate extensions of the K-function and its estimator were proposed by Lotwick and Silverman (1982). For a stationary, isotropic process let $\lambda_j : j = 1, \ldots, m$ denote the intensity of type j events. Define functions $K_{ij}(s) = \lambda_j^{-1} E_{ij}(s)$, where $E_{ij}(s)$ is the expected number of further type j events within distance s of an arbitrary type i event. Note that $K_{ij}(s) = K_{ji}(s)$. Although this equality is not obvious from the above definitions, it follows immediately from the multivariate analogue of our earlier definition (1.1) of $K(s)$ as an integrated version of the pair correlation function. However, direct extension of (1.4) to the multivariate case leads to two different estimates $\tilde{K}_{ij}(s)$ and

$\tilde{K}_{ji}(s)$ which, following Lotwick and Silverman (1982), we can combine to give the single estimate

$$\hat{K}_{ij}(s) = \{n_i\tilde{K}_{ij}(s) + n_j\tilde{K}_{ji}(s)\}/(n_i + n_j). \tag{1.5}$$

Two useful benchmark results for multivariate K-functions are:

1. If type i and type j events form independent processes, then $K_{ij}(s) = \pi s^2$;
2. If type i and type j events form a random labelling of a univariate process with K-function $K(s)$, then $K_{ii}(s) = K_{jj}(s) = K_{ij}(s) = K(s)$.

1.2.3 Functional summary statistics for the amacrines data

Figure 1.7 shows estimates $\hat{K}_{ij}(s) - \pi s^2$ for the amacrine cell data. Our interpretation of the three estimates is as follows. First, the near-equality of $\hat{K}_{11}(s)$ and $\hat{K}_{22}(s)$ suggests that the underlying biological process may be the same for both types of cell. Informally, the difference between $\hat{K}_{11}(s)$ and $\hat{K}_{22}(s)$ gives an upper bound to the size of the sampling fluctuations in the estimates. Second, both estimates show a strong inhibitory effect, with no two cells of the same type occurring within a distance of around 30 μm. Third, the

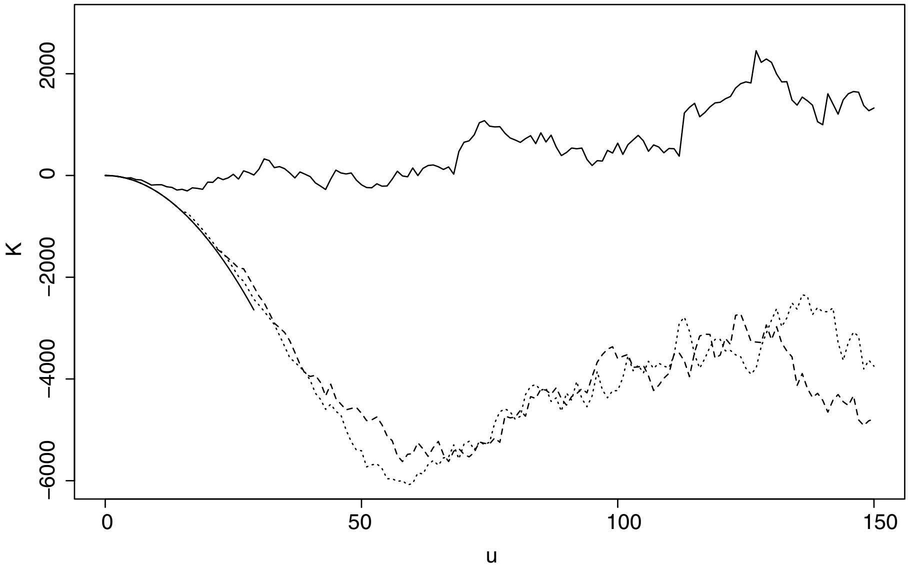


Figure 1.7 Estimates of the K-functions for the amacrine cell data. Each plotted function is $\hat{K}(s) - \pi s^2$. The dashed line corresponds to $\hat{K}_{11}(s)$ (on cells), the dotted line to $\hat{K}_{22}(s)$ (off cells), and the solid line to $\hat{K}_{12}(s)$. The parabola $-\pi s^2$ is also shown as a solid line.

estimate $\hat{K}_{12}(s)$ fluctuates around a value close to zero at small distances s, suggesting that the two component patterns are approximately independent. More specifically, $\hat{K}_{12}(s)$ does not show the strong inhibitory effect exhibited by both $\hat{K}_{11}(s)$ and $\hat{K}_{22}(s)$.

Collectively, these results are consistent with the first of the two developmental hypotheses for these data, namely that the component patterns of on and off cells form initially in two separate layers which later fuse to form the mature retina. Specifically, the separate layer hypothesis would imply statistical independence between the two component patterns; hence $K_{12}(s) = \pi s^2$. In fact, as we discuss below, the component patterns cannot strictly be independent because of the physical space required by each cell body. The data are clearly not compatible with random labelling of an initially undifferentiated pattern, as this would require all three estimated K-functions to be equal to within sampling variation. Furthermore, it is difficult to imagine how any biologically plausible labelling process could preserve strict inhibition between any two cells of the same type without imposing a similar constraint on two cells of opposite type. Hence, the analysis summarised in Figure 1.7 strongly favours the separate layer hypothesis.

1.2.4 Likelihood-based methods

Classical maximum likelihood estimation is straightforward for Poisson processes, but notoriously intractable for other point process models. Two more tractable alternatives are maximum pseudo-likelihood and Monte Carlo maximum likelihood. Both are particularly well suited to estimation in a class of models known as pairwise interaction point processes, and it is in this context that we discuss them here.

A third variant of likelihood-based estimation uses a *partial likelihood*. This method is best known in the context of survival analysis (Cox, 1972, 1975). We describe its adaptation to spatio-temporal point processes in Section 1.3.2.2.

1.2.4.1 Pairwise interaction point processes

Pairwise interaction processes form a sub-class of Markov point processes (Ripley and Kelly, 1977). They are defined by their likelihood ratio, $f(\cdot)$, with respect to a Poisson process of unit intensity. Hence, if $\chi = \{x_1, \ldots, x_n\}$ denotes a configuration of n points in a spatial region A, then $f(\chi)$ measures in an intuitive sense how much more likely is the configuration χ than it would be as a realisation of a Poisson process of unit intensity. For a pairwise interaction process, we need to specify a parameter β which governs the mean number of events per unit area and an *interaction function* $h(r)$, where r denotes distance. Intuitively, $h(r)$ is related to the likelihood that the model will generate pairs of events separated by a distance r, in the sense that the likelihood for a particular configuration of events depends on the product of $h(||x_i - x_j||)$ over all distinct pairs of events. Hence, for example, a value $h(r) = 0$ for all $r < \delta$ would imply that no two events can be separated by