SPATIAL CLUSTER MODELLING

Edited by
Andrew B. Lawson
David G.T. Denison

CHAPMAN & HALL/CRC

SPATIAL CLUSTER MODELLING

Edited by
Andrew B. Lawson
Department of Mathematical Sciences
University of Aberdeen
Aberdeen, UK

David G.T. Denison
Department of Mathematics
Imperial College of Science, Technology and Medicine
London, UK

CHAPMAN & HALL/CRC
A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Lawson, Andrew (Andrew B.)
Spatial cluster modeling / edited by Andrew B. Lawson, David G.T. Denison.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-266-2 (alk. paper)
1. Cluster analysis. 2. Spatial analysis. I. Denison, David G. T. II. Title.

QA278 .L39 2002
519.5′3—dc21 2002019297

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Visit the CRC Press Web site at www.crcpress.com

International Standard Book Number 1-58488-266-2
Library of Congress Card Number 2002019297
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

SPATIAL CLUSTER MODELLING

List of Contributors

Adrian Baddeley	University of Western Australia
Dankmar Böhning	Free University of Berlin
Simon D. Byers	AT&T Labs, New Jersey
Allan B. Clark	University of Aberdeen
Murray K. Clayton	University of Wisconsin, Madison
David G.T. Denison	Imperial College, London
José Tomé A.S. Ferreira	Imperial College, London
Jürgen Gallinat	Free University of Berlin
Ronald B. Gangnon	University of Wisconsin, Madison
Fred Godtliebsen	University of Tromsø
Christopher M. Hans	Duke University
Christopher C. Holmes	Imperial College, London
John T. Kent	University of Leeds
Hyoung Moon Kim	Texas A&M University
Andrew B. Lawson	University of Aberdeen
Marc Loizeaux	Florida State University, Tallahassee
Ian W. McKeague	Florida State University, Tallahassee

Bani K. Mallick	Texas A&M University
Kanti V. Mardia	University of Leeds
J. Steve Marron	University of North Carolina, Chapel Hill
Jesper Møller	Aalborg University
Stephen M. Pizer	University of North Carolina, Chapel Hill
Adrian E. Raftery	University of Washington, Seattle
Peter Schlattmann	Free University of Berlin
David A. van Dyk	Harvard University
Marie-Colette van Lieshout	CWI, Amsterdam
Rasmus P. Waagepetersen	Aalborg University
Christopher K. Wikle	University of Missouri–Columbia

Preface

The development of statistical methodology for the analysis of spatial data has seen considerable advances since the publication of the seminal work of Cressie (1993). In particular, the development of fast computational algorithms for sampling of complex Bayesian models (most notably Markov chain Monte Carlo algorithms) has allowed a wide range of problems to be addressed which hitherto could not be directly analysed. Many spatial problems can be considered within the paradigm of hierarchical Bayesian modelling and so the emphasis within this volume will lie within that area.

The aim of this volume is not to present a general review of spatial statistical modelling but rather to focus on the area of spatial *cluster* modelling. Hence the theme of this work is the highlighting of the diverse approaches to the definition of clusters and clustering in space (and its adjunct space-time), and to present state-of-the-art coverage of the diverse modelling approaches which are currently available. In Chapter 1 we provide a brief historical introduction to the subject area and, in particular, compare conventional and spatial clustering. In addition this chapter introduces the notation and different areas of study explored. After this initial chapter the volume is split into 3 parts, each relating to a specific area of cluster modelling. Part I deals with point and object process modelling, Part II involves spatial process modelling, while Part III contains papers relating to spatio-temporal models.

One of the features of modelling spatial data is the need to use fast computational algorithms to be able to evaluate the complex posterior distributions or likelihood surfaces which arise in spatial applications. The 1980s saw the development of Markov chain Monte Carlo algorithms based on the Gibbs and Metropolis-Hastings samplers, and witnessed rapid development of models for complex spatial problems. Not only could existing models be sampled from but newer more sophisticated models could also be developed and applied. Often these models are of a hierarchical form so this naturally leads to the Bayesian paradigm being of importance in a great deal of the work.

As the potential fields of application for spatial methods are so wide we cannot hope to cover all of them. Nevertheless the chapters here do make reference to data in astrophysics (Chapter 10), spatial epidemiology (Chapters 5,7,8,14), ecology (Chapters 4,11), imaging (Chapter 13), geology and the geosciences (Chapters 6,4,7,9,12). In addition, the volume provides a useful insight into the current issues and methodology used for spatial cluster modelling. We have specifically included the burgeoning area of spatio-temporal modelling as an important extension to standard spatial data analysis and Chapters 12,13,14 specifically deal with this topic.

Finally we would like to thank all the contributors for their timely and thoughtful articles. In addition, we acknowledge the help of the staff at CRC Press, in particular Kirsty Stroud, Jasmin Naim and Helena Redshaw for their continued support and encouragement in the production of this volume.

Contents

CHAPTER 1

Spatial Cluster Modelling: An Overview

A.B. Lawson   D.G.T. Denison

1.1 Introduction

When analysing spatial data one is often interested in detecting deviations from the expected. For instance, we may be interested in the answers to questions like, "Is there an unusual aggregation of leukemia cases around a nuclear power station ?" or, "Where is it likely that the air pollution level is above the legally allowed limit ?". In both cases the focus is on finding regions in (usually two-dimensional) space in which higher than expected counts, or readings, are observed. We shall call such areas *clusters* and determining their nature forms the focus of this work.

This volume brings together a collection of papers on the topic of spatial cluster modelling and gives descriptions of various approaches which begin to solve the problem of detecting clusters. The papers are statistical in nature but draw on results in other fields as diverse as astrophysics, medical imaging, ecology and environmental engineering.

Two examples of the sort of spatial processes that we shall consider here are displayed in Figs. 1.1-1.2. Fig. 1.1 is an example of a point process, where each dot is an "event" (in this case the occurrence of a cancer). Here it is of interest to determine whether the cases are more aggregated, or *clustered*, than expected and whether the clustering relates to the locations of any possible pollution sources. To assess this a background control disease map (which is not shown) is often used to represent the expected variation in the distribution of cases; this is often a function of the relative population density.

Fig. 1.2 is an example of a dataset that consists of observations of an underlying spatial process at a number of locations. The usual aim of an analysis of this type of data is to determine the value of the spatial process at all the locations in the domain of interest, assuming that each measurement is only observed in the presence of a random error component. However, we may also be interested in determining areas where the process is above some predefined limit or even, in some sense, above average.

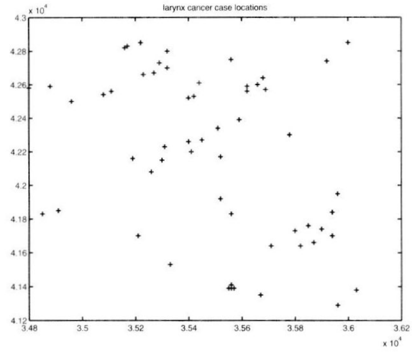

Figure 1.1 *An example of a point process: The Larynx cancer case event map (Diggle 1990) relating to cases in Lancashire, UK from 1974 to 1983.*

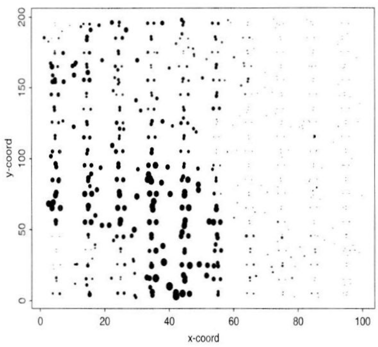

Figure 1.2 *An example of observations of a spatial process: The Piazza Road data (Higdon et al. 1999). Each dot represents a location where a measurement of the dioxin concentration in an area around the Piazza Road is taken. The size of the dots gives an indication of the observed measurements, with larger dots representing higher concentrations.*

Spatio-temporal data will also be looked at in this volume. In these examples either measurements or point processes are observed over time and similar questions arise but now clusters can occur in space, in time, or even in space and time jointly.

The analysis of spatial clustering has had a varied history with developments often resulting from particular applications, and so these developments have related to varying interest over time in different applications. For example, much work in the 1960s and 1970s on clustering developed from ecology applications (e.g. Diggle 1983, Ripley 1981), whereas an increased interest in image analysis in the 1980s led to associated advances in image segmentation and object recognition (e.g. Besag et al. 1991). Increased public and scientific interest in environmental hazards and health in the late 1980s and 1990s, has led to increased emphasis on cluster detection in small area health studies (Lawson et al. 1999, Elliott et al. 1999, Lawson 2001). The context of this historical development in relation to methodological progress is discussed more fully in the next section.

1.2 Historical Development

The analysis of clustering has a long history in statistical science. In one dimension the analysis of aggregation of data around a preferred location is at the heart of much statistical work, whether the focus is on the mean tendency of the aggregation of data or on its spread or variance. In addition in cluster studies the location of the maximum aggregation may also be of interest (modal property). In two dimensions, the natural extension of these ideas is to a two-dimensional aggregation, perhaps around a single point. In this case the centre of the aggregation may be defined either by mean or modal properties while the variance or spread of the aggregation can be defined around the putative centre which is now a two-dimensional location. In the case of spatial data these quantities have obvious interpretations as cluster centre location and cluster spread.

In the spatial domain, it is possible to view a clustered pattern in different ways depending on the focus of the analysis. First, it may be possible to conceive of the pattern as the realisation of a random process which produces aggregations as a result of the global structure of the process, whereby a small number of parameters control the scale and frequency of aggregations but the defined process does not parameterise the *locations* of the aggregations. This is akin to the geostatistical view of random processes, where the intensity or local density of events is defined by, for example, a spatial Gaussian process. The peaks of this process would correspond with local aggregations, but no parameterisation of the locations is made. Recent examples of this approach can be found in Cressie (1993) and Diggle et al. (1998). Essentially, this approach regards the aggregations as produced by random effects which are governed by global model parameters, i.e. the degree of aggregation and spread of the aggregations would be controlled by a small number of global parameters. This form of random effect modelling is at the heart of much hierarchical Bayesian modelling in this context, and in the literature of spatial applications the term *clustering* or *cluster modelling* is used to refer to such random effect modelling. An example from small area health studies is Clayton and Bernardinelli (1992).

The second approach to the modelling of clusters is to include within the modelled process specific elements which relate to cluster location and how these locations relate to the surrounding data. Much of the early work in stochastic geometry relating to point processes examined clustering of point processes and models such as the Neyman-Scott and Cox cluster processes. The fundamental feature of these processes was the definition of a set of cluster centres around which the *offspring* (data) lie. The term offspring comes from the idea that the clustering could arise from a multigenerational process. The variation in the local aggregation of data is thought to be summarised by the local density of events around the cluster centre set. In the case of a Neyman-Scott or Poisson cluster process the local density is