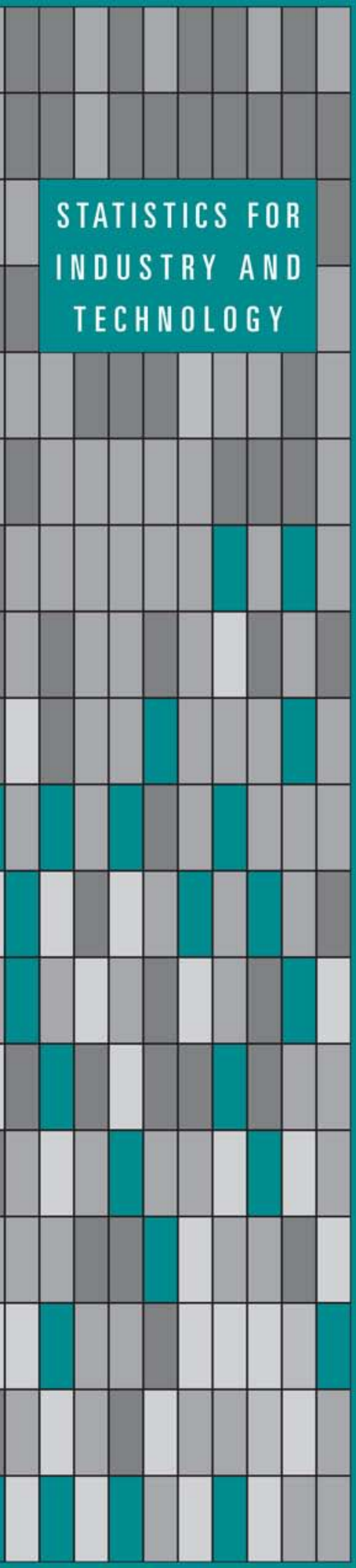


STATISTICS FOR INDUSTRY AND TECHNOLOGY

Joseph Glaz
Vladimir Pozdnyakov
Sylvan Wallenstein
Editors

Scan Statistics

Methods and Applications

Birkhäuser

Statistics for Industry and Technology

Series Editor

N. Balakrishnan
McMaster University
Department of Mathematics and Statistics
1280 Main Street West
Hamilton, Ontario L8S 4K1
Canada

Editorial Advisory Board

Max Engelhardt
EG&G Idaho, Inc.
Idaho Falls, ID 83415

Harry F. Martz
Group A-1 MS F600
Los Alamos National Laboratory
Los Alamos, NM 87545

Gary C. McDonald
NAO Research & Development Center
30500 Mound Road
Box 9055
Warren, MI 48090-9055

Kazuyuki Suzuki
Communication & Systems Engineering Department
University of Electro Communications
1-5-1 Chofugaoka
Chofu-shi
Tokyo 182
Japan

Scan Statistics

Methods and Applications

Joseph Glaz
Vladimir Pozdnyakov
Sylvan Wallenstein
Editors

Birkhäuser
Boston • Basel • Berlin

Editors
Joseph Glaz
Department of Statistics, U-4120
University of Connecticut
215 Glenbrook Rd.
Storrs, CT 06269-4120, USA
joseph.glaz@uconn.edu

Vladimir Pozdnyakov
Department of Statistics, U-4120
University of Connecticut
215 Glenbrook Rd.
Storrs, CT 06269-4120, USA
vladimir.pozdnyakov@uconn.edu

Sylvan Wallenstein
Department of Community
& Preventive Medicine
Box 1057
Mount Sinai School of Medicine
1 Gustave Levy Place
New York, NY 10029, USA
sylvan.wallenstein@mssm.edu

ISBN 978-0-8176-4748-3 e-ISBN 978-0-8176-4749-0
DOI 10.1007/978-0-8176-4749-0

Library of Congress Control Number: 2009926299

Mathematics Subject Classification (2000): 60C05, 60D05, 60G30, 60G35, 60G55, 60G63, 60G70, 60J22, 60M02, 62P10, 62P12, 62P25, 62P30, 62M30, 62N05

Printed on acid-free paper

Birkhäuser Boston is part of Springer Science+Business Media (www.birkhauser.com)

In honor of Joseph I. Naus

Contents

Preface

In the last ten years the area of scan statistics has risen to prominence in the field of applied probability and statistics. A recent search with Google Scholar lists 1780 references to scan statistics, 988 of which are from the last five years. It is quite impressive that about 200 articles on scan statistics are published each year. About 60 percent of the articles focus on spatial scan statistics and their applications. In addition to challenging theoretical problems, the area of scan statistics has exciting applications in many areas of science and technology, including: archaeology, astronomy, bioinformatics, biosurveillance, computer science, electrical engineering, epidemiology, food sciences, genetics, geography, material sciences, molecular biology, physics, reconnaissance, reliability and quality control, and telecommunication.

This volume has been edited in honor of Joseph Naus's seventieth birthday. The leading chapter, "Joseph Naus: Father of the Scan Statistic," by Sylvan Wallenstein, provides a comprehensive and interesting historical account of the early stages of research in the area of scan statistics, initiated by Joseph Naus almost half a century ago. The rest of the chapters have been arranged in alphabetical order of surnames of their leading authors.

In this volume, we have gathered a group of experts in the field of probability and statistics that have made significant contributions to the area of scan statistics, to review major developments in this area over the last ten years and to present recent or new results as well as point out new directions for future research. The contents of this volume illustrate the depth and the diversity of the methods and applications of the area. We hope that this volume will provide a comprehensive survey of the major recent developments in this area of research and will serve as a valuable reference and source for researchers in applied probability and statistics and in many other areas of science and technology. Graduate students interested in this area of research will find this volume to be of great value, as it points out many interesting and challenging research directions that they could pursue. This volume is suitable for use in teaching a graduate-level seminar course in applied probability and statistics.

Our sincere thanks go to all the authors, who showed great enthusiasm and support for this project. We appreciate their cooperation throughout the course of the project in submitting their articles on time and their help in reviewing the manuscripts. Additional thanks go to Mrs. Debbie Iscoe for her

Joseph Naus

support with issues related to typesetting this volume. Our special thanks go to N. Balakrishnan, Series Editor of *Statistics for Industry and Technology*, Regina Gorenshteyn, Associate Editor, and Tom Grasso, Editor, Computational Sciences and Engineering, Birkhäuser Boston (Springer) for their continual support and encouragement throughout the preparation of this volume.

Joseph Glaz thanks his wife, Sarah, and his son, Ron, for their continual loving support and encouragement. Vladimir Pozdnyakov thanks his mother, Valentina, and his late father, Ivan Ivanovich, as many called him, for nurturing Vladimir's interest in mathematics. Sylvan Wallenstein thanks his wife, Helene, for her love and encouragement, as well as for proofreading.

Storrs, CT, USA **J. Glaz**

Storrs, CT, USA **V. Pozdnyakov**

New York, NY, USA **S. Wallenstein**

Contributors

Balakrishnan, N. McMaster University, Hamilton, Ontario, Canada
bala@univmail.cis.mcmaster.ca

Boutsikas, M. University of Piraeus, Piraeus, Greece
mbouts@unipi.gr

Chan, H.P. National University of Singapore, Singapore, Republic of Singapore
stachp@nus.edu.sg

Chen, J. University of Massachusetts, Boston, MA, USA
jie.chen@umb.edu

Cooper, G.F. University of Pittsburgh, Pittsburgh, PA, USA
gfc@pitt.edu

Costa, M.A. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
macosta@ufmg.br

Das, K. Carnegie Mellon University, Pittsburgh, PA, USA
kaustav@cs.cmu.edu

Duarte, A.R. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
andersonrd@ufmg.br

Duczmal, L. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
duczmal@ufmg.br

Glaz, J. University of Connecticut, Storrs, CT, USA
joseph.glaz@uconn.edu

Fu, J.C. University of Manitoba, Winnipeg, Manitoba, Canada
fu@cc.umanitoba.ca

Haiman, G. UFR de Mathématiques, Université de Lille 1, Lille, France
haiman@ccr.jussieu.fr

Hoh, J. Yale University, New Haven, CT, USA
josephine.hoh@yale.edu

Jiang, X. University of Pittsburgh, Pittsburgh, PA, USA
xij6@pitt.edu

Joshi, S.W. Slippery Rock University of Pennsylvania, Slippery Rock, PA, USA
sharadchandra.joshi@sru.edu

Koli, R.E. Watershed Surveillance and Research Institute, Jalgaon, India
rek.jalasri@gmail.com

Koutras, M. University of Piraeus, Piraeus, Greece
mkoutras@unipi.gr

Kulldorff, M. Harvard University, Boston, MA, USA
martin_kulldorff@hms.harvard.edu

Lou, W.Y.W. University of Toronto, Toronto, Ontario, Canada
wendy.lou@utoronto.ca

Milienos, F. University of Piraeus, Piraeus, Greece
fmilien@unipi.gr

Myers, W.L. Pennsylvania State University, University Park, PA, USA
wlm@psu.edu

Neill, D.B. Carnegie Mellon University, Pittsburgh, PA, USA
neill@cs.cmu.edu

Ng, H.K.T. Southern Methodist University and Baylor Research Institute, Dallas, TX, USA
ngh@mail.smu.edu

Ott. J. Beijing Institute of Genomics, Beijing, China
ottjurg@yahoo.com

Patil, G.P. Pennsylvania State University, University Park, PA, USA
gpp@stat.psu.edu

Perone-Pacifico, M. Sapienza University of Rome, Rome, Italy
marco.peronepacifico@uniroma1.it

Pozdnyakov, V. University of Connecticut, Storrs, CT, USA
vladimir.pozdnyakov@uconn.edu

Preda, C. Faculté de Médecine, Université de Lille 2, Lille, France
cpreda@univ-lille2.fr

Robin, S. AgroParisTech/INRA, Paris, France
stephane.robin@agroparistech.fr

Schbath, S. INRA, Jouy-en-Josas, France
schbath@jouy.inra.fr

Schneider, J. Carnegie Mellon University, Pittsburgh, PA, USA
schneide@cs.cmu.edu

Stefanov, V.T. University of Western Australia, Crawley, Australia
stefanov@maths.uwa.edu.au

Steele, J.M. University of Pennsylvania, Philadelphia, PA, USA
steele@wharton.upenn.edu

Tango, T. National Institute of Public Health, Wako-shi, Japan
tango@niph.go.jp

Tavares, R. Universidade Federal de Ouro Preto, Ouro Preto, Brazil
tavares@iceb.ufop.br

Tu, I-P. Academia Sinica, Taipei, Taiwan
iping@stat.sinica.edu.tw

Verdinelli, I. Sapienza University of Rome, Rome, Italy, and
Carnegie Mellon University, Pittsburgh, PA, USA
isabella@stat.cmu.edu

Wallenstein, S. Mount Sinai School of Medicine, New York, NY, USA
sylvan.wallenstein@mssm.edu

Zhang, N.R. Stanford University, Stanford, CA, USA
nzhang@stanford.edu

List of Tables

List of Figures

1

Joseph Naus: Father of the Scan Statistic

Sylvan Wallenstein

Department of Community and Preventive Medicine, Mount Sinai School of Medicine, New York, NY, USA

Abstract: Currently, the literature on the scan statistic is vast, growing exponentially in diverse directions, with contributions by many researchers and groups. As time goes on, the early history of the problem bears telling. Joseph Naus, the father of the scan statistic, originated the modern work on the topic. The process took almost twenty years to reach maturity; I have chosen Naus (1982) as the definition of this maturity. The very name "scan statistic" does not appear to have become attached to the problem for fifteen years, and the interconnections to what is now one problem, in both statement of the problem and common methods of solution, was far from obvious originally. This chapter will not attempt a full review of all of Naus's statistical contributions, or even a full review of his contributions as they concern the scan statistic. Instead, it will focus on a few themes that had already originated in Naus's first twenty years of written research (1962–1982), and briefly continue with those threads to the present. Since these early themes include such general issues as applications of the scan statistic, mentoring graduate students, and specific methodological issues, the review will encompass a significant portion of Dr. Naus's research, without making claim to being exhaustive regarding either his research or the much broader topic of research he influenced on the scan statistic.

This chapter is divided into five parts:

1. Naus (1963), Naus's Ph.D. thesis, and the state of the art prior to 1965.
2. Naus's six singly authored first papers, covering all aspects of the problem, and focusing on exact solutions.
3. The first jointly published papers with Naus and his first five Ph.D. students working on the scan, focusing on exact values.
4. Two key publications in 1979–1982 that brought various strands together.
5. A shorter description of "later" work focusing on themes previously introduced.

Keywords and phrases: Scan statistic

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_1,

1.1 Naus (1963): Ph.D. Thesis

Joseph Naus graduated from the City College of New York in 1959 with a BBA in Economics. He began graduate study in Economics at Harvard the following year, where his advisor was Robert Dorfman. He was advised, as preparation for his graduate studies in economics, to broaden his knowledge in several areas — one of them being statistics. One of his first courses was taught by Arthur Dempster. The field intrigued him and seemed (and perhaps was, at the time) appreciably more manageable than the seemingly broader field of economics. At some point within his first year, he switched to the Statistics Department.

In his third year of graduate study (his second in statistics), Naus was spending an appreciable portion of the time in the Applied Science Division of the Operations Evaluation Group (OEG), which was funded through a contract awarded to MIT from the Navy. Naus notes in the preface to his thesis that on April 10, 1961, Jacinto Steinhardt outlined "Two Probability Problem Areas of Immediate Concern" to the OEG group. The first problem is stated as arising from "naval needs," which was probably motivated with the Navy wanting to know something about future buildup of naval forces in one region of the ocean. Nevertheless, from what Joe remembers, the problem was stated in general terms, though apparently with some emphasis on the two-dimensional problem. Naus, as a member of the OEG, began work on the problem in the fall of 1961 and on June 29, 1962 wrote up his results, Naus (1962), in ASD (Applied Science Division) Paper 8. This technical report, written before the thesis, is referenced in a footnote in Ederer, Myers, and Mantel (1964), which is apparently the first citation of Naus's work.

This line of research continued in a later contract with the Navy and culminated in a thesis approved in October 1963, under the direction of Frederick Mosteller in the Department of Statistics at Harvard, titled "Clustering of Random Points in the Line and Plane." The thesis acknowledged appreciation to Jerome Klotz, who had an appointment in the Business School.

The one-dimensional aspect of the problem, as stated in the thesis, concerns N points independently drawn from a random variable X on $[0, 1)$, with cumulative distribution $F(x)$. $P(k; N, w|F(x))$ is the probability that as some subinterval of length w scans the interval $[0, 1)$, it contains at least k of the N points on that larger interval. (Naus (1963) used the notation n instead of k, and sometimes referred to the problem as the "big N/little n" problem, but in keeping with later literature, this paper will use k for the size of the cluster.) When no argument is given for $F(x)$, X is assumed to follow a uniform distribution, so that $P(k; N, w)$ is the probability that given N points uniformly distributed on $[0, 1)$, there exists a subinterval of width w containing k or more points. As will

be pointed out below, this problem was but one of four parts of the "general" (one-dimensional) problem that would eventually emerge. The four subdivisions of the problem are formed by (i) conditioning, or not, on the total number of points in the interval, and (ii) considering discrete or continuous events. Various aspects of the problem would be studied for two decades, with some of the problems addressed in their own papers giving exact solutions. It was not until Naus (1982) that all four problems were put in the same framework and a single generic approximation was given for all four cases. Perhaps twenty years seems like a long time, but it should be noted that in addition to Naus's students and readers, giants of the field such as Mosteller and Karlin who dealt with various aspects of the four-fold problem also "missed" the global connection. In addition, it took considerable time and effort to lay the foundation to find exact values for the probabilities.

Some special cases of the problem had been previously considered. For $k = N$, the problem was one of finding the distribution of the range with the solution given by Burnside (1928, p. 22); for $k = 2$, the problem relates to the smallest distance between N points with the solution given by Parzen (1960, p. 304). Naus (1963) cites Feller (1958), who had noted the problem but stated that it involves complicated sample spaces, and thus implicitly did not have a simple solution.

The other papers most directly related to Naus's thesis project were Silberstein (1945), Berg (1945), and Mack (1948, 1950). These investigators were apparently the first to address the clustering problem beyond the special cases. They focused on the expected number of clusters, a topic that was apparently not to be addressed again for over thirty years when Glaz and Naus (1983) addressed the issue.

Mosteller, Naus's advisor, had worked on the discrete problem, which a decade or so later would be linked to the yet unnamed scan statistic, but for the first decade the link would remain unexplored. In the early 1960s, the two natural extensions to previous work were to $k = 3$ and to $k = N-1$, with $k = 3$ being the more promising. Naus recalls that another student of Mosteller, Tom Lehrer, who would later achieve fame as a well-known musical satirist, worked on the problem for $k = 3$. Apparently unbeknownst to Naus, and to this author ten years later, was a paper by Elteren and Gerrits (1961) that "nibbled" on the $k = 3$ problem by using a direct integration approach for $N = 6, 7, 8$.

One approach of Silberstein and Mack that Naus apparently used was the polynomial approach. Silberstein (1945) had noted that $P(k, N, w)$ is a polynomial in w of order N. Mack (1948) notes that the polynomial expression may change in different regions of $[0, 1)$. Naus exploits this observation in his thesis, as a lemma that helps move from a derivation for a particular value of w (typically $w = 1/L$, L an integer) to all w.

Naus's ground-breaking approach, which perhaps appears obvious in retrospect (but is not really, for one must know its limitations), was to phrase the problem in terms of paths, particularly what he termed 2-paths and L-paths, and then use combinatoric techniques, particularly the reflection principle, which allowed an exact solution to be computed. Whether an event of interest occurred, depended on whether the move of a path down preceded, or followed, a move up. This involved an analogy between points dropping in and out of an interval, and a cluster of points. Two different parts of the distribution were thus tackled: $w = 1/2$, and $k > N/2$. But it is perhaps even harder to realize the situation in which the argument fell apart, and why the condition $k > N/2$ is so critical. This is summarized as a footnote in both Naus (1963) and Naus (1965a).

Chapter 2 of the thesis found $P(k; N, w|F(x))$ for $k > N/2$, and Chapter 3 found limiting distributions. Chapter 4 explored two-dimensional generalizations, while the last chapter gave applications. It is probably the topic in the second chapter that sparked the greatest progress in subsequent papers and in research in the field up to about 1990. Already in the thesis, Naus showed an interest in a wide range of applications, for example, relating his work to work of Daniel Bernoulli concerning the "mutual inclinations of the planets."

In the thesis and in a later paper, Naus contrasts this "scan" approach with that based on a "fixed grid." The contrast can best be illustrated when $w = 1/L$, L an integer, in which case the fixed grid approach is based on the maximum number of events in any of the L intervals, while the "scan approach" is based on the maximum number of cases as the interval of length w scans the $[0, 1)$ interval. Naus seems to have used "scan" in this restricted context, more than in an attempt to label the statistic.

The beginning of Joe's work on the scan coincided with the start of his married life. During this period, he married Sarah Rosen who was originally from New Jersey. They had met after Joe's first year of graduate study. Their first daughter, Alisa, was born in 1962 while Joe was at Harvard, and their second daughter, Laura, was born in 1965 while Joe was at Rutgers. At Harvard, Joe remembers living in a small apartment with minimal extras, and commuting to Harvard by bicycle.

Joseph Irwin Naus's thesis was approved in September 1963, and the Ph.D. degree was awarded officially in January of the following year. In the 1963–1964 academic year, he continued his work full time as an operations research analyst at the Institute of Naval Studies.

1.2 The Early Papers Touching All Aspects of the Problem: 1965–1968

This section covers Naus's first five papers. In addition to the references previously cited, Naus was by this time aware of the asymptotic distribution for a scan-like statistic in Menon (1964). He found that this asymptotic approximation was not adequate. This problem would continue to plague approximations of the scan based on asymptotic theory and continue to provide justification for the search for exact values. Later, approximations, as opposed to asymptotic values, would be used with some measure of success.

Naus's first position after Harvard was as an assistant professor of statistics at Rutgers, joining the department in 1964 and having an appointment there in 1964–1966. As will be noted in this section, the professional collaborations and informal discussions between Naus and fellow faculty members were to be productive.

1.2.1 Maximum cluster of points on a line, Naus (1965a)

Interestingly, this first paper of Naus cites only Berg (1945), Mack (1948), Silberstein (1945), and Naus (1963). Since, as noted above, the contribution of the cited articles involved at most integration methods, the ideas in the paper were generated entirely by Naus, with possible help from his mentors at Harvard, and possibly later, Rutgers.

To understand the context of Naus's work, we introduce only a little notation, almost all in this paragraph. As noted above, derivations are simplified by considering the case $w = 1/L$, L an integer, so that the $[0, 1)$ interval can be viewed as divided into L parts. The event A denotes that one of these L intervals has k or more points, i.e. that at least one of the L cell occupancy numbers is at least k. The event B_i denotes the event that (i) A^c, all the L subintervals contain fewer than k points, and that (ii) there exists an interval of length w that overlaps the ith and $i + 1$st disjoint intervals that contains k or more points. Setting $B = \cup B_i$, $P(k, N, w) = P(A) + P(B)$. As Naus implicitly realizes, calculation of $P(B)$ becomes more complicated to the extent that more intersections of the B_i's have to be considered. The probabilities of intersections become rapidly more complicated, as the number of events increase, particularly for consecutive i's. By keeping k large relative to N, (i) the number of possible intersections of B_i is limited, and (ii) simpler methods can be used to calculate the probabilities needed. Specifically, for $k > N/2$, the only events possible are B_i and $B_i \cap B_{i+1}$. As would not be noted until much later, under this restriction, the probability for the latter event is the sum of two, rather than six, terms.

In Naus (1965a), the derivation of $P(B_i)$ is based on the reflection principle, which finds the number of paths starting at one point, and ending at another, that exceeds a certain level. A more complicated argument is used to derive $P(B_i \cap B_{i+1})$ involving permuting both "free" and "fixed" points. A critical footnote explains why the complex argument breaks down when $k < N/2$, and thus there are too many "free" points. The arguments are subtle; the following extract from a key footnote explains why the method breaks down:

> Permuting a free $(i+2)$nd interval point with [one in] the ith interval or $(i+1)$st interval leads to a distinguishable arrangement. Permuting a free $(i+2)$nd interval point with a fixed [one in] the ith interval or $(i+1)$st interval does not lead to a distinguishable arrangement. However when a free $(i+2)$nd intervals point falls among fixed $i+2$ interval points, then one of the fixed $(i+1)$st interval points is free to vary. ...Note that this argument is only valid so long as there are more fixed (i+1)st interval points than free points...

It is a fortunate "coincidence" that the same restriction, $k \geq N/2$, accomplishes two objectives: (i) freeing one from considering three- and higher-way intersections, and (ii) limiting the two-way intersections to the case where there are enough "fixed" points to make this subtle argument work.

It would require an apparently unrelated corollary of Barton and Mallows (1965) to a theorem of Karlin and McGregor (1959) to resolve this apparently general intractable problem, or even the simpler one of merely finding $P(B_i \cap B_{i+1})$. To complete the derivation, the argument of a polynomial form is used to extend the derivations from $w = 1/L$, L an integer, to arbitrary w. A brief comment at the end of the article introduces, as a type of afterthought, a "counter problem," the unconditional problem based on a Poisson process. This topic apparently did not appear in the thesis, and is described in more detail first in Naus (1982). Here, the conditional and unconditional probabilities are related through a somewhat difficult-to-implement summation of conditional probabilities from $N = k$ to infinity.

1.2.2 Clustering in two dimensions, Naus (1965b)

In the same year, Naus published the first paper for the two-dimensional problem — which as noted was the cause of his research into the scan statistic, and one that has become more active recently. In this formulation of the problem, the scanning area is rectangular. Naus derives upper and lower bounds for the probability of a cluster, and shows that these bounds converge, as both the length and width of the scanning interval approach zero. He shows that, contrary to the two previous papers on the subject in the mid-1940s, the shape of the scanning interval does matter. An example in the paper concerns the probability that, given 5 ships within the same 20° longitude and 30° latitude, 4 out

of 5 of these ships are in a 10° longitude × latitude square of the ocean. (Presumably, given the funding history above, this was the basis of Naus's exposure to the problem.) He bounds this probability by (.0086, .0278) and indicates an approximation of 0.01.

1.2.3 Power comparisons, Naus (1966a)

This paper, titled "A Power Comparison of Two Tests of Non-Random Clustering," is longer and allows a fuller discussion of the problem with implications that foreshadow recent developments, and places the problem in the context of statistics as well as probability. It begins by contrasting the "scan test" with the "disjoint" test (previously termed the fixed grid) based on the maximum of events in disjoint "cells." (In the terminology above, the disjoint test hinges on the event A, while the scan test hinges on both A and B.) Both statistics test the null hypothesis of a uniform distribution against two alternatives suggestive of clustering. Perhaps the most important contribution, in light of later developments and current focus, is the result in Section 5, which shows that the scan statistic is a generalized likelihood ratio test of the null hypothesis against the alternative that the density is a step function with two levels, a high constant density on $(b, b + w)$, b unknown, and a lower constant density elsewhere.

Equation (5.16) of this paper provides an asymptotic p-value, which Naus uses to prove consistency of the test based on the (yet unnamed) scan statistic. Naus also gives approximations for power, though these have not been exploited. The paper continues Naus's interest in the two-dimensional case, supplementing the previously described results with simulation, to construct a small table tabulating probabilities for $N = 10$.

Naus next opens up a new area of research, the Kolmogorov–Smirnov (KS) statistic. The simplest case of the scan that corresponds to the traditional KS statistic is based on $L = 2$, or equivalently, $w = 1/2$. The formulation is extendable to a version of the L-group problem, in which the groups can be ordered, and interest is focused on comparing adjacent groups. Naus notes that the KS problem can be expressed as a scan statistic problem conditional on cell occupancy numbers.

These first three papers, although they cover a substantially larger range than the topics in Naus (1963), can be thought of as a "fleshing-out" of the various ideas in the thesis.

1.2.4 Application of Karlin–McGregor (1959) theorem, Naus (1966b)

Naus's next paper, somewhat vaguely titled as "Some Probabilities, Expectations, and Variances for the Size of Largest Clusters and Smallest Intervals,"

solves the theoretical problem of calculating the distribution of the "scan statistic" and reduces much future work on the topic to further exploitations of these results. The essence of the paper is a six-line proof, which applied a corollary of Barton and Mallows (1965) concerning "election results" to the scan. The cited corollary concerns the amount of lead (extra votes) of candidate i over candidate $i+1$, in an L-candidate election. (To make sense of this problem, the candidates might have to be arranged in order, for example on the basis of very conservative to very liberal voting records, age, etc.) The corollary itself was based on a theorem of Karlin and McGregor (1959).

To one not immersed in the problems, the concern of the amount of lead in an election with L candidates does not appear to be closely related to a problem with the scan statistic. But based on a close reading of Naus's earlier papers, one can, in retrospect, see a conceptual connection. (As noted below, Karlin himself, when starting work on the scan in the 1990s, was not immediately aware of the connection.) Awareness of the connection, by Naus's account, was hastened when Arthur Cohen, a colleague at Rutgers, in a presumably offhand remark, pointed out the paper as something Naus might be interested in.

Naus (1966b) shows that, conditional on the cell occupancy numbers, the probability distribution of the scan statistic could be expressed as the summation of many determinants of an $L \times L$ matrix. (The summation was over all cell occupancy numbers, $\{n_1, n_2, ..., n_L\}$ with $n_i < k$.) Computational problems ensued both in computing such a determinant when L was large, and because the summation would be over $O(k^L)$ terms.

Based on this theorem, Naus gives an explicit formula for the distribution of the scan statistic when $w = 1/3$. Implicit in the formula was the basis of computing $P(B_i \cap B_{i+1})$, even when $k < N/2$. The paper tabulates some values based on computations and simulations by Naus's Ph.D. student, Richard Larsen. Implicit in such a limited tabulation, based in part on simulation, is the recognition that the formula was not, at that time, that easy to apply!

The title of the article alludes to a table of moments of the scan for $N = 1(1)10$, $w = .1(.1).9$. Curiously to today's reader (but easily explainable given the computer resources of the time), even though an exact formula had just been given for the window width of the form $w = 1/L$, L an integer (in particular for $w = 1/10$ and $w = 1/5$), these moments (as well as those for $w = 0.3, 0.4$, for which no formulas were available) were based on simulation. Whether realized at the time or not, at least by the computing standards of the 1960s (and no one has apparently bothered to see what the case would be today), the method was to serve more as an impetus for future theoretical work, rather than as a direct computational tool.

1.2.5 Birthday problem #1, Naus (1968)

On the surface, this paper is not directly linked to the scan statistic. The paper appears more related to Naus's interest in coincidences, to be realized in the Naus (1979) bibliography, than to his work on the scan. The article also reflects Naus's continuing interest in teaching. The purpose of the paper can be posed as engendering some interest in coincidences in a class so small that it is unlikely that two children have the same birthday. As Naus (1968) shows, even in a small class, it is still likely that two students have birthdays within a very few days of each other. For example, if the class size were 15, the probability would be only 0.223 of two (or more) children having birthdays on the same day, 0.537 of two children having birthdays on two consecutive days, and 0.957 of having birthdays within 6 days.

One aspect of the problem is ascribed to Mosteller, who as noted above was Naus's advisor. Naus also acknowledges Saul Blumenthal, who was his colleague at Rutgers. Naus and Blumenthal apparently communicated about common interests in "coincidences." A more formal connection between the birthday problem and the scan was not made until Naus (1974), possibly motivated by a cross-fertilization of ideas between Naus and Blumenthal's student at NYU, Saperstein, and maybe also Naus's student, Huntington.

At this point in his career, Naus was impacted by what seems to be a peculiar policy of the dean at the time. To justify promotion in the Statistics Department, it was strongly suggested that another institution should also offer a similar or higher level position. A policy of not promoting an institution's own Ph.D. graduates was more common then (and possibly now), but this extra "extension" of the policy appears unusual. In any case, Naus and Blumenthal both went to different New York institutions, with Naus, at least technically, on a leave of absence.

1.3 Joseph Naus's Students in 1967–1978, Exploitation of Ballot Problem Results, Broadening of Problem

Having been given the appropriate promotion at Baruch, the Business School of what was then City College, Naus returned to Rutgers from his leave of absence as an associate professor of statistics in 1967. He was subsequently promoted to professor in 1974. He was acting director of the Statistics Center of Rutgers in 1973–1974, and graduate director in 1974–1977. His third daughter, Julie, was born in 1968, and a son, Mark, was born in 1970.

Naus supervised eight Ph.D. theses from 1967 to 1978. Four theses focused entirely on the scan, and one partly so. Thus, during this time period, nearly an average of one Ph.D. student was graduating each year with Naus as the thesis supervisor, a substantial percentage of the output of the department at the time. While this account focuses on the scan statistic, and thus to students working on this topic, to the best of my memory, students (at least I) chose Naus as thesis advisor because of his overall reputation as a mentor, and his expertise in the area of applied probability, rather than the scan statistic specifically. In the early 1970s, Bell Labs and AT&T Long Lines were perceived as the major employers of potential graduates of the Rutgers program in statistics, and applied probability was perceived to be the subject area of their interest. (They continued to be a very attractive employer for my contemporaries; the policy ending with the deregulation of the phone companies in the early 1980s. The pharmaceutical industry was not, as yet, a major employer.) Naus's students (years of graduation) were Ed Wolf (1967), Larry Rabinowitz (1968), Richard Larsen (1970), Sylvan Wallenstein (1971), Mark Nicholich (1974), Ray Huntington (1974), Norman Neff (1978), and Joseph Glaz (1978).

Joe Naus took detailed interest in each student, and had at least weekly meetings with each student to discuss that student's work. In addition to these weekly meetings, the door to Joe's office was always open, and a student could just walk in and talk about the joint projects one had been working on. I can still remember the precise setup of his office. A non-statistical aspect, scheduling these meetings, has for some reason stuck in my mind. Joe often had car-pool responsibilities in the early afternoon, which probably based on both of our schedules, was an ideal time for us to talk, and he to car-pool. His two oldest girls were at that time enrolled at schools fairly close to both his home and his office (then in New Brunswick), traffic especially outside of rush hours was lighter than now, and the trips merely a relatively short distraction. By coincidence, scheduling of my single visit to talk to Joe about this article, revolved around his car-pooling a grandchild and also helping out Laura, who had just given birth.

To give some perspective to results that would have been available to graduate students in the late 1960s and very early 1970s, it should be noted that some additional work had begun to appear concerning the scan, to supplement the meager results previously available. Naus, who had a keen eye for connections to the scan, became aware of an abstract, Menon (1964) "Clusters in a Poisson Process," and the work of Newell (1963) which included an asymptotic result for a scan-like statistic, an apparently unrelated work, Ozols (1956), having to do with paths in three dimensions, and Ikeda (1965). Joe gave me this corpus of work, together with Ederer, Myers, and Mantel (1964), some more recent unpublished work by Mantel, and his thesis, when I became his student in perhaps the fall of 1969. At the time, paper copies, particularly of articles in relatively obscure journals and proceedings, were a valuable and hard-to-obtain

resource. There were other papers on coincidences possibly available at the time [e.g. Tackacs (1961)], which appear in Naus (1979), but it is unclear if Naus was aware of all the connections that would later emerge. This author is fairly confident that the term "scan statistic" had not been coined at the time, and finding a relevant paper would have taken much examination. (The lack of a good name for the problem, would explain the apparently uninformative titles of the papers appearing up to the late 1970s.).

The first three students and the fifth student worked in areas not directly linked to scan statistics. Naus's first student was Ed Wolf, who received his Ph.D. in 1967, and thus would have been doing research at about the same time Naus's above papers were published. The main published result, Wolf and Naus (1973), "Tables of Critical Values for a k-sample Kolmogorov-Smirnov Test Statistic," appeared in JASA. Since that paper focused on the L-sample Kolmogorov–Smirnov test, rather than the scan statistic, a detailed description is outside the purview of this article. In addition to this published work, Wolf contributed some methodology later used by Neff and Naus (1980) in their tabulation of probabilities, and started work on an approximation. Ed Wolf was later on the faculty at Baruch, where Naus had taught. Naus's next student was Richard Larsen (1970), who developed a non-parametric test. Naus' third student was Larry Rabinowitz, and the results of the dissertation, Rabinowitz and Naus (1975), in the *Annals of Probability*, concerned the moments of the number of components in random directed graphs. Mark Nicholich, who completed his dissertation in 1974, worked on the distributional pattern arising from multiple sources generating pollution according to a Gaussian plume model.

The overarching theme of the next group of four graduate students was research on the computation of exact probabilities of the scan statistic for the conditional, continuous case. Chapter 8 of Glaz, Naus, and Wallenstein (2001) contains a good unified description of these different results.

One result of Wallenstein's thesis is presented in Wallenstein and Naus (1974), which extended the range of N/k and of w for which calculations of exact probabilities could be performed relatively quickly. First, the results in Naus (1966b) concerning $P(B_i \cap B_{i+1})$ were slightly extended. Next, it was noted that for $2 < N/k < 3$, the additional terms that had to be computed were probabilities of three- and four-way intersections of the B_i's. If the intersections were consecutive, the probabilities could be obtained by the same theorem of Karlin and McGregor. If they were not, one could condition on cell occupancy numbers, obtain the probabilities, and sum over cell occupancy numbers, basically following the methods of Naus (1965). These observations essentially changed the problem from summing $O(k^L)$ determinants of $L \times L$ matrices to instead $O(k^3)$ summations of determinants of at most 5×5 matrices.

In Wallenstein and Naus (1973), the results in Naus (1966b) were extended to window widths that are rational numbers, i.e. $w = r/L$, r and L both integers.

To do so, r simultaneous processes were examined, which conditional on the cell occupancy numbers, were independent. As before, whether an event took place or not depended on the closeness of points to the beginning of the "cell."

Any possible usefulness of this paper would be limited, as it would be rendered obsolete by the work of Naus's next student, Huntington. Huntington and Naus (1975) greatly simplified calculations of exact probabilities when w was arbitrary. First, the exact probabilities for the scan statistic were generalized to any window widths, not just rational numbers. More importantly, instead of dealing with r simultaneous processes (for $w = r/L$), Huntington and Naus show that one only has to deal with two simultaneous processes. They use the same result involving the amount of lead in ballot problems. Huntington's thesis and subsequent research also shares Naus's renewed emphasis on the discrete scan, described in the next paragraph. The name Huntington would later have national prominence due to Huntington Learning Centers.

This sequence of joint papers with graduate students Naus mentored is now interrupted by Naus (1974), Naus's first formal foray into the birthday problem in the context of scan statistics, rather than coincidences. The informal genesis of this area can be attributed to Naus's casual conversations with Saul Blumenthal many years earlier. One of Blumenthal's students at NYU was Saperstein, who received his Ph.D. in 1969 and shortly thereafter published results [e.g. Saperstein (1972)] addressing the discrete problem using a different methodology than Naus had used. (Interestingly, the first use of the term "scan statistic" found in a Google search was in that paper, but by the chronology of subsequent use, it is doubtful that the paper is the source for the more frequent use of the term.)

Naus (1974) discussed the conditional problem: given A successes in N trials, what is the probability that some consecutive string of m trials contains k or more successes? Naus, dispensing with the buildup for the continuous case, gives the probability based on the same theorem of Karlin and McGregor (1959) via the corollary of Barton and Mallows (1965) for the quite broad case $N/m = L$, L an integer. He then extends the results to $N/m = L/r$, both L and r integers.

Towards the end of the paper, the case in which one does not condition on A is discussed. It would appear that this is now the more applied problem, but this was not then apparently obvious or true. In any case, Naus derives the unconditional probability by multiplying the conditional probabilities given cell occupancy numbers by the probability of observing these numbers, and adding over all possible cell occupancy numbers.

Naus's next student was Norman Neff, whose 1978 thesis found efficient ways to calculate the piecewise polynomial expressing $P(k, N, w)$ as a function of w. His thesis work, together with extensive collaboration with Naus, resulted in Neff and Naus (1980), a volume in "Selected Tables in Mathematical Statistics" published by IMS. The book is titled *"The Distribution of the Size of the*

Maximum Cluster of Points on the Line." It covered the cases $N < 20$ exhaustively. My copy of that book proved handy to several researchers in the field, who used it extensively. The book gives, for the first time, prominence to the unconditional problem, giving very extensive tables for the exact probability for this case. In addition to the exact probabilities, it gives coefficients for piecewise polynomials, as well as means and variances of the shortest intervals.

The book also gives an approximation for this unconditional case. This approximation was, as a departure from other approximations considered by Naus's students, a multiplicative one. Unfortunately, placing an approximation in a book devoted to exact values did not give it the exposure and influence it may have deserved. It is hard, at this point in time, to precisely evaluate the suggested approximation, but this type of thinking did eventually influence the Naus (1982) approximation. (Neff's thesis also used an additive approximation for the conditional case that was to morph into the Wallenstein–Neff (1987) approximation.) In any case, as with other work on the scan referenced above, whatever worth the approximation may have possessed became obsolete after Naus (1982). Neff then moved to the Department of Computer Sciences at Trenton State College.

Naus's last of the eight students in the period covered in this section was Joseph Glaz, with whom he continues to maintain a close collaboration, and who is actively involved in ongoing research on many aspects of the scan statistic. Glaz's (1978) thesis is titled "Multiple Coverage and Clusters on the Line," and the title suggests new directions for the scan that would be prominent in the 1980s. (In this section, we focus on work published before 1980.) Glaz and Naus (1979) derive the exact probability of covering the circle at least m times with a finite number of randomly placed arcs of equal length. This article solved an open problem for over thirty years, of multiple coverage of a circle by random arcs. Although Naus's students had worked with the problem on the circle and may have noted such research in their dissertations, this was the first paper focused on the topic.

We close this section by noting that Naus, together with colleagues, also started evaluating the closely linked problem of the distribution of the waiting time for a cluster in a stochastic process, which was to appear in Naus (1982). The topic was included in Glaz (1978), who continued work on the project. Naus also discussed this issue with Ester Samuel-Cahn when she was visiting Rutgers, resulting in Samuel-Cahn (1983). With the above papers and the publication of the tables in Neff and Naus (1980), the conceptual problem to calculate exact values was solved. The work had progressed in several directions. In terms of values of the triplets (k, N, and w), it had moved from the confines of the original solution in Naus (1965a) to larger value for N/k; it moved from $w = 1/L$, L an integer, to arbitrary values of w. In terms of the problems to be addressed, the scope of the problem had moved from the one- (and possibly two-)

dimensional continuous problem on the line conditional on the total number of events, to extensions to unconditional problems, the discrete version of the problem (but at this point mainly the conditional one), the circle, and waiting times.

Other investigations in the 1970s (not noted previously) with similar lines of research included Hwang (1977), several papers by Cressie such as Cressie (1977, 1979), and the limit results of Erdös and Rényi (1970). Methods for computations remained scattered over the literature, and were generally strewn with computational pitfalls, were others to attempt to apply them. Attention in the 1980s shifted to bounds, approximations, and other related directions of research.

1.4 Two Key Publications, 1979–1982

This section focuses on two papers by Naus that complete or complement the topics in the previous sections, specifically (1) approximations, (2) studying clusters and coincidences perse, beyond their connection with the scan, and (3) the relationship between different aspects of the problem.

1.4.1 Indexed bibliography

Dating back to his thesis, Naus had always shown an interest in the whole concept of coincidences, clusters, and clumps, irrespective of his work on the scan statistic. At a time when systematic collection of literature was difficult, and depended more on one's library (broadly defined) than on computer searches of the literature, he initiated efforts to collect, and list in a manageable manner, the completed work concerning coincidences. According to Naus, this involved, in addition to searches of the literature (which were quite difficult at the time and limited to only a few library tools such as Citation Index), a personal correspondence with those in the field. In the paper noted below, Naus acknowledges Cox, Cressie, F.N. David, Daley, Fienberg, Getis, Hammersley, Huntington, Hwang, Mantel, Melzak, Mood, Mosteller, Newell, Neyman, Pielou, E. Rothman, Saperstein, Takacs, and Watson.

Naus's interest in the topic, together with the above correspondence, culminated in the 30-page "An indexed Bibliography of Cluster, Clumps and Coincidences," published in 1979 in the *International Statistical Review*. The one-sentence summary describes the work: "This bibliography brings together and indexes an extensive and widely scattered literature on the probability of clustering of points in time and space".

The bibliography features a listing of approximately 1000 articles alphabetically by author. The article begins with 150 topics arranged alphabetically;

for example, those starting with "b" include ballot problem, birthdays, blocks, branching, bunch, bundles, burst, busy period; and those with "e" include earthquakes, empty cells, encounters, entropy, epidemics. These 150 topics are then associated with from one to many references; in the overwhelming number of cases, these take a line or two each. For example, birthdays begins with A4, an article by Abramson in the 1970 *American Mathematical Monthly*. The short two-page introduction ends with: "One of the goals of preparing the index is to simulate the interchange of methodologies and results from different areas of applications".

Interestingly, in this bibliography, the word "scan" appears only in articles by Cressie, an unpublished work by Saperstein, and works by Naus. The term "scan statistic" appears only in articles by Cressie beginning in 1977.

1.4.2 Approximations

Simple approximations and bounds, given in Ph.D. theses referenced above, were known, but apparently not extensively utilized either in practice or in the literature. Although some of these approximations did a good job of estimating probabilities in the tail, they possibly all had the undesirable property of exceeding 1.0 on occasion (perhaps when the true probability was as small as .7), which would certainly make approximating means difficult.

Naus's landmark paper in 1982, like Naus (1966b), provided a new methodology, applicable in many contexts, and different from prior research, to approximate the distribution of the various scan statistics. In later papers, and in many contexts, this approximation continues to be used. Basically, the approximation has a type of Markovian property, stating that the probability of no cluster in an interval of width w depends mainly on the distribution of events in the previous interval of width w, but not on events further back in "time." (Different approximations developed later would use a related thought based on conditioning on m preceding events, m varying for different approximations.) Letting E_i be the event that there is no cluster (k or more points within an interval of width w) that begins in the ith subinterval, i.e. in $[(i-1)w, iw)$, the approximation is

$$P(E_{i+2}|E_{i+1} \cap E_i \cap E_{i-1}...) \approx P(E_{i+2}|E_{i+1}).$$

(In the notation above, $B_i = A^c \cap E_i^c$.)

This approximation allowed attention to be limited to $P(E_1)$ and $P(E_1 \cap E_2)$. The computation of these terms is similar to that of $P(B_1)$ and $P(B_1 \cap B_2)$, whose calculation is implicit in Naus (1965a) and Naus (1966b). The probabilities involving E's are a bit simpler than those involving B's because of the fewer restrictions on cell occupancy numbers. Thus, $P(E_1)$ (termed Q2) was based on a simple expression summed $O(k^2)$ times, and $P(B_1 \cap B_2)$, termed Q3, was the sum of a somewhat complicated expression over $O(k^3)$ terms.

Thus, for the first time, a formula was available that would have allowed scan probabilities to be relatively easily approximated for nearly all values of k, N, and w. Unlike most previously given approximations, the approximation would apparently not exceed 1.0 in any case. Had today's technology been available, the approximation would have been placed on a web site, and possibly some future work by Naus and others would have been unnecessary. (Sadly, there is no (well-known) web site for either this or any other approximation for the scan statistic.) Such a task would not be trivial, and would require some computational tricks familiar to workers in the field, including methods to evaluate the exact and cumulative binomial and Poisson distributions. In the absence of such a web site or similar resource, an unwary investigator would have to devote considerable time to programming Q2 and Q3, and take caution in numerical overflows for terms like $N!$. It is unclear to what extent future approximations surpassed the Naus (1982) approximation in accuracy, if the computational work were held constant. The concept behind the approximation continues to be used in more complicated contexts.

Naus actually motivates and applies this new approximation, not by the conditional case, which was the subject of most of the previous work, but by the unconditional problem and the waiting-time variant referred to previously. He then gives a possibly underappreciated approximation to the expected value of a waiting time. He applies a similar methodology to the unconditional discrete problem, termed here the generalized birthday problem. Naus then applies the approximation to the two problems previously noted: multiple coverage on the line, and on the circle. At the end of the paper, he notes that the methods can also be applied to the conditional problem, although he cautions that the approximation is "rougher." His caution in this regard seems to be a bit unfair to the approximation; there is little indication that it performs poorly except in the tail which is not of interest. Nevertheless, the approximation is better for the unconditional case.

1.5 Later Work, Briefly Noted

For the majority of the 1980s, Naus was chairman (1984–1986) or acting chairman (1981–1982, 1988–1989, 1992–1993) of the Department of Statistics at Rutgers. Naus was heavily involved in non-statistical issues at Rutgers: he was head of the University Senate Budget Committee, and of the Faculty Council Computer Policy Committee, and a member of the Strategic Planning Subcommittee on (implementing) Computing.

Joseph Naus became a Fellow of the American Statistical Association in 1998, in recognition for his work on scan statistics. The text of the award was:

For contributions to statistical theory and applications, particularly in the development of scan statistics and data editing techniques, and for leadership in promoting statistical science.

The multi-volume *Encyclopedia of Statistics* appeared in 1988, with the leading researchers in different fields being asked to contribute articles. Naus was naturally asked to contribute the article on the scan statistic, (Naus (1988)). A revised version, Glaz and Naus (2005), appeared in the second edition of the *Encyclopedia*. He also published an invited review of the scan statistic (Naus (2006)), in the *Handbook of Engineering Statistics*.

Naus had continued to work with Glaz on topics concerning the scan statistic. Glaz and Naus (1983) extended Glaz's Ph.D. work and investigated the topic of multiple clusters of ordered occurrences of events on a line. Two models were considered: the uniform model and the exponential inter-arrival times model. Exact formulae and approximations were derived for the expected number of clusters and the variance of the number of clusters. These approximations were based on using Markov-like approximations for intersections of dependent events. Glaz and Naus (1986) derived approximations and waiting times of first passage in a special Gaussian process.

Glaz and Naus (1991) began work on a different approach involving bounds. Bounds were potentially of great benefit, since exact values continued to be computationally difficult, and approximations were inherently unsatisfying since their accuracy could not generally be proven. Glaz and Naus (1991) also proposed more accurate approximations, and give an algorithm for implementing the calculations. The scope of the scan statistic was again broadened to include integer-valued observations, in addition to the continuous and discrete cases that had been addressed previously.

Naus began working, and meeting, usually in his home, with Glaz and Wallenstein in the early 1990s. The collaboration resulted in several papers as well as a book. Wallenstein, Glaz, and Naus (1993) used a variant of the Q2/Q3 approximation in Naus (1982) to obtain power for a pulse-like alternative, and gave special attention to simplifying the computations. Glaz, Naus, Roos, and Wallenstein (1994) proposed a compound Poisson approximation for the distribution of scan statistics, and showed that it is more accurate than previously given Poisson approximations. The derivation of the approximation is based on ordered m-spacings for independent and identically distributed uniform observations.

The collaboration with Glaz and Wallenstein culminated in a book *"Scan Statistics"* published by Springer-Verlag in 2001. The book was divided into two parts: "Methods and Applications" and "Scan Distribution Theory and its Developments." The exceptional aspect of the book, and the reason for the positive reviews, was the first part written almost completely by Naus. This first

half contains a wealth of applications of the scan in different areas, reflecting a more substantial review of these applications than the 1979 bibliography. Particularly interesting to this writer is a description of star clusters—a problem addressed by Fisher (1959), and one that reflects Naus's continuing interest in the two-dimensional problem. Naus's description of the problem goes back to 1767 when Reverend Mitchell noted the visual closeness of six of the stars of the Pleiades. This section of the book closes with a chapter on the use of the scan in DNA and protein sequence analysis.

In the 1990s, possibly motivated by the topic described above, Naus, in addition to his membership in the Statistics Department, joined the Computational Molecular Biology Group at Rutgers. In a description of that group, Naus's interests are listed as

> Probabilistic and statistical approaches to looking for unusual clustering of patterns within DNA and protein sequences, and measuring the unusualness of matching of patterns between multiple DNA sequences. Statistical approaches to data editing, survey sampling, and applied statistics.

Related to this working group, Naus also participated in activities at DIMACS, the Center for Discrete Mathematics and Theoretical Computer Science, a consortium of Rutgers, Princeton University, AT&T Bell Laboratories, and Bellcore. Naus was one of about five Rutgers faculty members of the Steering Committee for the DIMACS special year of 1994–1995, termed "Mathematical Support for Molecular Biology."

The work at DIMACS had become important in the 1990s, due to the effort of sequencing the human genome. As part of his interest in this area, Naus worked on the connection between the genetic sequencing problems and the unconditional, discrete scan statistic. Karlin, who had become very involved in this area, invited Naus to come to California to discuss his work on the scan—work that was motivated by Karlin's own prior work cited previously.

Naus's work in this area led to new mentorship roles with two Ph.D. students, Vatsala Karwe and Ke-Ning Sheng, who both received their Ph.D.'s in 1993. Karwe's thesis included work on the maximum net charge of DNA sequences. Some results were given in Karwe and Naus (1997).

Sheng's work, which began on matching DNA sequences, resulted in four joint publications, two of which are described here. Motivated by the problem of comparing two DNA sequences, Sheng and Naus (1994) give the probability distribution of the longest matching word in two different sequences. The sequences need not be perfectly aligned, and mismatches are allowed. Sheng and Naus (1996) describe the two-dimensional analogue of this problem in which R $N \times N$ lattices are scanned looking for a common rectangular set of letters using an analogue of the Naus (1982) Q2/Q3 approximation.

Naus kept on extending the scan in light of new potential applications. For example, Naus and Wartenberg (1997) introduce the double scan, which evaluates clustering of two related events. This can be applied both when there is no order to the events, as in quality-control applications, and when there is some order as in homicide/suicide clusters.

In the current decade, work on the scan owes much to Kulldorff. For example, Kulldorff (1997) describes how the scan statistic can be implemented to detect spatial clustering, and Kulldorff (2001) describes the implementation for surveillance. Part of Kulldorff's contribution could be viewed as extending the context for which the scan statistic is the optimal statistic for clustering. Naus (1966) had demonstrated optimality for the narrow one-dimensional case with a fixed window and a uniform density, while Cressie (1979) and Loader (1991) extended it somewhat for the case of w varying. Kulldorff and Williams (1997) make the computations easily accessible via a program, SaTScan, and Kulldorff and co-workers continuously (at times, ingeniously) updated and expanded this resource. The actual calculation of probabilities, based on simulation, goes into a completely different direction of research than the computation of exact values, approximations, and bounds described here. Exactly how to implement these simulation-based methods is often a difficult task; the difficulty is of a different nature than that described here. The huge advantage, of course, of the simulation-based approach is an incredibly wider scope of applications.

Some of Naus's research in this decade could be viewed as using probabilistic arguments to compute probabilities and approximations for applications broader than considered previously, but not as broad as allowed by these more recent methods. Naus and Wallenstein (2004) focus on finding p-values when the assumption of a constant window and perhaps a constant density is relaxed. Naus and Wallenstein (2006), using an analogue of the Naus (1982) Q2/Q3 approximation, apply the scan statistic or its variants to surveillance of bioterrorism, without the restriction of w fixed.

Naus continues to work on scan statistics, having been awarded a grant this summer (2008) by the National Security Agency. His latest Ph.D. student, though his work was not connected with the scan, was Ken Ganning in 2005.

Probably the best place to find twentieth century work on the scan summarized by a wide range of authors is in Glaz and Balakrishnan (1999), *"Scan Statistics and Applications."* Practically all of the authors cited Naus in their work, although Naus's contribution was not the theme of the volume. Other recent books are those by Balakrishnan and Koutras (2002) and Fu and Lou (2003).

We have already noted that the phrase "scan statistic" is not necessarily used when discussing Naus's work, and that the term was only used once from 1962 to 1977, and only by one author before 1980. A search on Google Scholar (on June 26, 2008) indicated 48 uses of "scan statistic" in the 1980s, 161 in

the 1990s, and 1160 in the current decade. The majority of these post-2000 citations (my best guess is about 1000) were based on "spatial clustering" or its variants, but 233 did not use the term "spatial," and examination of the titles suggests that about 200 did not deal at all with spatially related methods. Thus, research on the topic, both in terms of numbers and breadth of applications, continues to be robust. Much of this work owes its origin to Joe Naus's continuing examination of the problem.

References

1. Balakrishnan, N. and Koutras, M.V. (2002). *Runs and Scans with Applications*, Wiley, New York.

2. Barton, D.E. and Mallows, C.L. (1965). Some aspects of the random sequence, *Annals of Mathematical Statistics*, **36**, 236–260.

3. Berg, W. (1945). Aggregates in one- and two-dimensional random distributions, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **36**, 319–336.

4. Burnside W. (1928). *Theory of Probability*, Cambridge University Press, Cambridge.

5. Cressie, N. (1977). On some properties of the scan statistic on the circle and the line, *Journal of Applied Probability*, **14**, 272–283.

6. Cressie, N. (1979). An optimal statistic based on higher order gaps, *Biometrika*, **66**, 619–627.

7. Ederer, F., Myers, M.H., and Mantel, N. (1964). A statistical problem in space and time: Do leukemia cases come in clusters? *Biometrics*, **20**, 626–636.

8. Elteren, Van P.H. and Gerrits, H.J.M. (1961). Een wachtprobleem voorkomende bij drempelwaardemetingen aan het oof, *Statistica Neerlandica*, **15**, 385–401.

9. Erdös, P. and Rényi, A. (1970). On a new law of large numbers, *Journal d'Analyse Mathématique*, **23**, 103–111.

10. Feller, W. (1958). *An Introduction to Probability Theory and its Applications*, Vol. I, 2nd Edition, John Wiley & Sons, New York.

11. Fisher, R.A. (1959). *Statistical Methods and Scientific Inference*, Hafner, New York.

12. Fu, J.C. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and Its Applications*, World Scientific, Singapore.

13. Glaz, J. (1978). *Multiple Coverage and Clusters on the Line*, Ph.D. thesis, Rutgers University, New Brunswick, NJ.

14. Glaz, J. and Balakrishnan, N., Editors (1999). *Scan Statistics and Applications*, Birkhäuser, Boston, MA.

15. Glaz, J. and Naus, J. (1979). Multiple coverage of the line, *Annals of Probability*, **7**, 900–906.

16. Glaz, J. and Naus, J. (1983). Multiple clusters on the line, *Communications in Statistics—Theory and Methods*, **12**, 1961–1986.

17. Glaz, J. and Naus, J. (1986). Approximating probabilities of first passage in a particular Gaussian process, *Communications in Statistics*, **15**, 1709–1722.

18. Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *Annals of Applied Probability*, **1**, 306–318.

19. Glaz, J. and Naus, J. (2005). Scan Statistics and Applications, *Encyclopedia of Statistical Sciences*, 2nd Edition, S. Kotz, N. Balakrishnan, C.B. Read and B. Vidacovic, eds., 7463–7471, Wiley, New York.

20. Glaz, J., Naus, J., Roos, M., and Wallenstein, S. (1994). Poisson approximations for the distribution and moments of ordered m-spacings, *Journal Applied Probability*, **31A**, 271–281.

21. Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan Statistics*, Springer-Verlag, New York.

22. Greenberg, M., Naus, J., Schneider, D., and Wartenberg, D. (1991). Temporal clustering of homicide and suicide among 15–24 year old white and black Americans, *Ethnicity and Disease*, **1**, 342–350.

23. Huntington, R. and Naus, J.I. (1975). A simpler expression for kth nearest neighbor coincidence probabilities, *Annals of Probability*, **3**, 894–896.

24. Hwang, F.K. (1977). A generalization of the Karlin-McGregor theorem on coincidence probabilities and an application to clustering, *Annals of Probability*, **5**, 814–817.

25. Ikeda, S. (1965). On Bouman-Velden-Yamamoto's asymptotic evaluation formula for the probability of visual response in a certain experimental research in quantum biophysics of vision, *Annals of the Institute of Statistics and Mathematics*, **17**, 295–310.

26. Karlin, S. and McGregor, G. (1959). Coincidence probabilities, *Pacific Journal of Mathematics*, **9**, 1141–1164.

27. Karwe, V.V. and Naus, J. (1997). New recursive methods for scan statistic probabilities, *Computational Statistics and Data Analysis*, **23**, 389–402.

28. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics, A—Theory and Methods*, **26**, 1481–1496.

29. Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic, *Journal of Royal Statistical Society A*, **164**, 61–72.

30. Kulldorff, M. and Williams, G. (1997). *SaTScan v. 1.0, Software for the Space and Space-Time Scan Statistics*, National Cancer Institute, Bethesda, MD.

31. Loader, C. (1991). Large deviation approximations to the distribution of scan statistics, *Advances in Applied Probability*, **23**, 751–771.

32. Mack, C. (1948). An exact formula for $Q_k(n)$, the probable number of k-aggregates in a random distribution of n points, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **39**, 778–790.

33. Mack, C. (1950). The expected number of aggregates in a random distribution of n points, *Proceedings Cambridge Philosophical Society*, **46**, 285–292.

34. Menon, M.V. (1964). Clusters in a Poisson process [abstract], *Annals of Mathematical Statistics*, **35**, 1395.

35. Naus, J. (1962). The distribution of the maximum number of points on the line, *ASD Paper 8*.

36. Naus, J. (1963). *Clustering of Random Points in the Line and Plane*, Ph.D. thesis, Rutgers University, New Brunswick, NJ.

37. Naus, J. (1965a). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association*, **60**, 532–538.

38. Naus, J. (1965b). Clustering of random points in two dimensions, *Biometrika*, **52**, 263–267.

39. Naus, J. (1966a). A power comparison of two tests of non-random clustering, *Technometrics*, **8**, 493–517.

40. Naus, J. (1966b). Some probabilities, expectations, and variances for the size of largest clusters, and smallest intervals, *Journal of the American Statistical Association*, **61**, 1191–1199.

41. Naus, J. (1968). An extension of the birthday problem, *American Statistician*, **22**, 27–29.

42. Naus, J. (1974). Probabilities for a generalized birthday problem, *Journal of the American Statistical Association*, **69**, 810–815.

43. Naus, J. (1979). An indexed bibliography of clusters, clumps and coincidences, *International Statistical Review*, **47**, 47–78.

44. Naus, J. (1982). Approximations for distributions of scan statistics, *Journal of the American Statistical Association*, **77**, 177–183.

45. Naus, J. (1988). Scan statistics, *Encyclopedia of Statistical Sciences*, Vol. 8, 281–284, N.L. Johnson and S. Kotz, eds., Wiley, New York.

46. Naus, J. (2006). Scan Statistics, *Handbook of Engineering Statistics*, H. Pham, ed., Chapter 43, 775-790. Springer-Verlag, New York.

47. Naus, J. and Sheng K.N. (1996). Screening for unusual matched segments in multiple protein sequences, *Communications in Statistics: Simulation and Computation*, **25**, 937–952.

48. Naus, J. and Sheng K.N. (1997). Matching among multiple random sequences, *Bulletin of Mathematical Biology*, **59**, 483–496.

49. Naus, J. and Wartenberg D. (1997). A double scan statistic for clusters of two types of events, *Journal of the American Statistical Association*, **92**, 1105–1113.

50. Naus, J. and Wallenstein, S. (2004). Simultaneously testing for a range of cluster or scanning window sizes, *Methodology and Computing in Applied Probability*, **6**, 389–400.

51. Naus, J. and Wallenstein S. (2006). Temporal surveillance using scan statistics, *Statistics in Medicine*, **25**, 311–324.

52. Neff, N. and Naus, J. (1980). The distribution of the size of the maximum cluster of points on a line, *IMS Series of Selected Tables in Mathematical Statistics*, Vol. VI, AMS, Providence, RI.

53. Newell, G.F. (1963). Distribution for the smallest distance between any pair of kth nearest-neighbor random points on a line, *Time series analysis, Proceedings of a conference held at Brown University*, M. Rosenblatt editor, pp. 89–103, John Wiley & Sons, New York.

54. Ozols, V. (1956). Generalization of the theorem of Gnedenko-Korolyuk to three samples in the case of two one-sided boundaries, *Latvijas PSR Zinatnu Akad. Vestis*, **10** (111), 141–152.

55. Parzen, E. (1960). *Modern Probability Theory and its Applications*, John Wiley & Sons, New York.

56. Rabinowitz, L. and Naus, J. (1975). The expectation and variance of the number of components in random linear graphs, *Annals of Probability*, **3**, 159–161.

57. Samuel-Cahn, E. (1983). Simple approximations to the expected waiting time for a cluster of any given size for point processes, *Advances in Applied Probability*, **15**, 21–38.

58. Saperstein, B. (1972). The generalized birthday problem, *Journal of the American Statistical Association*, **67**, 425–428.

59. Sheng. K.N. and Naus, J. (1994). Pattern matching between two non-aligned random sequences, *Bulleting of Mathematical Biology*, **56**, 1143–1162.

60. Sheng, K.N. and Naus, J. (1996). Matching fixed rectangles in 2-dimensions, *Statistics and Probability Letters*, **26**, 83–90.

61. Silberstein, L. (1945). The probable number of aggregates in random distributions of points, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **36**, 319–336.

62. Takacs, L. (1961). On a coincidence problem concerning particle counters, *Annals of Mathematical Statistics*, **32**, 739–756.

63. Wallenstein S.R. and Naus, J. (1973). Probabilities for the kth nearest neighbor problem on the line, *Annals of Probability*, **1**, 188–190.

64. Wallenstein S. and Naus, J. (1974). Probabilities for the size of largest clusters and smallest intervals, *Journal of the American Statistical Association*, **69**, 690–697.

65. Wallenstein S., Naus, J., and Glaz, J. (1993). Power of the scan statistic for the detection of clustering, *Statistics in Medicine*, **12**, 1829–1843.

66. Wallenstein, S. and Neff, N. (1987). An approximation for the distribution of the scan statistic, *Statistics in Medicine*, **6**, 197–207.

67. Wolf E. and Naus, J. (1973). Tables of critical values for a k-sample Kolmogorov-Smirnov test statistic, *Journal of the American Statistical Association*, **68**, 994–997.

2

Precedence-Type Tests for the Comparison of Treatments with a Control

Narayanaswamy Balakrishnan[1] **and Hon Keung Tony Ng**[2,3]

[1] *Department of Mathematics and Statistics, McMaster University, Hamilton, Canada*
[2] *Department of Statistical Science, Southern Methodist University, Dallas, TX, USA*
[3] *Institute for Health Care Research and Improvement, Baylor Research Institute, Dallas, TX, USA*

Abstract: Precedence-type tests are proposed for comparing several treatments with a control. The null distributions of these test statistics are derived, and critical values for some combination of sample sizes are then presented. Next, the exact power function of these tests under the Lehmann alternative is derived and used to compare the power properties of the proposed test procedures. Finally, an example is presented to illustrate all the test procedures discussed here.

Keywords and phrases: Precedence test, Wilcoxon rank-sum test, life-testing, level of significance, power, Lehmann alternative

2.1 Introduction

In life-testing and reliability experiments, it is natural to compare several treatments with a standard treatment (control). For example, a manufacturer of electronic components may wish to compare $(k-1)$ new production processes with the standard process and then determine whether any of these new processes would produce more reliable components than the standard process. In many cases, the costs of production for the new processes are relatively high because they are under development, and so it would be desirable to have a statistical test procedure which allows the experimenter to make a decision early on in the life-test.

The precedence test, first proposed by Nelson (1963), is a distribution-free two-sample life-test (i.e., a special case when $k = 2$) based on the order of early failures. Assume that a random sample of n_1 units from distribution F_X and another independent sample of n_2 units from distribution F_Y are placed

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_2,

simultaneously on a life-testing experiment. Suppose the null hypothesis is that the two lifetime distributions are equal, and the alternative hypothesis of interest is that one distribution is stochastically larger than the other, say, F_X is stochastically larger than F_Y. This alternative corresponds to the situation wherein the Y-units are more reliable than the X-units. The experiment is terminated as soon as the r-th failure from the Y-sample is observed. Then, the precedence test statistic $P_{(r)}$ is defined simply as the number of failures from the X-sample that precede the r-th failure from the Y-sample. It is obvious that large values of $P_{(r)}$ lead to the rejection of the hypothesis that $F_X = F_Y$ and in favor of the above-mentioned alternative hypothesis. The precedence test will be useful (i) when a life-test involves expensive units as the units that had not failed could be used for some other testing purposes, and (ii) to make quick and reliable decisions early on in the life-testing experiment. Many authors have studied the power properties of the precedence test and have also proposed some alternative tests; see, for example, Eilbott and Nadler (1965), Shorack (1967), Nelson (1986, 1993), Lin and Sukhatme (1992), Balakrishnan and Frattina (2000), Balakrishnan and Ng (2001), Ng and Balakrishnan (2002, 2004), and van der Laan and Chakraborti (2001). A brief review of all these precedence-type tests is first presented in Section 2.2, while an elaborate discussion of precedence-type tests and their variants can be found in the review articles by Chakraborti and van der Laan (1996, 1997) and also in the recent book by Balakrishnan and Ng (2006).

In this work, different precedence-type test procedures are proposed for the k-sample problem. Specifically, suppose we have $(k-1)$ treatments that we wish to compare with a control, or $(k-1)$ new processes that we wish to compare with the standard process. With $F_1(x)$ denoting the lifetime distribution associated with the control (or the standard process) and $F_{i+1}(x)$ denoting the lifetime distribution associated with the i-th treatment (or the i-th new process) for $i = 1, 2, \ldots, k-1$, our null hypothesis is simply

$$H_0 : F_1(x) = F_2(x) = \cdots = F_k(x) \text{ for all } x. \tag{2.1}$$

We are specifically concerned with a stochastically ordered alternative of the form

$$H_1 : \{F_2(x) \le F_1(x)\} \cup \{F_3(x) \le F_1(x)\} \cup \cdots \cup \{F_k(x) \le F_1(x)\} \text{ for all } x, \text{ with at least one holding strictly for some } x. \tag{2.2}$$

Suppose k independent random samples of sizes $n_1, n_2, \ldots, n_k$ from $F_1(x)$, $F_2(x), \ldots, F_k(x)$, respectively, are placed simultaneously on a life-testing experiment. The experiment is terminated as soon as the r-th failure from $F_1(x)$ is observed. Then, the number of failures from $F_i(x)$, $i = 2, \ldots, k$, in between the failures from $F_1(x)$ are counted and their functions are used as test statistics for testing the hypothesis in (2.1).

The chapter is organized as follows. In Section 2.2, we review some results on the precedence-type tests which are considered in the subsequent sections. In Section 2.3, we propose the precedence-type tests, which include tests based on the precedence, weighted maximal precedence and minimum Wilcoxon rank-sum precedence test statistics, for testing the hypothesis in (2.1). The exact null distributions of the proposed test statistics are derived in Section 2.3, and critical values for some selected choices of sample sizes are also tabulated. Exact power properties of these tests under Lehmann alternatives are derived in Section 2.4. We then compare the power properties of the proposed precedence-type tests under Lehmann alternatives. Finally, an example is presented to illustrate all the tests discussed here.

2.2 Review of Precedence-Type Tests

The precedence-type test allows a simple and robust comparison of two distribution functions. Suppose there are two failure time distributions F_X and F_Y and that we are interested in testing

$$H_0^* : F_X = F_Y \quad \text{against} \quad H_1^* : F_X > F_Y. \tag{2.3}$$

Note that some specific alternatives such as the location-shift alternative and the Lehmann alternative are subclasses of the stochastically ordered alternative considered in (2.3).

Assume that a random sample of n_1 units from distribution F_X and another independent sample of n_2 units from distribution F_Y are placed simultaneously on a life-testing experiment. Let $X_1, \ldots, X_{n_1}$ denote the sample from F_X, and $Y_1, \ldots, Y_{n_2}$ denote the sample from F_Y. Let us denote the order statistics from the X- and Y-samples by $X_{1:n_1} \leq \cdots \leq X_{n_1:n_1}$ and $Y_{1:n_2} \leq \cdots \leq Y_{n_2:n_2}$, respectively. Further, let M_1 denote the number of X-failures before $Y_{1:n_2}$ and M_i the number of X-failures between $Y_{i-1:n_2}$ and $Y_{i:n_2}$, $i = 2, 3, \ldots, r$. Figure 2.1 gives a schematic representation of this precedence setup.

Note here that the idea of precedence-type test is closely related to that of a run, which is defined as an uninterrupted sequence. Wald and Wolfowitz (1940) used runs to establish a two-sample test for testing the hypothesis in (2.3). They suggested that one should combine the two samples, arrange the $n_1 + n_2$ observations in increasing order of magnitude, and replace the ordered values by 0 or 1 depending on whether it originated from the X-sample or the Y-sample, respectively. For example, in Figure 2.1, we have a binary sequence (1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1). Then, the total number of runs in that binary sequence is used as a test statistic to test the hypothesis in (2.3). Instead of

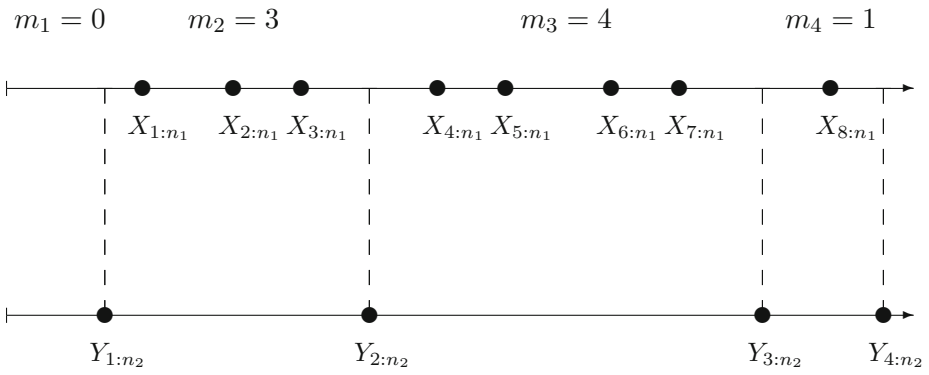


Figure 2.1. Schematic representation of a precedence life-test.

using the number of runs in the binary sequence, the precedence-type tests use the length of the runs of 0's (i.e., M_i, $i = 1, \ldots, n_2$) and their functions as test statistics for testing the hypotheses in (2.3). For extensive reviews on runs and applications, one may refer to Balakrishnan and Koutras (2002) and Fu and Lou (2003).

2.2.1 Precedence test

The precedence test statistic $P_{(r)}$ is defined simply as the number of failures from the X-sample that precede the r-th failure from the Y-sample, i.e.,

$$P_{(r)} = \sum_{j=1}^{r} M_j.$$

Large values of $P_{(r)}$ lead to the rejection of H_0^* and in favor of H_1^* in (2.3). In other words, H_0^* is rejected if $P_{(r)} \geq s$, where s is the critical value of the precedence test statistic for specific values of n_1, n_2, r and level of significance (α). For example, from Figure 2.1, with $r = 4$, the precedence test statistic takes on the value $P_{(4)} = \sum_{i=1}^{4} M_i = 0+3+4+1 = 8$. If we have $n_1 = n_2 = 10$ and we use the precedence test with $r = 4$, the near 5% critical value will be $s = 8$ with exact level of significance 0.035, in which case H_0^* would be rejected if there were at least 8 failures from the X-sample before the fourth failure from the Y-sample. Therefore, the null hypothesis that the two distributions are equal is rejected based on the precedence test in this example.

From Balakrishnan and Ng (2006, Theorem 4.1), we have the joint probability mass function of $(M_1, \ldots, M_r)$, under $H_0^* : F_X = F_Y$, to be

$$\Pr\left(M_1 = m_1, M_2 = m_2, \ldots, M_r = m_r \mid H_0 : F_X = F_Y\right) = \frac{\binom{n_1 + n_2 - \sum_{j=1}^{r} m_j - r}{n_2 - r}}{\binom{n_1 + n_2}{n_2}}. \tag{2.4}$$

The null distribution and critical values of the precedence test statistic $P_{(r)}$ can be readily computed from (2.4). The critical values and their exact levels of significance (as close as possible to 5% and 10%) for different choices of r and the sample sizes n_1 and n_2 are presented, for example, in Balakrishnan and Ng (2006).

2.2.2 Weighted maximal precedence test

Balakrishnan and Frattina (2000) observed that a masking effect is present in the precedence test which has an adverse effect on its power properties. The maximal precedence test proposed by Balakrishnan and Frattina (2000) and Balakrishnan and Ng (2001) was specifically to avoid this masking problem. It is a test procedure based on the maximum number of failures occurring from the X-sample before the first, between the first and the second, ... , between the $(r-1)$-th and the r-th failures from the Y-sample. Then, Ng and Balakrishnan (2005) proposed the weighted maximal precedence test by giving a decreasing weight to m_j as j increases, which is given by

$$M_{(r)} = \max_{1 \le j \le r} (n_2 - j + 1) M_j. \tag{2.5}$$

It is also a test procedure suitable for testing the hypotheses in (2.3) with large values of $M_{(r)}$ leading to the rejection of H_0^* and in favor of H_1^* in (2.3). The null distribution of the weighted maximal precedence test statistic $M_{(r)}$ can also be obtained from (2.4). The critical values and their exact levels of significance (as close as possible to 5% and 10%) for different choices of r and the sample sizes n_1 and n_2 are presented, for example, in Balakrishnan and Ng (2006). For example, if we refer to Figure 2.1, with $r = 4$ and with $n_1 = n_2 = 10$, the critical value is 42 with exact level of significance 0.043 and the weighted maximal precedence test statistic is $M_{(4)} = \max(10 \times 0, 9 \times 3, 8 \times 4, 7 \times 1) = \max(0, 27, 32, 7) = 32$. Therefore, the null hypothesis that the two distributions are equal is not rejected based on the weighted maximal precedence test in this example.

2.2.3 Minimal Wilcoxon rank-sum precedence test

The Wilcoxon rank-sum test is a well-known nonparametric procedure for testing the hypotheses in (2.3) based on complete samples. For testing the

hypotheses in (2.3), if complete samples of sizes n_1 and n_2 are available from F_X and F_Y, respectively, one can use the standard Wilcoxon's rank-sum statistic, proposed by Wilcoxon (1945), which is simply the sum of ranks of X-observations in the combined sample.

Ng and Balakrishnan (2002, 2004) proposed the Wilcoxon-type rank-sum precedence tests for testing the hypotheses in (2.3) in the context of precedence test described earlier, i.e., when the Y-sample is Type-II right censored. This test is a variation of the precedence test and a generalization of the Wilcoxon rank-sum test. In order to test the hypotheses in (2.3), instead of using the maximum of the frequencies of failures from the X-sample between the first r failures of the Y-sample, one could use the sum of the ranks of those failures. More specifically, suppose that $M_1, M_2, \ldots, M_r$ denote the number of X-failures that occurred before the first, between the first and the second, ... , between the $(r-1)$-th and the r-th Y-failures, respectively; see Figure 2.1. Let W be the rank-sum of the X-failures that occurred before the r-th Y-failure. The Wilcoxon's rank-sum test statistic will be smallest when all the remaining $\left(n_1 - \sum_{j=1}^{r} M_j\right)$ X-failures occur between the r-th and $(r+1)$-th Y-failures. The test statistic in this case would be

$$\begin{aligned} W_{(r)} &= W + \left[\left(\sum_{j=1}^{r} M_j + r + 1\right) + \left(\sum_{j=1}^{r} M_j + r + 2\right) + \cdots + (n_1 + r)\right] \\ &= \frac{n_1(n_1 + 2r + 1)}{2} - \sum_{j=1}^{r} (r - j + 1) M_j. \end{aligned}$$

This is called the *minimal rank-sum statistic*. Note that in the special case of $r = n_2$ (that is, when we observe a complete sample), $W_{(n_2)}$ is equivalent to the classical Wilcoxon's rank-sum statistic. Small values of $W_{(r)}$ lead to the rejection of H_0^* and in favor of H_1^* in (2.3). The null distribution of the minimal Wilcoxon-type rank-sum precedence test statistic can once again be obtained from (2.4). The critical values and their exact levels of significance (as close as possible to 5% and 10%) for different choices of r and the sample sizes n_1 and n_2 are presented, for example, in Balakrishnan and Ng (2006).

For example, from Figure 2.1, when $n_1 = n_2 = 10$ and $r = 4$, we have

$$W_{(4)} = 2 + 3 + 4 + 6 + 7 + 8 + 9 + 11 + 13 + 14 = 77$$

and the critical value of the test is 81 with exact level of significance 0.050. Therefore, the null hypothesis that the two distributions are equal is not rejected based on the minimal Wilcoxon rank-sum precedence test in this example.

Ng and Balakrishnan (2002, 2004) observed that the large-sample normal approximation for the null distribution of these statistics is not satisfactory in

the case of small or moderate sample sizes. For this reason, they developed an Edgeworth expansion to approximate the significance probabilities. They also derived the exact power function under the Lehmann alternative and examined the power properties of the minimal Wilcoxon-type rank-sum precedence test.

2.3 Test Statistics for Comparing $k-1$ Treatments with Control

Suppose k independent random samples of sizes $n_1, n_2, \ldots, n_k$ from $F_1(x)$, $F_2(x), \ldots, F_k(x)$, respectively, are placed simultaneously on a life-testing experiment. When the sample sizes are all equal, we have a balanced case which usually provides a favorable setting for carrying out a precedence-type procedure for testing H_0 in (2.1) against the alternative in (2.2); however, the test can be carried out even in the unbalanced case, although the power of the test may be adversely affected in this case.

A precedence-type test procedure, for this specific testing problem, may be constructed as follows. After pre-fixing an r ($\leq n_1$), the life-test continues until the r-th failure in the sample from the control group. We then observe $\boldsymbol{M}_2 = (M_{12}, M_{22}, \ldots, M_{r2}), \ldots, \boldsymbol{M}_k = (M_{1k}, M_{2k}, \ldots, M_{rk})$ from the $(k-1)$ treatments, where $M_{1i}, M_{2i}, \ldots, M_{ri}$ are the numbers of failures in the sample from the $(i-1)$-th treatment (for $i = 2, 3, \ldots, k$) before the first failure, between the first and second failures, ... , and between the $(r-1)$-th and r-th failures from the control group, respectively. The observed value of $\boldsymbol{M}_i$ is denoted by $\boldsymbol{m}_i$, $i = 2, \ldots, k$.

2.3.1 Tests based on precedence statistic

Let us consider

$$P_{(r)i} = \sum_{j=1}^{r} M_{ji} \quad \text{for} \quad i = 2, 3, \ldots, k \tag{2.6}$$

for the precedence statistic corresponding to the sample from the $(i-1)$-th treatment. For convenience of notation, let $M_{j\bullet} = \sum_{i=2}^{k} M_{ji}$ and denote its observed value by $m_{j\bullet}$, $j = 1, \ldots, r$. We may then propose the following precedence-type test statistics:

$$P_1 = \sum_{i=2}^{k} P_{(r)i} = \sum_{i=2}^{k} \sum_{j=1}^{r} M_{ji} = \sum_{j=1}^{r} M_{j\bullet} \tag{2.7}$$

and

$$P_2 = \min_{2\le i\le k} P_{(r)i} = \min_{2\le i\le k} \left\{ \sum_{j=1}^{r} M_{ji} \right\}. \tag{2.8}$$

The rationale for the use of the statistics in (2.7) and (2.8) is that, under the stochastically ordered alternative H_1 in (2.2), we would expect some of the precedence statistics $P_{(r)i}$ in (2.6) to be too small. Consequently, we will tend to reject H_0 in (2.1) in favor of H_1 in (2.2) for small values of P_1 and P_2 in which the critical values can be determined for specific values of $k, r, n_i, i = 1, 2, \ldots, k$, and pre-fixed level of significance α. Specifically, $\{0 \le P_1 \le c_{P_1}\}$ and $\{0 \le P_2 \le c_{P_2}\}$ will serve as critical regions, where c_{P_1} and c_{P_2} are determined such that

$$\Pr(P_1 \le c_{P_1}|H_0) = \alpha \quad \text{and} \quad \Pr(P_2 \le c_{P_2}|H_0) = \alpha. \tag{2.9}$$

The null distributions of the test statistics P_1 and P_2 can be expressed as

$$\begin{aligned}&\Pr(P_1 = p_1|H_0)\\ &\quad = \sum_{p_{(r)2}=0}^{n_2} \cdots \sum_{p_{(r)k}=0}^{n_k} \Pr(P_{(r)i} = p_{(r)i}, i = 2, \ldots, k|H_0) I\left(\sum_{i=2}^{k} p_{(r)i} = p_1\right)\end{aligned} \tag{2.10}$$

for $p_1 = 0, 1, \ldots, \sum_{i=2}^{k} n_i$, and

$$\begin{aligned}&\Pr(P_2 = p_2|H_0)\\ &\quad = \sum_{p_{(r)2}=0}^{n_2} \cdots \sum_{p_{(r)k}=0}^{n_k} \Pr(P_{(r)i} = p_{(r)i}, i = 2, \ldots, k|H_0) I\left(\min_{2\le i\le k} p_{(r)i} = p_2\right)\end{aligned} \tag{2.11}$$

for $p_2 = 0, 1, \ldots, \min_{2\le i\le k} n_i$, where $I(A)$ is the indicator function defined by

$$I(A) = \begin{cases} 1 \text{ if } A \text{ is true,} \\ 0 \text{ otherwise,} \end{cases}$$

and

$$\begin{aligned}&\Pr(P_{(r)i} = p_{(r)i}, i = 2, \ldots, k|H_0)\\ &\quad = \sum_{\boldsymbol{m}_2} \cdots \sum_{\boldsymbol{m}_k} \delta(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k) I\left(\sum_{j=1}^{r} m_{ji} = p_{(r)i}, i = 2, \ldots, k\right)\end{aligned} \tag{2.12}$$

with

$$\sum_{\boldsymbol{m}_i} \overset{def.}{=} \sum_{m_{1i}=0}^{n_i} \sum_{m_{2i}=0}^{n_i - m_{1i}} \cdots \sum_{m_{ri}=0}^{n_i - \sum_{j=1}^{r-1} m_{ji}} \quad \text{for } i = 2, \ldots, k$$

and $\delta(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k)$ is the probability mass function of $(\boldsymbol{M}_2, \ldots, \boldsymbol{M}_k)$ under H_0 (see Appendix A)

$$\begin{aligned}\delta(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k) &= \Pr(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k | H_0 : F_1 = F_2 = \cdots = F_k) \\ &= \frac{1}{\binom{\sum_{i=1}^k n_i}{n_1, \ldots, n_k}} \left\{ \prod_{j=1}^r \binom{m_{j\cdot}}{m_{j2}, \ldots, m_{jk}} \right\} \\ &\quad \times \binom{\sum_{i=1}^k n_i - \sum_{j=1}^r m_{j\cdot} - r}{n_1 - r, n_2 - \sum_{j=1}^r m_{j2}, \ldots, n_k - \sum_{j=1}^r m_{jk}},\end{aligned}$$

where

$$\binom{a_1 + \cdots + a_l}{a_1, \ldots, a_l} = \frac{(a_1 + \ldots + a_l)!}{a_1! \ldots a_l!}.$$

From Equations (2.9)–(2.12), the critical values c_{P_1}, c_{P_2} and their exact levels of significance as close as possible to $\alpha = 5\%$ for $k = 3, 4$ with equal sample sizes $n_1 = \cdots = n_k = n$ and $r = 4(1)n$ were computed and are presented in Tables 2.1 and 2.2; similarly, for the unequal sample sizes $n_1 = 10, n_2 = \cdots = n_k = 15$; $n_1 = 15, n_2 = \cdots = n_k = 20$ and $r = 4(1)n_1$, the values are presented in Tables 2.3 and 2.4. Due to the heavy computational demand in going through all the possible outcomes, the critical values of the tests discussed in this section were obtained from the exact null distribution for $r \leq 8$ and through 20,000,000 Monte Carlo simulations for $r > 8$.

2.3.2 Tests based on weighted maximal precedence statistic

We can proceed similarly and propose weighted maximal precedence-type statistics for the testing problem discussed here. Once again, we terminate the life-test when the r-th failure occurs in the sample from the control group. Then, with $\boldsymbol{M}_i = (M_{1i}, M_{2i}, \ldots, M_{ri})$, for $i = 2, \ldots, k$, being observed from the $(k-1)$ treatments, where M_{ji} denotes the number of failures in the sample from the $(i-1)$-th treatment between the $(j-1)$-th and j-th failures from the control group, we may set

$$M_{(r)i} = \max_{1 \leq j \leq r} (n_1 - j + 1) M_{ji} \quad \text{for} \quad i = 2, 3, \ldots, k$$

for the weighted maximal precedence statistic corresponding to the sample from the $(i-1)$-th treatment. We may then propose the weighted maximal precedence-type test statistics as

$$T_1 = \sum_{i=2}^k M_{(r)i} = \sum_{i=2}^k \max_{1 \leq j \leq r} (n_1 - j + 1) M_{ji} \tag{2.13}$$

Table 2.1. Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 3$, $n_1 = n_2 = n_3 = n = 10, 15$ and 20.

$n = 10$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	1	0.031	0	0.079	10	0.047	0	0.079	186	0.058	95	0.079
5	3	0.056	0	0.031	17	0.047	6	0.050	202	0.051	104	0.050
6	4	0.039	1	0.052	19	0.045	7	0.050	216	0.049	112	0.045
7	6	0.045	2	0.063	21	0.048	8	0.061	228	0.048	119	0.044
8	8	0.045	3	0.062	22	0.050	8	0.043	237	0.052	124	0.052
9	11	0.062	4	0.051	23	0.051	8	0.037	244	0.051	128	0.054
10	13	0.038	5	0.029	23	0.051	8	0.037	248	0.050	131	0.048
$n = 15$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	1	0.036	0	0.090	15	0.053	0	0.090	357	0.042	180	0.090
5	3	0.068	0	0.040	26	0.047	0	0.040	383	0.046	195	0.040
6	4	0.052	1	0.073	30	0.053	11	0.042	407	0.052	208	0.042
7	5	0.039	1	0.033	35	0.047	12	0.043	430	0.050	220	0.048
8	7	0.048	2	0.046	39	0.052	13	0.044	451	0.050	232	0.045
9	9	0.054	3	0.055	41	0.051	14	0.055	470	0.050	242	0.051
10	11	0.056	4	0.059	42	0.051	14	0.042	487	0.050	252	0.047
11	13	0.055	5	0.059	43	0.048	15	0.058	502	0.049	260	0.050
12	15	0.051	6	0.054	44	0.053	15	0.048	514	0.050	267	0.051
13	17	0.043	7	0.045	44	0.052	15	0.045	524	0.049	273	0.049
14	20	0.050	8	0.032	44	0.052	15	0.045	530	0.051	277	0.049
15	23	0.045	10	0.037	44	0.052	15	0.045	534	0.050	279	0.050
$n = 20$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	1	0.038	0	0.096	20	0.057	0	0.096	577	0.045	290	0.096
5	2	0.036	0	0.044	36	0.051	0	0.044	613	0.050	310	0.044
6	4	0.059	0	0.019	39	0.049	16	0.048	648	0.048	328	0.048
7	5	0.046	1	0.041	49	0.048	17	0.051	681	0.050	345	0.057
8	7	0.059	2	0.059	54	0.052	18	0.054	712	0.053	362	0.055
9	8	0.044	2	0.029	57	0.048	19	0.057	742	0.051	378	0.054
10	10	0.051	3	0.038	60	0.050	20	0.059	770	0.050	394	0.047
11	12	0.056	4	0.044	63	0.048	20	0.047	796	0.050	408	0.049
12	14	0.059	5	0.049	65	0.048	22	0.045	820	0.050	421	0.050
13	15	0.041	6	0.051	67	0.051	25	0.048	842	0.049	433	0.051
14	18	0.059	7	0.051	67	0.046	26	0.052	861	0.051	444	0.051
15	20	0.056	8	0.048	68	0.051	26	0.048	879	0.049	454	0.050
16	22	0.051	9	0.044	68	0.050	27	0.051	894	0.049	462	0.052
17	24	0.044	10	0.037	68	0.049	27	0.050	906	0.050	470	0.049
18	27	0.052	12	0.057	68	0.049	27	0.050	915	0.051	475	0.051
19	30	0.056	13	0.039	68	0.049	27	0.050	922	0.050	479	0.050
20	33	0.048	15	0.041	68	0.049	27	0.050	925	0.051	481	0.051

and

$$T_2 = \min_{2\le i\le k} M_{(r)i} = \min_{2\le i\le k} \left\{ \max_{1\le j\le r} (n_1 - j + 1) M_{ji} \right\}. \tag{2.14}$$

Here again, the rationale for the use of the statistics in (2.13) and (2.14) is that, under the stochastically ordered alternative H_1 in (2.2), we would expect some of the weighted maximal precedence statistics $M_{(r)i}$ in (2.12) to be too

Table 2.2. Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 4$, $n_1 = n_2 = n_3 = n_4 = n = 10, 15$ and 20.

$n = 10$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	3	0.052	0	0.109	22	0.050	0	0.109	278	0.050	95	0.109
5	5	0.050	0	0.044	28	0.051	0	0.044	301	0.051	105	0.044
6	7	0.042	1	0.073	32	0.049	6	0.043	322	0.047	113	0.043
7	10	0.049	1	0.027	35	0.050	7	0.040	339	0.049	120	0.045
8	13	0.048	2	0.030	36	0.049	8	0.061	353	0.049	126	0.045
9	17	0.058	4	0.070	37	0.052	8	0.053	363	0.049	130	0.050
10	21	0.052	5	0.040	37	0.052	8	0.053	368	0.051	133	0.046
$n = 15$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	3	0.058	0	0.125	30	0.052	0	0.125	533	0.056	180	0.125
5	5	0.060	0	0.056	42	0.048	0	0.056	572	0.050	195	0.056
6	7	0.056	0	0.024	52	0.051	11	0.059	608	0.052	208	0.059
7	9	0.049	1	0.047	57	0.048	12	0.061	642	0.050	221	0.050
8	12	0.059	2	0.064	62	0.050	12	0.039	673	0.049	233	0.048
9	14	0.047	2	0.028	65	0.048	13	0.039	701	0.050	244	0.048
10	17	0.050	3	0.033	68	0.052	14	0.059	726	0.049	254	0.047
11	20	0.049	4	0.035	69	0.049	14	0.043	747	0.051	262	0.053
12	23	0.045	5	0.033	70	0.052	14	0.035	765	0.050	270	0.049
13	27	0.052	7	0.062	70	0.051	15	0.065	779	0.051	276	0.049
14	31	0.054	8	0.044	70	0.050	15	0.064	789	0.050	280	0.050
15	35	0.042	10	0.050	70	0.050	15	0.064	794	0.051	282	0.052
$n = 20$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	3	0.062	0	0.132	39	0.048	0	0.132	864	0.046	290	0.132
5	4	0.039	0	0.063	57	0.052	0	0.063	917	0.054	310	0.063
6	6	0.040	0	0.028	71	0.051	15	0.044	969	0.050	329	0.044
7	9	0.057	1	0.057	79	0.050	16	0.046	1018	0.051	347	0.043
8	11	0.051	1	0.027	87	0.050	17	0.049	1065	0.050	364	0.047
9	13	0.044	2	0.041	93	0.050	18	0.051	1109	0.050	380	0.052
10	16	0.051	3	0.053	97	0.048	19	0.053	1150	0.050	396	0.047
11	19	0.056	4	0.062	101	0.050	20	0.066	1188	0.051	410	0.052
12	21	0.045	5	0.067	104	0.051	20	0.052	1224	0.049	424	0.049
13	24	0.046	5	0.033	106	0.052	22	0.049	1256	0.050	436	0.052
14	27	0.046	6	0.034	107	0.049	22	0.042	1285	0.050	448	0.049
15	30	0.044	8	0.066	108	0.052	24	0.054	1310	0.050	458	0.050
16	34	0.052	9	0.060	108	0.050	24	0.049	1332	0.050	467	0.049
17	37	0.045	10	0.051	108	0.050	25	0.050	1350	0.050	474	0.051
18	41	0.048	11	0.039	108	0.050	25	0.050	1364	0.050	480	0.050
19	45	0.045	13	0.053	108	0.050	25	0.050	1374	0.050	484	0.051
20	50	0.046	15	0.055	108	0.050	25	0.050	1379	0.050	486	0.051

small. Therefore, we would reject H_0 in (2.1) in favor of H_1 in (2.2) for small values of T_1 and T_2 in which the critical values can be determined for specific values of $k, r, n_i, i = 1, 2, \ldots, k$, and pre-fixed level of significance α. Specifically, $\{0 \le T_1 \le c_{T_1}\}$ and $\{0 \le T_2 \le c_{T_2}\}$ will serve as critical regions, where c_{T_1} and c_{T_2} are determined such that

$$\Pr(T_1 \le c_{T_1} | H_0) = \alpha \quad \text{and} \quad \Pr(T_2 \le c_{T_2} | H_0) = \alpha. \tag{2.15}$$

Table 2.3. Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 3$, $n_1 = 10$, $n_2 = n_3 = 15$ and $n_1 = 15$, $n_2 = n_3 = 20$.

$n_1 = 10$, $n_2 = n_3 = 15$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	3	0.052	0	0.031	19	0.050	7	0.052	353	0.050	179	0.052
5	5	0.050	1	0.041	25	0.050	8	0.046	376	0.051	192	0.041
6	7	0.042	2	0.040	28	0.050	9	0.039	397	0.047	203	0.048
7	10	0.049	3	0.033	30	0.049	10	0.054	414	0.049	213	0.046
8	13	0.048	5	0.050	32	0.052	10	0.037	428	0.049	221	0.046
9	17	0.058	7	0.058	32	0.049	12	0.063	438	0.049	226	0.053
10	21	0.052	9	0.048	32	0.049	12	0.063	443	0.051	230	0.048
$n_1 = 15$, $n_2 = n_3 = 20$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	2	0.041	0	0.048	27	0.050	0	0.048	575	0.043	290	0.048
5	4	0.052	1	0.072	35	0.047	12	0.047	609	0.050	308	0.047
6	6	0.055	1	0.029	41	0.051	13	0.046	641	0.052	325	0.051
7	8	0.054	2	0.035	45	0.050	14	0.043	671	0.052	341	0.054
8	10	0.049	3	0.036	49	0.048	16	0.051	699	0.050	356	0.053
9	12	0.043	4	0.034	52	0.051	18	0.045	724	0.050	370	0.050
10	15	0.051	6	0.060	54	0.053	21	0.053	746	0.051	382	0.052
11	18	0.057	7	0.049	55	0.050	21	0.045	765	0.051	393	0.051
12	20	0.042	8	0.036	55	0.047	22	0.053	781	0.051	402	0.051
13	24	0.056	10	0.046	56	0.054	22	0.051	794	0.050	410	0.048
14	27	0.047	12	0.050	56	0.054	22	0.051	803	0.050	415	0.049
15	31	0.043	14	0.041	56	0.054	22	0.051	808	0.049	418	0.048

Table 2.4. Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 3$, $n_1 = 10$, $n_2 = n_3 = n_4 = 15$ and $n_1 = 15$, $n_2 = n_3 = n_4 = 20$.

$n_1 = 10$, $n_2 = n_3 = n_4 = 15$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	5	0.047	0	0.043	32	0.049	0	0.043	528	0.050	180	0.043
5	8	0.045	1	0.056	40	0.051	7	0.038	562	0.051	192	0.058
6	12	0.051	2	0.054	45	0.047	9	0.054	592	0.051	204	0.050
7	16	0.051	3	0.045	49	0.049	9	0.033	618	0.049	214	0.050
8	20	0.044	5	0.067	51	0.050	10	0.052	638	0.051	222	0.053
9	26	0.055	6	0.039	51	0.047	10	0.042	653	0.050	229	0.046
10	32	0.050	9	0.063	51	0.047	11	0.042	661	0.050	232	0.050
$n_2 = 15$, $n_2 = n_3 = n_4 = 20$												
	P_1		P_2		T_1		T_2		W_1		W_2	
r	c_{P_1}	l.o.s.	c_{P_2}	l.o.s.	c_{T_1}	l.o.s.	c_{T_2}	l.o.s.	c_{W_1}	l.o.s.	c_{W_2}	l.o.s.
4	4	0.046	0	0.067	43	0.052	0	0.067	860	0.052	290	0.067
5	7	0.055	0	0.025	56	0.050	11	0.041	911	0.052	309	0.041
6	10	0.057	1	0.041	66	0.050	12	0.040	959	0.051	326	0.052
7	13	0.055	2	0.049	73	0.049	14	0.061	1004	0.049	343	0.046
8	16	0.050	3	0.050	79	0.050	15	0.056	1045	0.049	358	0.050
9	20	0.056	4	0.047	83	0.050	16	0.050	1082	0.049	372	0.051
10	23	0.046	5	0.041	85	0.049	18	0.050	1114	0.051	385	0.049
11	27	0.046	7	0.066	87	0.050	20	0.055	1143	0.050	396	0.050
12	32	0.054	8	0.049	88	0.051	20	0.049	1166	0.051	406	0.048
13	36	0.046	10	0.062	88	0.050	20	0.045	1185	0.050	413	0.051
14	42	0.056	11	0.035	88	0.050	21	0.055	1198	0.050	419	0.048
15	48	0.054	14	0.054	88	0.050	21	0.055	1205	0.050	422	0.048

The null distributions of the test statistics T_1 and T_2 can be expressed as

$$\begin{aligned}&\Pr(T_1 = t_1|H_0)\\ &= \sum_{m_{(r)2}=0}^{n_2} \cdots \sum_{m_{(r)k}=0}^{n_k} \Pr(M_{(r)i} = m_{(r)i}, i = 2, \ldots, k|H_0) I\left(\sum_{i=2}^{k} m_{(r)i} = t_1\right)\end{aligned} \tag{2.16}$$

for $t_1 = 0, 1, \ldots, \sum_{i=2}^{k} n_i$, and

$$\begin{aligned}&\Pr(T_2 = t_2|H_0)\\ &= \sum_{m_{(r)2}=0}^{n_2} \cdots \sum_{m_{(r)k}=0}^{n_k} \Pr(M_{(r)i} = m_{(r)i}, i = 2, \ldots, k|H_0) I\left(\min_{2\le i\le k} m_{(r)i} = t_2\right)\end{aligned} \tag{2.17}$$

for $t_2 = 0, 1, \ldots, \min_{2\le i\le k} n_i$, where $\Pr(M_{(r)i} = m_{(r)i}|H_0)$ is

$$\begin{aligned}&\Pr(M_{(r)i} = m_{(r)i}, i = 2, \ldots, k|H_0)\\ &= \sum_{\boldsymbol{m}_2} \cdots \sum_{\boldsymbol{m}_k} \delta(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k) I\left(\max_{1\le j\le r} (n_1 - j + 1) m_{ji} = m_{(r)i}, i = 2, \ldots, k\right).\end{aligned} \tag{2.18}$$

From Equations (2.15)–(2.18), the critical values c_{T_1}, c_{T_2} and their exact levels of significance as close as possible to $\alpha = 5\%$ for $k = 3, 4$ with equal sample sizes $n_1 = \cdots = n_k = n$ and $r = 4(1)n$ were computed and are presented in Tables 2.1 and 2.2; similarly, for the unequal sample sizes $n_1 = 10, n_2 = \cdots = n_k = 15$; $n_1 = 15, n_2 = \cdots = n_k = 20$ and $r = 4(1)n_1$, the values are presented in Tables 2.3 and 2.4.

2.3.3 Tests based on minimal Wilcoxon rank-sum precedence statistic

Similarly, we propose test procedures based on minimal Wilcoxon rank-sum precedence statistic for the testing problem discussed here. We set

$$W_{(r)i} = \frac{n_i(n_i + 2r + 1)}{2} - \sum_{j=1}^{r} (r - j + 1) M_{ji} \quad \text{for} \quad i = 2, 3, \ldots, k \tag{2.19}$$

for the minimal Wilcoxon rank-sum precedence statistic corresponding to the sample from the $(i-1)$-th treatment. We may then propose the minimal Wilcoxon rank-sum precedence statistics as

$$W_1 = \sum_{i=2}^{k} W_{(r)i}$$

and

$$W_2 = \max_{2\le i\le k} W_{(r)i}.$$

Under the stochastically ordered alternative H_1 in (2.2), we would expect some of the minimal Wilcoxon rank-sum precedence statistics $W_{(r)i}$ in (2.19) to be large. Therefore, we would reject H_0 in (2.1) in favor of H_1 in (2.2) for large values of W_1 and W_2 in which the critical values can be determined for specific values of $k, r, n_i, i = 1, 2, \ldots, k$, and pre-fixed level of significance α. Specifically, $\{W_1 \ge c_{W_1}\}$ and $\{W_2 \ge c_{W_2}\}$ will serve as critical regions, where c_{W_1} and c_{W_2} are determined such that

$$\Pr(W_1 \ge c_{W_1}|H_0) = \alpha \quad \text{and} \quad \Pr(W_2 \ge c_{W_2}|H_0) = \alpha. \tag{2.20}$$

The null distributions of the test statistics W_1 and W_2 can be expressed as

$$\begin{aligned}&\Pr(W_1 = w_1|H_0)\\ &= \sum_{w_{(r)2}=l_2}^{u_2} \cdots \sum_{w_{(r)k}=l_k}^{u_k} \Pr(W_{(r)i} = w_{(r)i}, i = 2, \ldots, k|H_0) I\left(\sum_{i=2}^{k} w_{(r)i} = w_1\right)\end{aligned} \tag{2.21}$$

for $w_1 = \min_{2\le i\le k} l_i, \ldots, \max_{2\le i\le k} u_i$, with $l_i = n_i(n_i+1)/2$, $u_i = (r+n_i)(r+n_i+1)/2 - r(r+1)/2$, and

$$\begin{aligned}&\Pr(W_2 = w_2|H_0)\\ &= \sum_{w_{(r)2}=l_2}^{u_2} \cdots \sum_{w_{(r)k}=l_k}^{u_k} \Pr(W_{(r)i} = w_{(r)i}, i = 2, \ldots, k|H_0) I\left(\max_{2\le i\le k} w_{(r)i} = w_2\right)\end{aligned} \tag{2.22}$$

for $w_2 = \min_{2\le i\le k} l_i, \ldots, \min_{2\le i\le k} u_i$, where $\Pr(W_{(r)i} = w_{(r)i}|H_0)$ is given by

$$\begin{aligned}&\Pr(W_{(r)i} = w_{(r)i}, i = 2, \ldots, k|H_0)\\ &= \sum_{\boldsymbol{m}_2} \cdots \sum_{\boldsymbol{m}_k} \delta(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k)\\ &\qquad \times I\left(\frac{n_i(n_i+2r+1)}{2} - \sum_{j=1}^{r}(r-j+1)m_{ji} = w_{(r)i}, i = 2, \ldots, k\right).\end{aligned} \tag{2.23}$$

From Equations (2.20)–(2.23), the critical values c_{W_1}, c_{W_2} and their exact levels of significance as close as possible to $\alpha = 5\%$ for $k = 3, 4$ with equal sample sizes $n_1 = \cdots = n_k = n$ and $r = 4(1)n$ were computed and are presented in

Tables 2.1 and 2.2; similarly, for the unequal sample sizes $n_1 = 10, n_2 = \cdots = n_k = 15$; $n_1 = 15, n_2 = \cdots = n_k = 20$ and $r = 4(1)n_1$, the values are presented in Tables 2.3 and 2.4.

2.4 Exact Power Under Lehmann Alternative

The Lehmann alternative $H_1 : [F_i(x)]^{\gamma_i} = F_1(x)$ for some γ_i, $i = 2, \ldots, k$, which was first proposed by Lehmann (1953), is a subclass of the alternative $H_1 : F_i(x) > F_1(x)$ when at least one $\gamma_i \in (0, 1)$ (see Gibbons and Chakraborti, 2003). In this section, we will derive an explicit expression for the power functions of the proposed test procedures under the Lehmann alternative.

When $\gamma_2 = \cdots = \gamma_k = \gamma$, for some $\gamma \in (0, 1)$, under the Lehmann alternative $H_1 : [F_i(x)]^{\gamma} = F_1(x)$, the probability mass function of $(\boldsymbol{M}_2, \ldots, \boldsymbol{M}_k)$ is (see Appendix B)

$$
\begin{aligned}
&\delta^*(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k) \\
&= \Pr(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k | H_1 : [F_i]^{\gamma} = F_1, i = 2, \ldots, k) \\
&= \frac{\gamma^r n_1!}{(n_1 - r)!} \left\{ \prod_{i=2}^{k} \binom{n_i}{m_{1i}, m_{2i}, \ldots, m_{ri}, n_i - \sum_{j=1}^{r} m_{ji}} \right\} \\
&\quad \times \left\{ \prod_{j=1}^{r-1} B\left(m_{1\bullet} + \ldots + m_{j\bullet} + j\gamma, m_{j+1\bullet} + 1\right) \right\} \\
&\quad \times \left\{ \sum_{l=0}^{n_1 - r} \binom{n_1 - r}{l} (-1)^l B\left(\sum_{j=1}^{r} m_{j\bullet} + (r + l)\gamma, \sum_{i=2}^{k} n_i - \sum_{j=1}^{r} m_{j\bullet} + 1 \right) \right\},
\end{aligned}
\tag{2.24}
$$

where $B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1} dx$ is the complete beta function. Note that the exact distribution of $(\boldsymbol{M}_2, \ldots, \boldsymbol{M}_k)$ under the general Lehmann alternative $H_1 : [F_k(x)]^{\gamma_k} = [F_{k-1}(x)]^{\gamma_{k-1}} = \cdots = [F_2(x)]^{\gamma_2} = [F_1(x)]$ can also be obtained. For the purpose of illustration, we present the result for $k = 3$ in Appendix B.

Under the Lehmann alternative, the probability mass functions of P_1, P_2, T_1, T_2, W_1 and W_2 can be computed from Equations (2.10), (2.11), (2.16), (2.17), (2.21) and (2.22), respectively, by replacing $\delta(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k)$ with $\delta^*(\boldsymbol{m}_2, \ldots, \boldsymbol{m}_k)$ in Equations (2.12), (2.18) and (2.23). Here, we computed the power values of the proposed test procedures for $k = 3, 4$ with $n_1 = \cdots = n_k = 10$, $\gamma = 0.2(0.2)1.0$, $i = 2, \ldots, k$. Note that when $\gamma = 1.0$, the power values are precisely the exact levels of significance. These results are presented in Tables 2.5 and 2.6.

Table 2.5. Power values under Lehmann alternative for $k = 3$, $n_1 = n_2 = n_3 = 10$, $r = 4(1)10$ and $\gamma_2 = \gamma_3 = \gamma = 0.2(0.2)1.0$.

$\gamma = 1.0$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.031	0.079	0.047	0.079	0.058	0.079
	5	0.056	0.031	0.047	0.050	0.051	0.050
	6	0.039	0.052	0.045	0.050	0.049	0.045
	7	0.045	0.063	0.048	0.061	0.048	0.044
	8	0.045	0.062	0.050	0.043	0.052	0.052
	9	0.062	0.051	0.051	0.037	0.051	0.054
	10	0.038	0.029	0.051	0.037	0.050	0.048
$\gamma = 0.8$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.084	0.163	0.112	0.163	0.147	0.163
	5	0.133	0.073	0.115	0.114	0.134	0.114
	6	0.095	0.113	0.105	0.114	0.132	0.106
	7	0.102	0.126	0.105	0.131	0.129	0.106
	8	0.094	0.118	0.104	0.094	0.134	0.120
	9	0.112	0.092	0.106	0.082	0.130	0.121
	10	0.065	0.051	0.106	0.081	0.125	0.108
$\gamma = 0.6$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.221	0.334	0.263	0.334	0.346	0.334
	5	0.302	0.181	0.271	0.261	0.329	0.261
	6	0.231	0.246	0.240	0.259	0.326	0.252
	7	0.229	0.255	0.229	0.277	0.319	0.252
	8	0.202	0.229	0.219	0.208	0.323	0.275
	9	0.208	0.174	0.219	0.181	0.311	0.273
	10	0.118	0.094	0.218	0.181	0.299	0.246
$\gamma = 0.4$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.524	0.632	0.564	0.632	0.691	0.632
	5	0.612	0.438	0.567	0.557	0.678	0.557
	6	0.514	0.515	0.510	0.548	0.673	0.554
	7	0.488	0.505	0.480	0.544	0.661	0.553
	8	0.422	0.445	0.445	0.435	0.657	0.575
	9	0.390	0.340	0.433	0.385	0.636	0.563
	10	0.224	0.186	0.432	0.383	0.617	0.522
$\gamma = 0.2$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.918	0.945	0.926	0.945	0.973	0.945
	5	0.940	0.859	0.917	0.923	0.970	0.923
	6	0.893	0.886	0.879	0.907	0.967	0.924
	7	0.858	0.859	0.845	0.877	0.962	0.920
	8	0.784	0.792	0.793	0.779	0.957	0.922
	9	0.703	0.658	0.757	0.712	0.948	0.910
	10	0.455	0.407	0.756	0.705	0.939	0.886

2.5 Discussion

The results in Tables 2.5 and 2.6 show that the test procedures can detect the difference between two distributions effectively in most cases early in the life-testing experiment. Note that the desired level of significance may be impossible to attain for some test statistics when r is small, especially for the tests based on

Table 2.6. Power values under Lehmann alternative for $k = 4$, $n_1 = \cdots = n_4 = 10, r = 4(1)10$ and $\gamma_2 = \gamma_3 = \gamma_4 = \gamma = 0.2(0.2)1.0$.

$\gamma = 1.0$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.052	0.109	0.050	0.109	0.050	0.109
	5	0.050	0.044	0.051	0.044	0.051	0.044
	6	0.042	0.073	0.049	0.043	0.047	0.043
	7	0.049	0.027	0.050	0.040	0.049	0.045
	8	0.048	0.030	0.049	0.061	0.049	0.045
	9	0.058	0.070	0.052	0.053	0.049	0.050
	10	0.052	0.040	0.052	0.053	0.051	0.046
$\gamma = 0.8$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.135	0.216	0.127	0.216	0.141	0.216
	5	0.128	0.102	0.128	0.102	0.146	0.155
	6	0.108	0.151	0.119	0.102	0.136	0.102
	7	0.114	0.063	0.118	0.094	0.140	0.107
	8	0.105	0.066	0.113	0.129	0.137	0.107
	9	0.110	0.122	0.118	0.113	0.128	0.116
	10	0.088	0.067	0.118	0.113	0.136	0.106
$\gamma = 0.6$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.326	0.416	0.309	0.416	0.358	0.416
	5	0.307	0.238	0.304	0.238	0.368	0.332
	6	0.263	0.310	0.279	0.244	0.351	0.247
	7	0.257	0.156	0.267	0.223	0.354	0.258
	8	0.225	0.152	0.251	0.269	0.344	0.255
	9	0.210	0.219	0.255	0.238	0.322	0.265
	10	0.152	0.120	0.254	0.238	0.330	0.244
$\gamma = 0.4$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.665	0.719	0.645	0.719	0.723	0.719
	5	0.633	0.525	0.621	0.525	0.733	0.645
	6	0.566	0.596	0.577	0.541	0.714	0.549
	7	0.533	0.381	0.541	0.489	0.710	0.564
	8	0.461	0.349	0.507	0.519	0.692	0.553
	9	0.400	0.400	0.500	0.468	0.662	0.556
	10	0.273	0.226	0.499	0.467	0.664	0.520
$\gamma = 0.2$	r	P_1	P_2	T_1	T_2	W_1	W_2
	4	0.964	0.969	0.960	0.969	0.981	0.969
	5	0.950	0.906	0.943	0.906	0.981	0.953
	6	0.918	0.921	0.916	0.914	0.977	0.922
	7	0.884	0.794	0.879	0.857	0.974	0.924
	8	0.814	0.730	0.841	0.833	0.968	0.913
	9	0.717	0.710	0.816	0.776	0.958	0.906
	10	0.510	0.456	0.815	0.772	0.956	0.882

extrema (viz., P_2, T_2 and W_2). For instance, for $k = 4$, $n_1 = n_2 = n_3 = n_4 = 20$ and $r = 4$, the minimum level of significance attainable by the tests based on P_2, T_2 and W_2 are all equal to 0.132. It is, therefore, not possible to test the hypotheses in (2.1) at 5% level in this setting based on P_2, T_2 and W_2. For this reason, the tests based on the extrema of the precedence statistics from the treatments may not be applicable for small values of r in practice.

From Tables 2.5 and 2.6, we can observe that the power values of the tests increase with the number of treatments (i.e., $k - 1$) as expected, but the power values do not increase with r under Lehmann alternatives. We can also see that

the tests based on precedence statistics (P_1 and P_2) suffer from the masking effect. In other words, the power values of P_1 and P_2 decrease as r increases and the information given by a larger value of r is thus being masked. The tests based on weighted maximal precedence statistics (T_1 and T_2) and minimal Wilcoxon rank-sum precedence statistics (W_1 and W_2) reduce the masking effect that affects the performance of P_1 and P_2.

In comparing the power performance of tests based on the sum of the precedence statistics from the treatments (viz., P_1, T_1 and W_1) with those based on the extrema of the precedence statistics from the treatments (viz., P_2, T_2 and W_2), we observe that the former have better power performance than the latter. Furthermore, among all the tests discussed here, the test based on the sum of minimal Wilcoxon rank-sum precedence statistics among treatments (viz., W_1) seems to give overall the best power performance under the Lehmann alternative, and hence is the one that we recommend for the problem discussed here.

Further, the decrease in power values with increasing r also suggests that the test procedures based on the order of early failures can be more powerful than those based on a complete sample. In fact, r ($\leq n_1$) need not be large to provide reliable comparison between treatments and the control. This can save both time and experimental units in a life-testing experiment, which are clear advantages of precedence-type tests. One may be interested in maximizing the power with respect to r, i.e., to determine the best choice of r in designing the experiment. When prior information about the alternative is available, this task can be achieved by comparing the power values for different values of r. For example, for $k = 4$, $n_1 = n_2 = n_3 = n_4 = 10$, if prior information suggests $\gamma = 0.4$ for the Lehmann alternative, we would recommend the use of W_1 with $r = 6$ based on the power values presented in Table 2.6.

2.6 Illustrative Example

Let us consider X_2, X_3 and X_1 samples to be the data on appliance cord life in flex tests 1, 2 and 3, respectively, of Nelson (1982, p. 510). These three tests were done using two types of cord, viz., B6 and B7, where flex tests 1 and 2 were done with cord type B6 and test 3 was done with cord type B7. Suppose cord B7 was the standard production cord and B6 was proposed as a cost improvement. We will then be interested in testing the equality of the lifetime distributions of these cords. For these data, we have $k = 3$, $n_1 = n_2 = n_3 = 12$. Had we fixed $r = 8$, the experiment would have stopped as soon as the eighth failure from the X_1-sample (cord B7) had been observed, i.e., at 128.7 hours. The data are presented in Table 2.7. The observed values of $(m_{1i}, \ldots, m_{8i})$ and the values of the statistics $P_{(8)i}$, $M_{(8)i}$ and $W_{(8)i}$, $i = 2, 3$, are presented in Table 2.8.

Table 2.7. Appliance cord life data from Nelson (1982, p. 510) (∗ denotes censored observations).

Test 1 (X_2) Cord B6	96.9	100.3	100.8	103.3	103.4	105.4	122.6	∗	∗	∗	∗	∗
Test 2 (X_3) Cord B6	57.5	77.8	88.0	98.4	102.1	105.3	∗	∗	∗	∗	∗	∗
Test 3 (X_1) Cord B7	72.4	78.6	81.2	94.0	120.1	126.3	127.2	128.7	∗	∗	∗	∗

Table 2.8. Values of $(m_{1i}, \ldots, m_{8i})$ and the statistics $P_{(8)i}$, $M_{(8)i}$ and $W_{(8)i}$ for $i = 2, 3$.

	m_{1i}	m_{2i}	m_{3i}	m_{4i}	m_{5i}	m_{6i}	m_{7i}	m_{8i}	$P_{(8)i}$	$M_{(8)i}$	$W_{(8)i}$
$i = 2$	0	0	0	0	6	1	0	0	7	48	147
$i = 3$	1	1	0	1	3	0	0	0	6	24	142

The near 5% critical values for $k = 3$, $n_1 = n_2 = n_3 = 12$ and $r = 8$ and their exact level of significance (in parentheses) for the test procedures discussed in the preceding sections are as follows:

P_1: 8 (0.061), P_2: 2 (0.033), T_1: 29 (0.048), T_2: 10 (0.044),
W_1: 317 (0.052), W_2: 164(0.056).

Then the test statistics and their p-values are

$P_1 = 13$ (p-value $= 0.363$), $P_2 = 6$ (p-value $= 0.491$),
$T_1 = 72$ (p-value $= 0.813$), $T_2 = 24$ (p-value $= 0.697$),
$W_1 = 289$ (p-value $= 0.398$), $W_2 = 147$ (p-value $= 0.507$),

and so we will not reject the null hypothesis that the lifetime distributions of these cords are equal. This means that the cord B6 is not better than the cord B7. Incidentally, this finding agrees with that of Nelson (1982), who analyzed these data by assuming a normal model.

Appendix A: Probability Mass Function of $(M_2, \ldots, M_k)$ Under the Null Hypothesis

Let the ordered failures from the control be $x_1 < x_2 < \cdots < x_r$. Consider the $(i-1)$-th treatment, conditional on the failures from the control. Then, the probability that there are m_{1i} failures from the treatment before x_1 and m_{ji} failures between x_{j-1} and x_j, $j = 2, \ldots, r$, is given by the multinomial probability

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_i = \boldsymbol{m}_i \middle| x_1, \ldots, x_r\right) \\
&\quad = \Pr\left(M_{1i} = m_{1i}, \ldots, M_{ri} = m_{ri} \middle| x_1, \ldots, x_r\right) \\
&\quad = \binom{n_i}{m_{1i}, \ldots, m_{ri}, n_i - \sum_{j=1}^{r} m_{ji}} \\
&\qquad\qquad \times [F_i(x_1)]^{m_{1i}} \left\{ \prod_{j=2}^{r} [F_i(x_2) - F_i(x_1)]^{m_{ji}} \right\} [1 - F_i(x_r)]^{\left(n_i - \sum\limits_{j=1}^{r} m_{ji}\right)}.
\end{aligned}
$$

For fixed values of $x_1 < x_2 < \cdots < x_r$, due to the independence of the samples from the $(k-1)$ treatments, we readily have the conditional joint probability as

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k \middle| x_1, \ldots, x_r\right) \\
&\quad = \left\{ \prod_{i=2}^{k} \binom{n_i}{m_{1i}, \ldots, m_{ri}, n_i - \sum_{j=1}^{r} m_{ji}} \right\} \\
&\qquad \times \left\{ \prod_{i=2}^{k} [F_i(x_1)]^{m_{1i}} \right\} \left\{ \prod_{i=2}^{k} \prod_{j=2}^{r} [F_i(x_j) - F_i(x_{j-1})]^{m_{ji}} \right\} \\
&\qquad \times \left\{ \prod_{i=2}^{k} [1 - F_i(x_r)]^{\left(n_i - \sum\limits_{j=1}^{r} m_{ji}\right)} \right\}.
\end{aligned}
$$

Now, we have the joint density of the first r order statistics from the control as

$$
f_{1,\ldots,r:n_1}(x_1, \ldots, x_r) = \frac{n_1!}{(n_1 - r)!} \left[\prod_{j=1}^{r} f_1(x_j) \right] [1 - F_1(x_r)]^{n_1 - r}, \quad x_1 < \ldots < x_r.
$$

As a result, we obtain the unconditional probability of $(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k)$ as

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k\right) \\
&\quad = C \int_{-\infty}^{\infty} \int_{-\infty}^{x_r} \cdots \int_{-\infty}^{x_2} \left\{ \prod_{i=2}^{k} [F_i(x_1)]^{m_{1i}} \right\} \left\{ \prod_{i=2}^{k} \prod_{j=2}^{r} [F_i(x_j) - F_i(x_{j-1})]^{m_{ji}} \right\} \\
&\qquad \times \left\{ \prod_{i=2}^{k} [1 - F_i(x_r)]^{\left(n_i - \sum\limits_{j=1}^{r} m_{ji}\right)} \right\} \\
&\qquad \times \left[\prod_{j=1}^{r} f_1(x_j) \right] [1 - F_1(x_r)]^{n_1 - r} dx_1 \cdots dx_r, \qquad (2.25)
\end{aligned}
$$

where

$$C = \frac{n_1!}{(n_1 - r)!} \prod_{i=2}^{k} \binom{n_i}{m_{1i}, \ldots, m_{ri}, n_i - \sum_{j=1}^{r} m_{ji}}.$$

Under the null hypothesis, $H_0 : F_1(x) = F_2(x) = \cdots = F_k(x)$, by denoting $m_{j\bullet} = \sum_{i=2}^{k} m_{ji}$, the expression in (2.25) becomes

$$\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k \middle| H_0\right) \\
&= C \int_{-\infty}^{\infty} \int_{-\infty}^{x_r} \cdots \int_{-\infty}^{x_2} \left\{ \prod_{i=2}^{k} [F_1(x_1)]^{m_{1i}} \right\} \left\{ \prod_{i=2}^{k} \prod_{j=2}^{r} [F_1(x_j) - F_1(x_{j-1})]^{m_{ji}} \right\} \\
&\quad \times \left\{ \prod_{i=2}^{k} [1 - F_1(x_r)] \left(n_i - \sum_{j=1}^{r} m_{ji} \right) \right\} \\
&\quad \times \left[\prod_{j=1}^{r} f_1(x_j) \right] [1 - F_1(x_r)]^{n_1 - r} dx_1 \cdots dx_r \\
&= C \int_{-\infty}^{\infty} \int_{-\infty}^{x_r} \cdots \int_{-\infty}^{x_2} [F_1(x_1)]^{m_{1\bullet}} \left\{ \prod_{j=2}^{r} [F_1(x_j) - F_1(x_{j-1})]^{m_{j\bullet}} \right\} \\
&\quad \times [1 - F_1(x_r)] \left(\sum_{i=1}^{k} n_i - \sum_{j=1}^{r} m_{j\bullet} - r \right) \left[\prod_{j=1}^{r} f_1(x_j) \right] dx_1 \cdots dx_r.
\end{aligned}$$

Upon setting $u_i = F_1(x_i)$ for $i = 1, \ldots, r$, the above expression becomes

$$\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k \middle| H_0\right) \\
&= C \int_{0}^{1} \int_{0}^{u_r} \cdots \int_{0}^{u_2} u_1^{m_{1\bullet}} \left[\prod_{i=2}^{k} (u_j - u_{j-1})^{m_{j\bullet}} \right] \\
&\quad \times (1 - u_r) \left(\sum_{i=1}^{k} n_i - \sum_{j=1}^{r} m_{j\bullet} - r \right) du_1 \cdots du_r.
\end{aligned}$$

Using the transformation $w_1 = u_1/u_2$, we have

$$\begin{aligned}
\int_{0}^{u_2} u_1^{m_{1\bullet}} (u_2 - u_1)^{m_{2\bullet}} du_1 &= u_2^{m_{1\bullet} + m_{2\bullet}} \int_{0}^{1} w_1^{m_{1\bullet}} (1 - w_1)^{m_{2\bullet}} dw_1 \\
&= u_2^{m_{1\bullet} + m_{2\bullet} + 1} B(m_{1\bullet} + 1, m_{2\bullet} + 1),
\end{aligned}$$

where, as before, $B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1} dx$ is the complete beta function. Proceeding similarly and using the transformations $w_l = u_l/u_{l+1}$ for $l = 2, \ldots, r - 1$, we obtain

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_2=\boldsymbol{m}_2,\ldots,\boldsymbol{M}_k=\boldsymbol{m}_k|\,H_0\right)\\
&\quad=C\left\{\prod_{j=1}^{r-1}B\left(m_{1\bullet}+\cdots+m_{j\bullet}+j,m_{j+1\bullet}+1\right)\right\}\\
&\qquad\times\int_0^1 u_r^{\left(\sum_{j=1}^{r}m_{j\bullet}+r+1\right)}(1-u_r)^{\left(\sum_{i=1}^{k}n_i-\sum_{j=1}^{r}m_{j\bullet}-r\right)}du_r\\
&\quad=C\left\{\prod_{j=1}^{r-1}B\left(m_{1\bullet}+\cdots+m_{j\bullet}+j,m_{j+1\bullet}+1\right)\right\}\\
&\qquad\times B\left(\sum_{j=1}^{r}m_{j\bullet}+r,\sum_{i=1}^{k}n_i-\sum_{j=1}^{r}m_{j\bullet}-r+1\right)\\
&\quad=\frac{n_1!}{(n_1-r)!}\left\{\prod_{i=2}^{k}\binom{n_i}{m_{1i},\ldots,m_{ri},n_i-\sum_{j=1}^{r}m_{ji}}\right\}\\
&\qquad\times\frac{\left(\sum_{i=1}^{k}n_i-\sum_{j=1}^{r}m_{j\bullet}-r\right)!m_{1\bullet}!\ldots m_{r\bullet}!}{\left(\sum_{i=1}^{k}n_i\right)!}\\
&\quad=\frac{1}{\binom{\sum_{i=1}^{k}n_i}{n_1,\ldots,n_k}}\left\{\prod_{j=1}^{r}\binom{m_{j\bullet}}{m_{j2},\ldots,m_{jk}}\right\}\\
&\qquad\times\binom{\sum_{i=1}^{k}n_i-\sum_{j=1}^{r}m_{j\bullet}-r}{n_1-r,n_2-\sum_{j=1}^{r}m_{j2},\ldots,n_k-\sum_{j=1}^{r}m_{jk}}.
\end{aligned}
$$

Appendix B: Probability Mass Function of $(\boldsymbol{M_2},\ldots,\boldsymbol{M_k})$ Under the Lehmann Alternative

Under the Lehmann alternative H_1: $[F_k(x)]^{\gamma_k}=[F_{k-1}(x)]^{\gamma_{k-1}}=\cdots=[F_2(x)]^{\gamma_2}=F_1(x)$, for some $\gamma_i\in(0,1)$, the expression in (2.25) can be expressed as follows:

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k | H_1 : F_k^{\gamma_k} = \cdots = F_2^{\gamma_2} = F_1\right) \\
&\quad = C\gamma_k^r \int_{-\infty}^{\infty} \int_{-\infty}^{x_r} \cdots \int_{-\infty}^{x_2} \left\{ \prod_{i=2}^{k} [F_k(x_1)]^{m_{1i}\gamma_k/\gamma_i} \right\} \\
&\qquad \times \left\{ \prod_{i=2}^{k} \prod_{j=2}^{r} [F_k^{\gamma_k/\gamma_i}(x_j) - F_k^{\gamma_k/\gamma_i}(x_{j-1})]^{m_{ji}} \right\} \\
&\qquad \times \left\{ \prod_{i=2}^{k} [1 - F_i^{\gamma_k/\gamma_i}(x_r)]^{\left(n_i - \sum_{j=1}^{r} m_{ji}\right)} \right\} \left[\prod_{j=1}^{r} F_k^{\gamma_k - 1}(x_j) \right] \\
&\qquad \times \left[\prod_{j=1}^{r} f_k(x_i) \right] [1 - F_k^{\gamma_k}(x_r)]^{n_1 - r} dx_1 \cdots dx_r. \qquad (2.26)
\end{aligned}
$$

In the special case when $\gamma_i = \gamma$ for $i = 2, \ldots, k$, the expression in (2.26) can be simplified as

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k | H_1 : F_k^{\gamma} = \cdots = F_2^{\gamma} = F_1\right) \\
&\quad = C\gamma^r \int_{-\infty}^{\infty} \int_{-\infty}^{x_r} \cdots \int_{-\infty}^{x_2} [F_k(x_1)]^{m_{1\bullet} + \gamma - 1} \\
&\qquad \times \left\{ \prod_{j=2}^{r} F_k^{\gamma-1}(x_j) [F_k(x_j) - F_k(x_{j-1})]^{m_{j\bullet}} \right\} \\
&\qquad \times [1 - F_k(x_r)]^{\left(\sum_{i=2}^{k} n_i - \sum_{j=1}^{r} m_{j\bullet}\right)} \left[\prod_{j=1}^{r} f_k(x_i) \right] [1 - F_k^{\gamma}(x_r)]^{n_1 - r} dx_1 \cdots dx_r.
\end{aligned}
$$

Upon setting $u_i = F_k(x_i)$ for $i = 1, \ldots, r$, the above expression becomes

$$
\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \ldots, \boldsymbol{M}_k = \boldsymbol{m}_k | H_1 : F_k^{\gamma} = \cdots = F_2^{\gamma} = F_1\right) \\
&\quad = C\gamma^r \int_{0}^{1} \int_{0}^{u_r} \cdots \int_{0}^{u_2} u_1^{m_{1\bullet} + \gamma - 1} \left\{ \prod_{j=2}^{r} u_j^{\gamma-1} (u_j - u_{j-1})^{m_{j\bullet}} \right\} \\
&\qquad \times (1 - u_r)^{\left(\sum_{i=2}^{k} n_i - \sum_{j=1}^{r} m_{j\bullet}\right)} (1 - u_r^{\gamma})^{n_1 - r} dx_1 \cdots dx_r.
\end{aligned}
$$

Adopting an approach similar to the one used in Appendix A, we obtain

$$\begin{aligned}
&\Pr\left(\boldsymbol{M}_2=\boldsymbol{m}_2,\ldots,\boldsymbol{M}_k=\boldsymbol{m}_k|H_1:F_k^{\gamma}=\cdots=F_2^{\gamma}=F_1\right)\\
&\quad=C\gamma^r\left\{\prod_{j=1}^{r-1}B\left(m_{1\cdot}+\cdots+m_{j\cdot}+j\gamma,m_{j+1\cdot}+1\right)\right\}\\
&\qquad\times\int_0^1 u_r^{\left(\sum_{j=1}^{r}m_{j\cdot}+r\gamma+1\right)}(1-u_r)^{\left(\sum_{i=2}^{k}n_i-\sum_{j=1}^{r}m_{j\cdot}\right)}(1-u_r^{\gamma})^{n_1-r}du_r\\
&\quad=C\gamma^r\left\{\prod_{j=1}^{r-1}B\left(m_{1\cdot}+\cdots+m_{j\cdot}+j\gamma,m_{j+1\cdot}+1\right)\right\}\\
&\qquad\times\left[\sum_{l=0}^{n_1-r}\binom{n_1-r}{l}(-1)^l\right.\\
&\qquad\qquad\left.\times\int_0^1 u_r^{\left(\sum_{j=1}^{r}m_{j\cdot}+r\gamma+1+l\gamma\right)}(1-u_r)^{\left(\sum_{i=2}^{k}n_i-\sum_{j=1}^{r}m_{j\cdot}\right)}\right]du_r\\
&\quad=C\gamma^r\left\{\prod_{j=1}^{r-1}B\left(m_{1\cdot}+\cdots+m_{j\cdot}+j\gamma,m_{j+1\cdot}+1\right)\right\}\\
&\qquad\times\sum_{l=0}^{n_1-r}\binom{n_1-r}{l}(-1)^l B\left(\sum_{j=1}^{r}m_{j\cdot}+(r+l)\gamma,\sum_{i=2}^{k}n_i-\sum_{j=1}^{r}m_{j\cdot}+1\right).
\end{aligned}$$

The exact distribution of $(\boldsymbol{M}_2,\ldots,\boldsymbol{M}_k)$, under the general Lehmann alternative $H_1:[F_k(x)]^{\gamma_k}=[F_{k-1}(x)]^{\gamma_{k-1}}=\cdots=[F_2(x)]^{\gamma_2}=F_1(x)$, can be derived in a similar manner by expanding each term by the binomial formula, and the final expression would then involve multiple summation. For purposes of illustration, we present the result for $k=3$. In this case, we have from Equation (2.26)

$$\begin{aligned}
&\Pr\left(\boldsymbol{M}_2=\boldsymbol{m}_2,\boldsymbol{M}_3=\boldsymbol{m}_3|H_1:F_3^{\gamma_3}=F_2^{\gamma_2}=F_1\right)\\
&\quad=C\gamma_3^r\int_{-\infty}^{\infty}\int_{-\infty}^{x_r}\cdots\int_{-\infty}^{x_2}[F_3(x_1)]^{m_{12}\gamma_3/\gamma_2}[F_3(x_1)]^{m_{13}}\\
&\qquad\times\left\{\prod_{j=2}^{r}[F_3^{\gamma_3/\gamma_2}(x_j)-F_3^{\gamma_3/\gamma_2}(x_{j-1})]^{m_{j2}}\right\}\left\{\prod_{j=2}^{r}[F_3(x_j)-F_3(x_{j-1})]^{m_{j3}}\right\}\\
&\qquad\times[1-F_3^{\gamma_3/\gamma_2}(x_r)]^{\left(n_2-\sum_{j=1}^{r}m_{j2}\right)}[1-F_3(x_r)]^{\left(n_3-\sum_{j=1}^{r}m_{j3}\right)}\\
&\qquad\times\left\{\prod_{j=1}^{r}[F_3(x_i)]^{\gamma_3-1}f_3(x_i)\right\}[1-F_3^{\gamma_3}(x_r)]^{n_1-r}dx_1\cdots dx_r.
\end{aligned}$$

Upon setting $u_i = F_3(x_i)$ for $i = 1, \ldots, r$, the preceding expression becomes

$$\begin{aligned}
&\Pr\left(\boldsymbol{M}_2 = \boldsymbol{m}_2, \boldsymbol{M}_3 = \boldsymbol{m}_3 | H_1 : F_3^{\gamma_3} = F_2^{\gamma_2} = F_1\right) \\
&\quad = C\gamma_3^r \int_0^1 \int_0^{u_r} \cdots \int_0^{u_2} u_1^{\left(\frac{m_{12}\gamma_3}{\gamma_2} + m_{13} + \gamma_3 - 1\right)} \\
&\qquad \times \left\{ \prod_{j=2}^{r} u_j^{\gamma_3 - 1} \left(u_j^{\gamma_3/\gamma_2} - u_{j-1}^{\gamma_3/\gamma_2}\right)^{m_{j2}} (u_j - u_{j-1})^{m_{j3}} \right\} \\
&\qquad \times \left(1 - u_r^{\gamma_3/\gamma_2}\right)^{\left(n_2 - \sum_{j=1}^{r} m_{j2}\right)} (1 - u_r)^{\left(n_3 - \sum_{j=1}^{r} m_{j3}\right)} \, du_1 \cdots du_r.
\end{aligned}$$

The first integral with respect to u_1 can be expressed as

$$\begin{aligned}
&\int_0^{u_2} u_1^{\left(m_{12}\frac{\gamma_3}{\gamma_2} + m_{13} + \gamma_3 - 1\right)} \left(u_2^{\gamma_3/\gamma_2} - u_1^{\gamma_3/\gamma_2}\right)^{m_{22}} (u_2 - u_1)^{m_{23}} \, du_1 \\
&\quad = \int_0^{u_2} u_1^{\left(\frac{m_{12}\gamma_3}{\gamma_2} + m_{13} + \gamma_3 - 1\right)} \\
&\qquad \times \left\{ \sum_{l_1=0}^{m_{22}} \binom{m_{22}}{l_2} (-1)^{l_1} u_2^{(m_{22} - l_1)\frac{\gamma_3}{\gamma_2}} u_1^{\left(\frac{l_1\gamma_3}{\gamma_2}\right)} \right\} (u_2 - u_1)^{m_{23}} \, du_1 \\
&\quad = u_2^{\left((m_{12} + m_{22})\frac{\gamma_3}{\gamma_2} + (m_{13} + m_{23}) + \gamma_3 - 1\right)} \\
&\qquad \times \sum_{l_1=0}^{m_{22}} \binom{m_{22}}{l_2} (-1)^{l_1} B\left((m_{12} + l_1)\frac{\gamma_3}{\gamma_2} + m_{13} + \gamma_3, m_{23} + 1\right).
\end{aligned}$$

Similarly, the j-th integral with respect to u_j $(j = 2, \ldots, r-1)$ becomes

$$\begin{aligned}
&u_{j+1}^{\left((m_{12} + \ldots + m_{(j+1)2})\frac{\gamma_3}{\gamma_2} + (m_{13} + \ldots + m_{(j+1)3}) + \gamma_3 - 1\right)} \\
&\quad \times \sum_{l_j=0}^{m_{(j+1)2}} \binom{m_{(j+1)2}}{l_j} (-1)^{l_j} \\
&\quad \times B\left((m_{12} + \cdots + m_{j2} + l_j)\frac{\gamma_3}{\gamma_2} + (m_{13} + \cdots + m_{j3})\gamma_3, m_{(j+1)3} + 1\right),
\end{aligned}$$

while the last integral with respect to u_r becomes

$$\begin{aligned}
&\int_0^{u_r} u_r^{\left(\left(\sum_{j=1}^{r} m_{j2}\right)\frac{\gamma_3}{\gamma_2}+\left(\sum_{j=1}^{r} m_{j3}\right)+\gamma_3-1\right)}(1-u_r^{\gamma_3/\gamma_2})^{\left(n_2-\sum_{j=1}^{r} m_{j2}\right)} \\
&\quad\times(1-u_r)^{\left(n_3-\sum_{j=1}^{r} m_{j3}\right)}(1-u_r^{\gamma_3})^{n_1-r}du_r \\
&= \sum_{l_r=0}^{n_2-\sum_{j=1}^{r} m_{j2}} \sum_{l=0}^{n_1-r} \binom{n_2-\sum_{j=1}^{r} m_{j2}}{l_r}\binom{n_1-r}{l}(-1)^{l_r+l} \\
&\quad\times\int_0^1 u_r^{\left(\left(\sum_{j=1}^{r} m_{j2}\right)\frac{\gamma_3}{\gamma_2}+\left(\sum_{j=1}^{r} m_{j3}\right)+\gamma_3-1+l_r\frac{\gamma_3}{\gamma_2}+l\gamma_3\right)}(1-u_r)^{n_3-\sum_{j=1}^{r} m_{j3}}du_r \\
&= \sum_{l_r=0}^{n_2-\sum_{j=1}^{r} m_{j2}} \sum_{l=0}^{n_1-r} \binom{n_2-\sum_{j=1}^{r} m_{j2}}{l_r}\binom{n_1-r}{l}(-1)^{l_r+l} \\
&\quad\times B\left(\left(\sum_{j=1}^{r} m_{j2}+l_r\right)\frac{\gamma_3}{\gamma_2}+\left(\sum_{j=1}^{r} m_{j3}\right)+(l+1)\gamma_3, n_3-\sum_{j=1}^{r} m_{j3}+1\right).
\end{aligned}$$

Combining all these expressions, we finally obtain

$$\begin{aligned}
&\Pr\left(\boldsymbol{M}_2=\boldsymbol{m}_2,\boldsymbol{M}_3=\boldsymbol{m}_3|H_1:F_3^{\gamma_3}=F_2^{\gamma_2}=F_1\right) \\
&= C\gamma_3^r \sum_{l_1=0}^{m_{22}}\cdots\sum_{l_{r-1}=0}^{m_{r2}} \sum_{l_r=0}^{n_2-\sum_{j=1}^{r} m_{j2}} \sum_{l=0}^{n_1-r}\left\{\prod_{j=2}^{r}\binom{m_{j2}}{l_{j-1}}\right\} \\
&\quad\times\binom{n_2-\sum_{j=1}^{r} m_{j2}}{l_r}\binom{n_1-r}{l}(-1)^{\left(\sum_{j=1}^{r} l_j+l\right)} \\
&\quad\times\left\{\prod_{j=2}^{r} B\left(\left(\sum_{l^*=1}^{j} m_{l^*2}+l_j\right)\frac{\gamma_3}{\gamma_2}+\left(\sum_{l^*=1}^{j} m_{l^*3}\right)\gamma_3, m_{(j+1)3}+1\right)\right\} \\
&\quad\times B\left(\left(\sum_{j=1}^{r} m_{j2}+l_r\right)\frac{\gamma_3}{\gamma_2}+\left(\sum_{j=1}^{r} m_{j3}\right)+(l+1)\gamma_3, n_3-\sum_{j=1}^{r} m_{j3}+1\right).
\end{aligned}$$

References

1. Balakrishnan, N., Koutras, M.V. (2002). *Runs and Scans with Applications*, John Wiley & Sons, New York.

2. Balakrishnan, N., Frattina, R. (2000). Precedence test and maximal precedence test, In Limnios, N., Nikulin, M. (Eds.), *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*, pp. 355–378, Birkhäuser, Boston, MA.

3. Balakrishnan, N., Ng, H. K. T. (2001). A general maximal precedence test, In Hayakawa, Y., Irony, T., Xie, M. (Eds.) *System and Bayesian Reliability—Essays in Honor of Prof. Richard E. Barlow on his 70th Birthday*, pp. 105–122, World Scientific, Singapore.

4. Balakrishnan, N., Ng, H. K. T. (2006). *Precedence-Type Tests and Applications*, John Wiley & Sons, Hoboken, NJ.

5. Chakraborti, S., van der Laan, P. (1996). Precedence tests and confidence bounds for complete data: An overview and some results, *The Statistician*, **45**, 351–369.

6. Chakraborti, S., van der Laan, P. (1997). An overview of precedence-type tests for censored data, *Biometrical Journal*, **39**, 99–116.

7. Eilbott, J., Nadler, J. (1965). On precedence life testing, *Technometrics*, **7**, 359–377.

8. Fu, J. C., Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and Its Applications*, World Scientific, Singapore.

9. Gibbons, J. D., Chakraborti, S. (2003). *Nonparametric Statistical Inference*, Fourth edition, Marcel Dekker, New York.

10. Lehmann, E. L. (1953). The power of rank tests, *Annals of Mathematical Statistics*, **24**, 23–42.

11. Lin, C. H., Sukhatme, S. (1992). On the choice of precedence tests, *Communications in Statistics—Theory and Methods*, **21**, 2949–2968.

12. Nelson, L. S. (1963). Tables of a precedence life test. *Technometrics*, **5**, 491–499.

13. Nelson, L. S. (1986). Precedence life test. In Kotz, S., Johnson, N. L. (Eds.), *Encyclopedia of Statistical Sciences*, 7, pp. 134–136, John Wiley & Sons, New York.

14. Nelson, L. S. (1993). Tests on early failures—The precedence life test, *Journal of Quality Technology*, **25**, 140–143.

15. Nelson, W. (1982). *Applied Life Data Analysis*, John Wiley & Sons, New York.

16. Ng, H. K. T., Balakrishnan, N. (2002). Wilcoxon-type rank-sum precedence tests: Large-sample approximation and evaluation, *Applied Stochastic Models in Business and Industry*, **18**, 271–286.

17. Ng, H. K. T., Balakrishnan, N. (2004). Wilcoxon-type rank-sum precedence tests, *Australia and New Zealand Journal of Statistics*, **46**, 631–648.

18. Ng, H. K. T., Balakrishnan, N. (2005). Weighted precedence and maximal precedence tests and an extension to progressive censoring, *Journal of Statistical Planning and Inference*, **135**, 197–221.

19. Shorack, R. A. (1967). On the power of precedence life tests, *Technometrics*, **9**, 154–158.

20. van der Laan, P., Chakraborti, S. (2001). Precedence tests and Lehmann alternatives, *Statistical Papers*, **42**, 301–312.

21. Wald, A., Wolfowitz, J. (1940). On a test whether two populations are from the same population, *Annals of the Institute of Statistical Mathematics*, **11**, 147–162.

22. Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics*, **1**, 80–83.

3

Extreme Value Results for Scan Statistics

Michael V. Boutsikas, Markos V. Koutras, and Fotios S. Milienos

Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece

Abstract: In the first part of this chapter we focus on the classical scan and multiple scan statistic, defined on a sequence of independent and identically distributed (i.i.d.) binary trials and review a number of bounds and approximations for their distributions which have been developed by the aid of distance measures. Moreover, we discuss briefly a number of asymptotic results that have been established by setting up appropriate conditions guaranteeing the convergence (to zero) of the distance measures' upper bounds. In the second part, we study a multiple scan statistic enumerating variable by considering a general threshold-based framework, defined on i.i.d. continuous random variables. More specifically, we first prove a compound Poisson approximation for the total number of fixed length overlapping moving windows containing a prespecified number of threshold exceedances. The classical scan and multiple scan statistic may be treated as a special case of this general model. Next we exploit the previous result to gain some new extreme value results for the scan enumerating statistic under the assumption that the continuous random variables belong to the maximum domain of attraction of one of the three extreme value distributions (Fréchet, reversed Weibull, Gumbel). Finally, we elucidate how the general results can be applied in a number of classical continuous distributions (Pareto, uniform, exponential and normal).

Keywords and phrases: Scan, multiple scan statistic, Poisson and compound Poisson approximation, Erdős–Rényi statistic, extreme value theory, maximum domain of attraction, moving sums and exceedances

3.1 Introduction

The discrete scan statistic $S_{n,k}$ in a sequence of n binary trials (1: success, 0: failure) has been defined as the maximum number of successes within any k consecutive trials (n and k are two positive integers with $k \leq n$). Due to its widespread

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_3,

applicability in an abundance of research areas such as quality control, actuarial science, reliability theory, molecular biology, etc. it has been an attractive subject of continuing research interest for the past few decades; see e.g. the monographs by Balakrishnan and Koutras (2002) Glaz, Naus and Wallenstein (2001) and the special issue edited by Glaz and Balakrishnan (1999).

An instance where $S_{n,k}$ arises in quite a natural way is in randomness tests when the null hypothesis of uniformity and independence of a sequence of binary observations X_i, $i = 1, 2, \ldots, n$ is to be tested against the alternative hypothesis of clustering of 1's due to local dependence between X_i, $i = 1, 2, \ldots, n$ or due to the existence of subsequences of consecutive X_i with $P(X_i = 1) > p$. As Glaz and Naus (1991) indicated, the generalized likelihood ratio test for checking the hypothesis of uniformity rejects the null hypothesis of uniformity whenever $S_{n,k} \geq c$, with the value of c being determined by the significance level of the test. Apparently, the evaluation of c such that a prespecified significance level is achieved calls for the distribution of the test statistic $S_{n,k}$. Since randomness tests are frequently applied to large data sets, theoretical developments related to the asymptotic distribution of $S_{n,k}$ (as $n, k \to \infty$) will play a primary role in the analysis of the test.

In an actuarial context, let us consider a portfolio with n daily claims and denote by X_i, $i = 1, 2, \ldots, n$ the binary variable describing whether the i-th claim exceeds a threshold $u > 0$ ($X_i = 1$) or not ($X_i = 0$). Then $S_{n,k}$ will describe the maximum number of "large claims" (i.e. claims exceeding threshold u) in a period of k consecutive days. The primary interest in this situation is also focused on extremely long periods ($n, k \to \infty$), and therefore one should look at the asymptotic distribution of $S_{n,k}$.

Exact results for the distribution of the scan statistic were discussed in Fu (2001), Balakrishnan and Koutras (2002) and Fu and Lou (2003). Since the evaluation of the exact distribution is computationally intractable, especially for large values of the parameters n, k, several approximations and bounds have been developed during the last decades.

Another random variable closely related to $S_{n,k}$ is the number of occurrences of k consecutive trials which contain at least r successes among them. If we denote by $W_{n,k,r}$ the resulting (overlapping) enumerating variable when a sequence of n trials is realized, it is clear that the probability mass function of $W_{n,k,r}$ at 0 coincides with the quantity $P(S_{n,k} < r)$ i.e.

$$P(W_{n,k,r} = 0) = P(S_{n,k} < r).$$

The statistic $W_{n,k,r}$ is referred to in the statistical literature under the name *multiple scan statistic*. Balakrishnan and Koutras (2002) have introduced two additional enumeration schemes for scan occurrence counting. They have used the terminology "type III enumeration" for the overlapping scheme', and the terms "type I" and "type II" for the non-overlapping counting schemes. In this chapter we shall restrict our discussion to the overlapping scheme only.

Although quite accurate approximations are available by now for the probability mass function of $W_{n,k,r}$ at 0 (for a review see Chen and Glaz (1999)), when the question comes to the whole distribution of $W_{n,k,r}$ the problem becomes extremely complex. Koutras and Alexandrou (1995) have described a method to obtain the exact distribution of $W_{n,k,r}$ by invoking a Markov chain embedding technique; however, this approach becomes unwieldy for k and r of moderate size, while its computational complexity for large k, r and n renders the whole procedure non-feasible. Therefore, the development of asymptotic results for the distribution of $W_{n,k,r}$ is of special interest.

In this chapter we review some asymptotic results on the scan and multiple scan statistic. In Section 3.2 we introduce all necessary notation and some notions that will be used in the subsequent sections. The scan and multiple scan statistic is introduced in a threshold-based framework, and the classical binary statistics are viewed as special cases of the general model. An alternative realization of the multiple scan statistic in terms of order statistics is given as well.

Section 3.3 deals with the (binary) discrete scan statistic and multiple scan statistic. We review first (Subsection 3.3.1) a number of bounds and approximations which have been developed by the aid of distance measures between discrete distributions. Note that we have confined ourselves to techniques offering error bounds, so that the establishment of convergence theorems will be easily achieved (Subsection 3.3.2). Section 3.3 is completed by providing a number of extreme value results for the discrete scan statistic (Subsection 3.3.3).

In Section 3.4 we present several results for the multiple scan statistic under the threshold exceedance framework. A compound Poisson approximation is first established (Subsection 3.4.1); this result is subsequently used in Subsection 3.4.2 to derive extreme value results for the multiple scan statistic under the assumption that the random variables (whose threshold exceedances are studied) are belonging to the maximum domain of attraction of the three classical extreme value distributions (Fréchet, reversed Weibull, Gumbel). Finally, in Subsection 3.4.3, we present applications of the general results in a number of typical continuous distributions (Pareto, uniform, exponential and normal).

3.2 Definitions and Notation

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of independent and identically distributed (i.i.d) continuous random variables with cumulative distribution function F and denote by $X_i(u)$ the indicator variable

$$X_i(u) = I_{(u,\infty)}(Y_i) = \begin{cases} 1, \text{ if } Y_i > u \\ 0, \text{ if } Y_i \leq u \end{cases}, \quad i = 1, 2, \ldots, n.$$

The quantity $u \in \Re$ is a fixed threshold which is exceeded by Y_i with probability

$$p = P(X_i(u) = 1) = E(X_i(u)) = P(Y_i > u) = \overline{F}(u),$$

where $\overline{F}$ denotes the tail probability of Y_i. Considering all the moving windows of length k in the sequence $Y_1, Y_2, \ldots, Y_n$, namely,

$$Y_i, Y_{i+1}, \ldots, Y_{i+k-1}, \quad i = 1, 2, \ldots, n-k+1,$$

we may introduce the k-scan exceedance process as follows:

$$S_k^{(i)}(u) = \sum_{j=i}^{i+k-1} X_j(u) = \sum_{j=i}^{i+k-1} I_{(u,\infty)}(Y_j), \quad i = 1, 2, \ldots, n-k+1.$$

Manifestly, $S_k^{(i)}$ denotes the number of random variables, among Y_i, Y_{i+1},..., Y_{i+k-1}, whose value exceeds threshold u, while

$$S_{n,k}(u) = \max_{1 \leq i \leq n-k+1} S_k^{(i)}(u) = \max_{1 \leq i \leq n-k+1} \sum_{j=i}^{i+k-1} X_j(u)$$

expresses the maximum number of exceedance occurrences among all possible moving windows of length k, in the sequence $Y_1, Y_2, \ldots, Y_n$. A closely related random variable is

$$W_{n,k,r}(u) = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_k^{(i)}(u)),$$

which enumerates the total number of overlapping moving windows of length k in which the threshold exceedances are at least r. It is plain from the previous definitions that

$$P(W_{n,k,r}(u) = 0) = P(S_{n,k}(u) < r).$$

There is an interesting connection between the variables $W_{n,k,r}(u)$, $S_{n,k}(u)$ and moving order statistics of the original sample $Y_1, Y_2, \ldots, Y_n$. To elucidate this, let us first arrange the observations of the i-th moving window (of length k) $Y_i, Y_{i+1}, \ldots, Y_{i+k-1}$ in descending order and denote by $Y_{r:k}^{(i)}$ the r-th larger observation, i.e.

$$Y_{1:k}^{(i)} \geq Y_{2:k}^{(i)} \geq \ldots \geq Y_{k:k}^{(i)}.$$

For fixed r, consider next the random variables $Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}$, arrange them again in descending order and denote by $Y_{m:r:k}$ the m-th larger among them, that is

$$Y_{1:r:k} \geq Y_{2:r:k} \geq \cdots \geq Y_{n-k+1:r:k}.$$

In particular, for $m = 1$ and $m = n - k + 1$ we may write

$$\begin{aligned} Y_{1:r:k} &= \max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}), \\ Y_{n-k+1:r:k} &= \min(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}). \end{aligned}$$

We mention that the parameter n has been suppressed in the last notation. If n is not fixed (e.g. if we wish to investigate the asymptotic behavior as $n \to \infty$), we shall use the notation $Y_{m:r:k}(n)$ instead of $Y_{m:r:k}$. The cumulative distribution function of $Y_{m:r:k}(n)$ can be readily expressed in terms of the distribution of $W_{n,k,r}(u)$ as follows:

$$\begin{aligned} P\left(Y_{m:r:k}(n) \leq u\right) &= P(\text{at most } m-1 \text{ of } Y_{r:k}^{(1)}, \ldots, Y_{r:k}^{(n-k+1)} \text{ exceed } u) \\ &= P\Big(\sum_{i=1}^{n-k+1} I_{(u,\infty)}(Y_{r:k}^{(i)}) < m\Big) \\ &= P\Big(\sum_{i=1}^{n-k+1} I_{[r,\infty)}\Big(\sum_{j=i}^{i+k-1} I_{(u,\infty)}(Y_j)\Big) < m\Big) \\ &= P\Big(\sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_k^{(i)}(u)) < m\Big) \\ &= P(W_{n,k,r}(u) < m). \end{aligned} \tag{3.1}$$

As a consequence, the cumulative distribution function of the maximum of the moving window order statistics $Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}$ is given by the probability mass function of $W_{n,k,r}(u)$ at zero, namely

$$P(\max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}) \leq u) = P(W_{n,k,r}(u) = 0). \tag{3.2}$$

Note that the maximum appearing above is over a set of dependent variables, since $Y_{r;k}^{(i)}$ involve order statistics over overlapping sets of variables.

Note also that the notation used here for the order statistics is slightly different from the one used in traditional order statistics books (see e.g. Arnold and Balakrishnan (1989) and David and Nagaraja (2003)); by our notation, the r-th order statistic refers to the r-th *largest* observation instead of the r-th *smallest*.

Let us next consider the special case where the random variables Y_i follow the uniform distribution on $(0, 1)$ and set $u = 1-p$ where $0 < p < 1$. Then $X_i(u)$

becomes a Bernoulli variable, say X_i, with success probability p and $S_{n,k}(u)$ reduces to the binary discrete scan statistic $S_{n,k}$ mentioned in the Introduction.

In this chapter we shall make use of the term *threshold-based scan statistic* to refer to the general model and *binary scan statistic* to refer to the special case where X_i are i.i.d. Bernoulli variables with common success probabilities p. In the latter case, the following notation will be practiced:

$$\begin{aligned} S_k^{(i)} &= \sum_{j=i}^{i+k-1} X_j, \quad i = 1, 2, \ldots, n-k+1, \\ S_{n,k} &= \max_{1 \le i \le n-k+1} S_k^{(i)}, \\ W_n &= W_{n,k,r} = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_k^{(i)}). \end{aligned}$$

The notation $\sim, O(\cdot)$ will assume their usual meaning, i.e.

$$\begin{aligned} f(t) \sim g(t) \text{ as } t \to t_0 \text{ if } \lim_{t \to t_0} \frac{f(t)}{g(t)} = 1, \\ f(t) = O(g(t)) \text{ if } \frac{f(t)}{g(t)} \text{ is bounded.} \end{aligned}$$

3.3 The Binary Scan Statistic

In this section we present several results for the binary discrete scan statistic $S_{n,k}$ and the binary multiple scan statistic $W_{n,k,r}$. In Subsection 3.3.1 we review several bounds and approximations that have been suggested for deriving accurate estimates of the distributions of the binary scan statistics when n is large. Most of them use an appropriate Poisson or compound Poisson approximation along with efficient error bounds for the distance between the exact and approximating distribution. Although another class of approximations which exploits a product-type formula is available in the literature, it will not be covered here, since usually these approximations do not offer error estimates and therefore no asymptotic (convergence) results can be achieved. The interested reader may refer to the monograph by Glaz, Naus and Wallenstein (2001) for a detailed presentation of product-type approximations.

A very popular method for establishing Poisson approximations for sums of (potentially) dependent Bernoulli random variables is the celebrated Chen–Stein method; see Arratia, Goldstein and Gordon (1990) or the monograph by Barbour, Holst and Janson (1992). This method can be used to compute an

upper bound for the total variation distance between the law $\mathcal{L}(W_{n,k,r})$ of the random variable $W_{n,k,r}$ and a Poisson distribution $Po(\lambda)$, i.e.

$$d_{TV}(\mathcal{L}(W_{n,k,r}), Po(\lambda)) = \sup_A |P(W_{n,k,r} \in A) - P(Z_\lambda \in A)|,$$

where Z_λ is a random variable following the Poisson distribution with $E(Z_\lambda) = \lambda > 0$, and A is any subset of non-negative integers.

Since the scan enumerating variable $W_{n,k,r}$ takes into account clusters of consecutive trials with a high concentration of successes and these events tend to occur in clumps, the compound Poisson distribution is an even more natural choice than the Poisson.

We recall that the term compound Poisson distribution with parameter λ and compounding distribution G (notation: $CP(\lambda, G)$) refers to the distribution of the random sum $U = \sum_{i=1}^{N} Z_i$, where N is a Poisson random variable with $\lambda = E(N)$, and Z_i are i.i.d. variables, independent of N, with cumulative distribution function G.

The probability generating function of U takes on the form

$$P_U(t) = E(t^U) = e^{-\lambda(1-E(t^Z))} = e^{-\lambda(1-P_Z(t))}. \tag{3.3}$$

where $P_Z(t)$ is the probability generating function of Z_i, $i = 1, 2, \ldots$. As a consequence, one could evaluate the probability mass function of U by considering the power series expansion of $P_U(t)$ (provided that an explicit expression is available for the probability generating function $P_Z(t)$ of the compounding distribution).

3.3.1 Bounds and approximations

Denote by $b(x; l, p)$ and $B(x; l, p)$ the probability mass function and cumulative distribution function of a binomial random variable X, i.e.

$$b(x; l, p) = P(X = x) = \binom{l}{x} p^x (1-p)^{l-x}, \quad x = 0, 1, \ldots, n,$$

$$B(x; l, p) = P(X \le x) = \sum_{j=0}^{\lfloor x \rfloor} b(j; l, p), \quad x \in \Re,$$

where the symbol $\lfloor x \rfloor$ indicates the integer part of x. In the following sections we shall make use of the quantities

$$f(s; k, p) = P(S_k^{(1)} < s, S_k^{(2)} < s, \ldots, S_k^{(k)} < s, S_k^{(k+1)} \ge s)$$
$$G(s; k, p) = P(S_k^{(1)} < s, S_k^{(2)} < s, \ldots, S_k^{(k+1)} < s),$$

which can be expressed in terms of $b(x;l,p)$ and $B(x;l,p)$ as follows (see Glaz and Naus (1991)):

$$f(s;k,p) = \frac{p}{s} b(s-1;k-1,p)[s(1-p)b(s-1;k-1,p) + (s-kp)B(s-2;k-1,p)], \quad (3.4)$$

$$G(s;k,p) = B(s-1;k,p)^2 - b(s;k,p)[(s-1)B(s-2;k,p) - kpB(s-3;k-1,p)],$$

with $1 \le s \le k$ (if $s > k$ or $s < 0$ then we set $f(s;k,p) = 0$).

We shall start with a result pertaining to the binary scan statistic $S_{n,k}$ which was established by Arratia, Gordon and Waterman (1990) with the aid of the Chen–Stein method. Let us introduce first the random variables

$$C_i = \begin{cases} 1, \text{ if } \sum_{j=i}^{i+k-1} X_j = r \\ 0, \text{ else} \end{cases}, i = 1, 2, \ldots, n-k+1$$

(convention: $C_i = 0$ for $i \le 0$) and then define an auxiliary variable W by summing up the quantities

$$D_i = C_i \prod_{j=1}^{k} (1 - C_{i-j}), \quad i = 1, 2, \ldots, n-k+1$$

i.e.

$$W = \sum_{i=1}^{n-k+1} C_i \prod_{j=1}^{k} (1 - C_{i-j}).$$

Arratia, Gordon and Waterman (1990) proved the following interesting result.

Theorem 3.3.1 *If $p < r/k < 1$ then*

$$\left| P(S_{n,k} < r) - e^{-E(W)} \right| \le 7kb(r;k,p) + (1 - B(r;k,p)).$$

That is, $P(S_{n,k} < r)$ is bounded below and above by the quantities $e^{-E(W)} \pm UB$, where UB denotes the quantity on the right-hand side (RHS) of the last inequality, and $E(W)$ equals $(n-k+1)E(D_i)$.

The same authors proved that $E(W)$ can be bounded as follows:

$$\frac{r}{k} - p \le \frac{E(W)}{(n-k+1)b(r;k,p)} \le (\frac{r}{k} - p) + 2(1 - \frac{r}{k})(1 - B(r;k,p)) \quad (3.5)$$

while an alternative upper bound is given by

$$\frac{E(W)}{(n-k+1)b(r;k,p)} \leq (\frac{r}{k}-p) + 2(1-\frac{r}{k})e^{-kH(r/k,\,p)},$$

where $H(\theta,p)$ denotes the Kullback–Leibler distance (or relative entropy)

$$\begin{aligned} H(\theta,p) &= \theta \ln\left(\frac{\theta}{p}\right) + (1-\theta)\ln\left(\frac{1-\theta}{1-p}\right) \\ &= \ln\frac{\theta^\theta(1-\theta)^{1-\theta}}{p^\theta(1-p)^{1-\theta}}, \qquad 0<p<\theta<1. \end{aligned} \tag{3.6}$$

For large values of k (with r/k kept fixed) the second summand on the RHS of (3.5) becomes negligible; therefore, the next simple approximation formula can be used:

$$E(W) \approx (\frac{r}{k}-p)(n-k+1)b(r;k,p).$$

Dembo and Karlin (1992) used the Chen–Stein method to establish upper bounds for the total variation distance between a scan process generated by a sequence of i.i.d. positive random variables (not necessarily binary) and a Poisson distribution. In the special case of i.i.d. binary variables, the following simpler results ensue.

Theorem 3.3.2 *If*

$$\lambda = (n-k+1)B(r-1;k,p), \mu = (n-k+1)(1-B(r-1;k,p)),$$

then

a. $d_{TV}(\mathcal{L}(W_{n,k,r}), Po(\mu)) \leq (1-e^{-\mu})[(2k-1)(1-B(r-1;k,p)) + 2\sum_{i=1}^{k-1} P(S_k^{(i+1)} \geq r | S_k^{(1)} \geq r)]$

b. $d_{TV}(\mathcal{L}(n-k+1-W_{n,k,r}), Po(\lambda)) \leq (1-e^{-\lambda})[(2k-1)B(r-1;k,p) + 2\sum_{i=1}^{k-1} B(r-1;i,p)].$

The conditional probabilities $P(S_k^{(i+1)} \geq r | S_k^{(1)} \geq r)$ appearing in the first upper bound can be easily expressed in terms of binomial probabilities as follows:

$$P(S_k^{(i+1)} \geq r | S_k^{(1)} \geq r) = \frac{1}{1-B(r-1;k,p)} \cdot \sum_{s=0}^{k-i} (1-B(r-s-1;i,p))^2 b(k-i;s,p)$$

with $B(x;l,p) = b(x;l,p) = 0$, for $x < 0$. The results of Theorem 3.3.2 can be exploited to bound the probability mass function or the cumulative distribution function of $W_{n,k,r}$ by an interval centered on the probability mass function or cumulative distribution function of a Poisson distribution, respectively, whose length is two times the error bounds provided above.

As already mentioned, due to the fact that windows of high concentration of successes tent to occur in clumps, one should expect that the Poisson approximations described in Theorem 3.3.2 would naturally provide poor results. As a matter of fact, the upper bound in (a) converges to zero only for $r = k$ and $p \to 0$ while the upper bound in (b) converges only for $r = 1$ and $p \to 1$.

Motivated by a sequence matching problem where the need of an accurate approximation of the discrete scan statistic was recognized, Goldstein and Waterman (1992) used the declumping variables

$$E_i = I_{[r,\infty)}(S_k^{(i)}) \prod_{j=1}^{\min(s,i-1)} (1 - I_{[r,\infty)}(S_k^{(i-j)})), \qquad i = 1, 2, \ldots, n-k+1$$

(where s is a fixed positive integer)[1] to establish total variation bounds for the compound Poisson approximation of $W_{n,k,r}$. The binary variables E_i indicate the occurrence of a clump where the starting point is the variable X_i, while the distribution of the number C of occurrences of the event $S_k^{(j)} \geq r$ within the clump is given by

$$P(C = c) = P(\sum_{j=i}^{\beta} I_{[r,\infty)}(S_k^{(j)}) = c | E_i = 1), \qquad c = 1, 2, \ldots,$$

where

$$\beta = \min(\gamma \geq i : I_{[r,\infty)}(S_k^{(\gamma)}) = 1, I_{[r,\infty)}(S_k^{(\gamma+1)}) = 0, \ldots, I_{[r,\infty)}(S_k^{(\gamma+s)}) = 0).$$

Considering $s = k$, the following result can be established with the aid of the Chen–Stein method (cf. Goldstein and Waterman (1992)).

Theorem 3.3.3 *The total variation distance between $W_{n,k,r}$ and a compound Poisson distribution $CP(\lambda, G)$ with $\lambda = (n-k+1)(1 - B(r-1;k,p))/E(C)$ and compounding distribution $G(x) = P(C \leq x)$ is bounded above as follows:*

$$d_{TV}(\mathcal{L}(W_{n,k,r}), CP(\lambda, G)) \leq 6\lambda^2(1 + E(C))\frac{k}{n-k} + 2\lambda P(C > k).$$

Note that the above expression for λ was deduced upon ignoring the boundary effects. The exact distribution of C is quite intricate; to overcome this, Goldstein

[1] Products of the form $\prod_{i=i_1}^{i_2} f(i)$ with $i_1 > i_2$ are conventionally set equal to 1.

and Waterman (1992) have suggested an alternative simpler but less accurate approximation which will not be presented here. If one is interested only in the distribution of $S_{n,k}$, and not the whole distribution of $W_{n,k,r}$, the following simple lower and upper bounds for $E(C)$ (given that $s = k$) may be useful:

$$\frac{r}{k} - p \leq \frac{1}{E(C)} \leq (\frac{r}{k} - p) + 2(1 - \frac{r}{k})e^{-kH(r/k,\, p)}.$$

A slightly different approach for deducing an upper bound for the distance between the distribution of $W_{n,k,r}$ and a compound Poisson distribution was practiced by Boutsikas and Koutras (2002). They used the "truncated" declumping variables

$$E_i' = (1 - I_{[r,\infty)}(S_k^{(i-1)})) \sum_{j=1}^{k} \prod_{l=i}^{i+j-1} (I_{[r,\infty)}(S_k^{(l)})), \qquad i = 1, 2, \ldots, n-k+1$$

and, exploiting a general result that they also developed [Boutsikas and Koutras (2001)], they proved the following theorem. For the presentation of the results up to the end of this section, the evaluation of the $S_k^{(i)}$'s is carried out by assuming that the sequence of trials $X_i, i = 1, 2, \ldots, n$ is extended for $i < 1$ and $i > n$. Note that the distance involved in this result is no longer the total variation distance, but the Kolmogorov distance, which is defined as follows (U, V are any random variables with cumulative distributions functions F_U and F_V, respectively):

$$d_K(\mathcal{L}(U), \mathcal{L}(V)) = d_K(F_U, F_V) = \sup_{-\infty < x < \infty} |F_U(x) - F_V(x)|.$$

Theorem 3.3.4 *Let* $\lambda = (n-k+1)P(E_1' > 0)$ *and* $G(x) = P(E_1' \leq x | E_1' > 0)$ $x = 0, 1, \ldots, k$. *Then*

$$\begin{aligned} d_K(\mathcal{L}(W_{n,k,r}), CP(\lambda, G)) \leq\ & (\lambda+1) \sum_{i=r}^{k} \binom{k}{i} p^i (1-p)^{k-i} \\ & + (\lambda(3k-1) + k - 1) \binom{k-1}{r-1} p^r (1-p)^{k-r+1} \\ & + (n-k) \sum_{b=2}^{k-1} \sum_{\substack{i= \\ \max\{0, r-k+b-1\}}}^{\min\{b-2, r-2\}} \binom{k-b}{r-i-1} \binom{b-2}{i} \\ & \binom{k-b}{r-i-2} p^{2r-i-1} (1-p)^{2k-b-2r+i+3}. \end{aligned}$$

The advantage of this result over the one stated in Theorem 3.3.3 is that both the upper bound for the distance and the parameters λ, G of the compound

Poisson distribution admit explicit expressions which are computationally tractable. It can be easily verified that $P(E_1' > 0) = \binom{k-1}{r-1} p^r q^{k-r+1}$ and therefore

$$\lambda = (n-k+1)P(E_1' > 0) = (n-k+1)\binom{k-1}{r-1} p^r q^{k-r+1}.$$

Moreover, the cumulative distribution function $G(x)$ of the compounding distribution has been computed by Boutsikas and Koutras (2002) as

$$\begin{aligned} G(x) = 1 - \sum_{j=\max\{0,r-x-1\}}^{\min\{k-x-1,r-1\}} & \left(\frac{\binom{x}{x-r+j+1}\binom{k-x-1}{j}}{\binom{k-1}{r-1}} \right) \\ & \cdot \left(\binom{x}{x-r+j+1} p^{r-j-1}(1-p)^{x-r+j+1} + \right. \\ & \left. + \left(1 - \frac{(x+1)(1-p)}{x-r+j+2}\right) \left(\sum_{i=0}^{x-r+j} \binom{x}{i} (1-p)^i p^{x-i-1} \right) \right) \end{aligned}$$

for $x = 1, 2, \ldots, k-1, G(0) = 0$, $G(k) = 1$. The upper bound described in Theorem 3.3.4 is of order $O(p)$ (for $r < k$); therefore it will produce reasonable lower and upper bounds for the cumulative distribution function of $W_{n,k,r}$ when $p \to 0$. Nonetheless, if p is fixed, the quality of the bounds will not be as good, and it is conceivable that no asymptotic results could be established with their use for fixed p and $n, k \to \infty$.

In order to cover the last case, the following family of declumping variables may be used:

$$E_i'' = \left(\prod_{j=i-k}^{i-1} (1 - I_{[r,\infty)}(S_k^{(j)})) \right) I_{[r,\infty)}(S_k^{(i)}) \left(\sum_{l=i}^{i+k} I_{[r,\infty)}(S_k^{(l)}) \right), \quad i = 1, \ldots.$$

The last bracket enumerates the number of scanning windows of length k that begin at positions $i, i+1, \ldots, i+k$ and contain at least r successes each. On the other hand, the first bracket guarantees that in the previous k positions $i-k, i-k+1, \ldots, i-1$ all scanning windows of length k contain fewer than r successes. As a matter of fact, it is the inclusion of this extra term that makes the construction of sharp bounds feasible. Exploiting the new family of declumping variables $E_i'', i = 1, 2, \ldots$, Boutsikas and Koutras (2006) arrived at the next result.

Theorem 3.3.5 *Let* $\lambda = (n-k+1)P(E_1'' > 0) = (n-k+1)f(r;k,p)$, *and*

$$G(x) = P(E_1'' \le x | E_1'' > 0)$$

$$= P\left(\sum_{l=k+1}^{2k+1} I_{[r,\infty)}(S_k^{(l)}) \le x \,\middle|\, I_{[r,\infty)}(S_k^{(l)}) = 0, l = 1, \ldots, k, I_{[r,\infty)}(S_k^{(k+1)}) = 1 \right),$$

where $x = 0, 1, \ldots, k$.

Then

$$d_K(\mathcal{L}(W_{n,k,r}), CP(\lambda, G))$$

$$\le (2k-1)\lambda p(1-p)b(r-1;k-1,p) + 3\lambda k f(r;k,p) + (\lambda+2)(1-G(r;k,p)),$$

where $f(r;k,p), G(r;k,p)$ *are given in (3.4).*

Although the evaluation of the compounding cumulative distribution function $G(x)$ is not easily accomplished, the above result is quite appealing for establishing the asymptotic behavior of $W_{n,k,r}$ for fixed p and $n, k \to \infty$. Moreover, if the interest is focused on $S_{n,k}$ and not the whole distribution function of $W_{n,k,r}$, then the cumulative distribution function of $S_{n,k}$, i.e. $P(S_{n,k} < r) = P(W_{n,k,r} = 0)$, can be effortlessly assessed by the aid of the following simpler result.

Corollary 3.3.1 *The cumulative distribution function of* $S_{n,k}$ *can be approximated by* $e^{-\lambda}$, $\lambda = (n-k+1)f(r;k,p)$ *with the error of approximation bounded above as follows:*

$$|P(S_{n,k} < r) - e^{-\lambda}| \le (2k-1)\lambda p(1-p)b(r-1;k-1,p) + 3\lambda k f(r;k,p) + (\lambda+2)(1-G(r;k,p)).$$

3.3.2 Asymptotic results

We are now presenting a number of asymptotic results pertaining to the Poisson and compound Poisson convergence of the discrete scan statistics $S_{n,k}$ and $W_{n,k,r}$. Although we are not going to present the technical details of the proofs of these results, alert readers may easily extract them by the aid of the bounds (on the total variation or Kolmogorov distances) described in the previous subsection.

In the light of Theorem 3.3.1 and the discussion following it, we may state the next result, which was provided by Arratia, Gordon and Waterman (1990).

Corollary 3.3.2 *If* n, k, r *are positive integers with* $p < r/k \ne 1$ *and* $\lambda = (n-k+1)(\frac{r}{k} - p)b(r;k,p)$ *then*

$$P(S_{n,k} < r) \text{ can be approximated by } e^{-\lambda}$$

with the approximation error of order $O(\frac{\ln n}{n})$.

In the next corollary, which can be easily proved by the aid of Theorem 3.3.4 (see Boutsikas and Koutras (2002) for details), a compound Poisson convergence for $W_{n,k,r}$ is established.

Corollary 3.3.3 *Assume that k, r are kept fixed, while $n \to \infty, p_n \to 0$ so that $\lambda_n = (n-k+1)\binom{k-1}{r-1}p_n^r(1-p_n)^{k-r+1} \to \lambda \in (0,\infty)$. Then $W_{n,k,r}$ converges weakly to a compound Poisson distribution with parameter λ and compounding distribution*

$$G(x) = \begin{cases} 0, & x \le 0 \\ 1 - \frac{\binom{k-x-1}{r-1}}{\binom{k-1}{r-1}}, & x = 1, 2, \ldots, k-r, \\ 1, & x \ge k-r+1. \end{cases}$$

Under the same assumptions for the cumulative distribution of $S_{n,k}$ we have

$$P(S_{n,k} < r) \approx e^{-\lambda}.$$

The probability mass function $g(x)$ of the compounding distribution $G(x)$ takes on the form

$$g(x) = G(x) - G(x-1) = \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \ldots, k-r+1.$$

It is of interest to note that, in the special case $r = 2 < k$, $g(x)$ reduces to the discrete uniform distribution on the integers $1, 2, \ldots, k-1$. Also, for $k = r$, the compounding distribution becomes degenerate (with all its mass concentrated at 1) and the limiting compound Poisson law $CP(\lambda, G)$ reduces to an ordinary Poisson distribution.

The probability generating function of the compound Poisson distribution described in Corollary 3.3.3 reads (cf. (3.3))

$$P(t) = E(t^{W_{n,k,r}}) = e^{-\lambda(1-E(t^Z))} = e^{-\lambda\left(1-\sum_{x=1}^{k-r+1} t^x g(x)\right)} = \exp\left(-\lambda\left(1 - \sum_{x=1}^{k-r+1} t^x \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}\right)\right)$$

and therefore we may easily compute the probabilities $P(W_{n,k,r} = i)$ considering the power series expansion of $P(t)$ and extracting the coefficients of i-th order term, $i = 0, 1, \ldots$. This can be done numerically (for specific values of the parameters) or analytically for small order terms. For example,

$$\begin{aligned} P(W_{n,k,r} = 0) &= \frac{1}{0!}P(0) = e^{-\lambda}, \\ P(W_{n,k,r} = 1) &= \frac{1}{1!}\left.\frac{dP(t)}{dt}\right|_{t=0} = \lambda g(1)e^{-\lambda} = \lambda\frac{r-1}{k-1}e^{-\lambda} \\ P(W_{n,k,r} = 2) &= \frac{1}{2!}\left.\frac{d^2P(t)}{dt^2}\right|_{t=0} = \lambda\frac{(r-1)(k-r)}{(k-1)(k-2)}e^{-\lambda} + \frac{\lambda^2}{2!}\left(\frac{r-1}{k-1}\right)^2 e^{-\lambda} \end{aligned} \tag{3.7}$$

etc. Alternatively, one may resort to the following recursive scheme (see Bowers *et al.* (1997)):

$$P(W_{n,k,r} = 0) = e^{-\lambda},$$

$$P(W_{n,k,r} = i) = \frac{\lambda k}{ri}\binom{k}{r}^{-1} \sum_{j=1}^{\min(k-r+1,i)} j\binom{k-j-1}{r-2} P(W_{n,k,r} = i-j), \qquad i = 1, 2, \ldots.$$

For $r < k$, the rate of convergence ascertained by Corollary 3.3.3 for the approximation $P(S_{n,k} < r) \approx e^{-\lambda}$ is of order $O(p)$. As a consequence, it is not applicable for fixed p and $n, k \to \infty$. The last case is covered by the next result which can be inferred from Theorem 3.3.5; the interested reader is referred to Boutsikas and Koutras (2006) for the technical details.

We shall use the symbol $h(\theta, p)$ to denote the derivative of the Kullback–Leibler distance $H(\theta, p)$ (3.6) with respect to θ, i.e.

$$h(\theta, p) = \frac{d}{d\theta} H(\theta, p) = \ln \frac{\theta(1-p)}{p(1-\theta)}, \qquad 0 < p < \theta < 1.$$

Corollary 3.3.4 *Let p be fixed, $\theta \in (p, 1)$, and k_n, r_n be two sequences satisfying the condition*

$$\lim_{n\to\infty} \frac{r_n - \theta k_n}{\sqrt{k_n}} = 0.$$

If $\rho_n = r_n - \theta k_n$ and the sequence

$$l_n = n\frac{(\theta - p)e^{-k_n H(\theta,p) - \rho_n h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k_n}}, \qquad n = 1, 2, \ldots$$

is bounded from above, then

$$P(S_{n,k} < r) \sim e^{-l_n}$$

with the rate of convergence of order $O(\frac{\rho_n^2+1}{k_n})$.

Practically speaking, the last corollary states that, for large values of n, k, r and $p < r/k \neq 1$, the cumulative distribution function of $S_{n,k}$ can be approximated by the aid of the formula (replace r_n, k_n, θ by $r, k, r/k$, respectively, in the formula of l_n):

$$P(S_{n,k} < r) \sim \exp\left(-n\frac{(r - kp)e^{-kH(r/k,p) - \rho h(r/k,p)}}{\sqrt{2\pi rk(k-r)}}\right).$$

3.3.3 Extreme value results

Extreme value results pertaining to moving sums of i.i.d. random variables (not necessarily binary) have been the subject of continuing interest for many decades. They are usually referred to as Erdős-Rényi laws and deal with the asymptotic distribution of

$$U_n = \max_{1\le i\le n-k+1} \sum_{j=i}^{i+k-1} Z_j,$$

where $Z_1, Z_2, \ldots$ is a sequence of i.i.d. random variables. The classical Erdős–Rényi (1970) theorem establishes the almost sure convergence to 1 of the sequence of random variables $U_n/(ak_n)$ for a large class of distributions for Z_i ($k = k_n = \lfloor c\ln(n)\rfloor$ for some positive constant c and $a > 0$ is a number depending on c and the distribution of Z_i). An extreme value theorem for U_n obtained by Deheuvels and Devroye (1987) states that, under the assumption that Z_i follow a non-lattice distribution with zero mean,

$$\lim_{n\to\infty} P(\frac{U_n - b_n}{a_n} \le x) = \Lambda(x),$$

where $\Lambda(x) = \exp(-e^{-x})$, $x \in \Re$ is the cumulative distribution of the Gumbel distribution, and $a_n, b_n \in \Re$ are appropriate sequences of normalizing constants.

We shall now present two extreme value results for the discrete scan statistic $S_{n,k}$; the fact that the original sequence of variables is binary (and therefore lattice) makes it possible to express the normalizing constants by explicit formulae.

The following theorem is a slight restatement of a result established by Arratia, Gordon and Waterman (1990).

Theorem 3.3.6 *Let $k > -(\ln p)^{-1}\ln n$, denote by $\theta = \theta(n,k,p) \in (p,1)$ the unique solution* [2] *of the equation*

$$H(\theta, p) = \frac{\ln n}{k},$$

and define the normalizing constants b_n as follows:

$$b_n = \theta\frac{\ln n}{H(\theta,p)} - \frac{1}{2h(\theta,p)}\ln(\ln n) - \frac{1}{2h(\theta,p)}\ln(\frac{2\pi\theta(1-\theta)}{H(\theta,p)}) + \frac{\ln(\theta-p)}{h(\theta,p)}.$$

Then for each $\varepsilon > 0$ such that $1+\varepsilon \le -(\ln n)^{-1}k\ln p \le 1/\varepsilon$ the following result holds true, uniformly for $n, k \to \infty$:

$$\sup_x |P(S_{n,k} - b_n < x) - \Lambda(h(\theta,p)x)| \to 0.$$

The supremum is evaluated over all $x \in \Re$ such that $x + b_n$ is a positive integer.

[2] Since $\frac{d}{d\theta}H(\theta,p) = h(\theta,p) > 0$, the quantity $H(\theta,p)$ varies monotonically from 0 to $-\ln p$; therefore, the equation $H(\theta,p) = c$ admits a unique solution $\theta \in (p,1)$. for $0 < c < -\ln p$.

Exploiting Corollary 3.3.4 it is not difficult to gain the following extreme value theorem, which also establishes the convergence to the Gumbel distribution of an appropriate normalized version of $S_{n,k}$ (cf. Boutsikas and Koutras (2006)).

Theorem 3.3.7 *For fixed $p \in (0,1)$, let θ be a number in the interval $(p,1)$ and define*

$$k_n = \lfloor \ln n / H(\theta, p) \rfloor ,$$
$$b_n = k_n \theta + \frac{1}{h(\theta,p)} \ln \frac{n(\theta-p)e^{-k_n H(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k_n}},$$
$$\epsilon_n(y) = \left(b_n + \frac{y}{h(\theta,p)} \right) - \left\lfloor b_n + \frac{y}{h(\theta,p)} \right\rfloor .$$

Then

$$\lim_{n\to\infty} \left[P\left(\frac{S_{n,k} - b_n}{1/h(\theta,p)} < y \right) - \Lambda(y - \epsilon_n(y) h(\theta,p)) \right] = 0$$

with convergence rate of order $O(\frac{(\ln k_n)^2}{k_n})$.

3.4 Scan Statistic Exceedances

In this section we shall present some new results for the multiple scan statistic under the threshold exceedance framework. More specifically, we consider a sequence $Y_1, Y_2, \ldots, Y_n$ of continuous i.i.d. random variables with cumulative distribution function F, and a threshold $u = u_n$ which varies with n. Exploiting Corollary 3.3.3, we establish first a compound Poisson convergence theorem for $W_{n,k,r}(u_{a_n})$, where a_n is an appropriate sequence of positive real numbers. Then we use this result to develop asymptotic results for $W_{n,k,r}(u_{a_n})$ or equivalently for the moving order statistic $Y_{m:r:k}(n)$ introduced in Section 3.2 under the assumption that the distribution of Y_i belongs to the domain of attraction of one of the three classical extreme type distributions (Weibull, Fréchet, Gumbel). Finally, application of the general results for typical continuous distributions along with illustrative graphs exhibiting the quality of convergence are provided.

3.4.1 Compound Poisson approximation for $W_{n,k,r}(u)$

Let us assume that the threshold u_n varies with n so that the event $Y_i > u_n$ becomes a rare event, i.e. $P(Y_i > u_n) \to 0$ as $n \to \infty$. A standard condition that yields non-degenerate results is the following:

$$\lim_{n\to\infty} nP(Y_i > u_n) = \lim_{n\to\infty} n\overline{F}(u_n) = \tau \in (0,\infty).$$

Denoting by a_n the sequence $a_n = n^{1/r}$, $n = 1, 2, \ldots$, it is clear that

$$\lim_{n\to\infty} n\overline{F}(u_{a_n})^r = \lim_{n\to\infty} (a_n\overline{F}(u_{a_n}))^r = \tau^r.$$

Viewing $X_i(u_{a_n}) = I_{(u_{a_n},\infty)}(Y_i)$, $i = 1, 2, \ldots$ as a sequence of binary trials with success probabilities $p_n = \overline{F}(u_{a_n})$, we may apply Corollary 3.3.3 with

$$\lambda_n = (n-k+1)\binom{k-1}{r-1}p_n^r(1-p_n)^{k-r+1} \to \binom{k-1}{r-1}\tau^r > 0$$

to conclude that $W_{n,k,r}(u_{a_n})$ converges to a compound Poisson distribution. Recalling formula (3.1) we arrive at the following interesting result.

Theorem 3.4.1 *If there exists a sequence u_n such that $\lim_{n\to\infty} n\overline{F}(u_n) = \tau > 0$, then the distribution of the number $W_{n,k,r}(u_{a_n})$ of (overlapping) moving windows of length k that contain at least r exceedances of the threshold u_{a_n} with $a_n = n^{1/r}$ (k and r are fixed positive integers) converges to a compound Poisson distribution with parameter $\lambda = \binom{k-1}{r-1}\tau^r$ and compounding distribution with probability mass function*

$$g(x) = \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}, \qquad x = 1, 2, \ldots, k-r+1. \tag{3.8}$$

Moreover, if we denote by f_{CP} the probability mass function of the compound Poisson distribution, we may write

$$\lim_{n\to\infty} P(Y_{m:r:k}(n) \le u_{a_n}) = \lim_{n\to\infty} P(W_{n,k,r}(u_{a_n}) < m) = \sum_{i=0}^{m-1} f_{CP}(i).$$

For the evaluation of the limiting distribution, see the discussion after Corollary 3.3.3.

Although u_{a_n} is typically defined over the set of positive integers n, hereafter we shall assume that such sequences can be extended in $\Re^+$, i.e. $u_x = u(x)$, $x \in \Re^+$. Under this assumption, it makes sense to write u_{a_n}, where a_n is any sequence of (not necessarily integer) numbers, a notation that will be met frequently in the following subsections. It is worth stressing that this technique works in a nice way for all the examples illustrated in Subsection 3.4.3. If the aforementioned extension is not feasible, one may use $u_{\lfloor a_n \rfloor}$ in place of u_{a_n} and the results provided hereafter remain valid as well.

Note that, for the determination of the limiting value $\binom{k-1}{r-1}\tau^r$ of λ_n, use was made of the fact that

$$(1-p_n)^{k-r+1} = \left(1 - \frac{a_n\overline{F}(u_{a_n})}{a_n}\right)^{k-r+1} \sim \left(1 - \frac{\tau}{a_n}\right)^{k-r+1} \underset{n\to\infty}{\to} 1.$$

However, this convergence is quite slow, even for small values of r. For example, if $k = 4, r = 2, \tau = 2$, then $q_{100}^{k-r+1} \approx 0.512, q_{1000}^{k-r+1} \approx 0.822, q_{10000}^{k-r+1} \approx 0.941$. Things get much worse when r is larger. Therefore, although the asymptotic result in Theorem 3.4.1 is valid for $n \to \infty$, the distribution of $W_{r,k,n}(u_{a_n})$ may be approximated more accurately by a compound Poisson distribution $CP(\lambda_n^*, G)$ with parameter

$$\lambda_n^* = \binom{k-1}{r-1}\tau^r \left(1 - \frac{\tau}{a_n}\right)^{k-r+1} = \lambda \left(1 - \frac{\tau}{a_n}\right)^{k-r+1} \tag{3.9}$$

(instead of λ). This remark is exploited in the numeric experimentation carried out in Subsection 3.4.3.

A result similar to the one stated in Theorem 3.4.1 has been given by Dudkiewicz (1998); however his result covers only the special case of moving minima and is applicable for different conditions on the parameters k, r.

In view of Theorem 3.4.1 and taking into account (3.2), we may write that

$$\lim_{n\to\infty} P(\max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}) \le u_{a_n}) = \lim_{n\to\infty} P(W_{n,k,r}(u_{a_n}) = 0) = e^{-\lambda}.$$

Apparently, the above formula displays the asymptotic behavior of the maximum of a family of dependent variables, namely $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$, whose marginal tails can be expressed through the tail of the binomial distribution with parameters k and $\overline{F}(u_{a_n})$, namely

$$\begin{aligned} nP(Y_{r:k}^{(i)} > u_{a_n}) &= n\sum_{i=r}^{k} \binom{k}{i} \overline{F}(u_{a_n})^i F(u_{a_n})^{k-i} \\ &\sim \binom{k}{r} (a_n \overline{F}(u_{a_n}))^r \to \binom{k}{r} \tau^r. \end{aligned}$$

If we assumed for the moment that $Y_{r:k}^{(i)}$ were independent, the limiting behavior of their maximum would reduce to

$$\lim_{n\to\infty} P(\max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}) \le u_{a_n}) = \lim_{n\to\infty} \left(1 - \frac{nP(Y_{r:k}^{(i)} > u_{a_n})}{n}\right)^n = e^{-\lambda_{ind}},$$

where $\lambda_{ind} = \binom{k}{r}\tau^r$. The ratio

$$\frac{\lambda}{\lambda_{ind}} = \frac{\binom{k-1}{r-1}\tau^r}{\binom{k}{r}\tau^r} = \frac{r}{k}$$

characterizes the extremal dependence between $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$. This quantity has been termed the *extremal index* by Embrechts *et al.* (1997) and, as stated therein, it provides information on the local dependence of the variables under study. According to the last result, the extremal index for the sequence $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$ decreases (and therefore the local dependence becomes stronger) as k increases or r decreases, a fact that can be easily interpreted intuitively.

3.4.2 Convergence of threshold-based scan statistics under maximum domain of attraction assumptions

The probabilistic extreme value theory focuses on the stochastic behavior of the maximum $M_n = \max(Z_1, Z_2, \ldots, Z_n)$ (and the minimum) of sequences of i.i.d. random variables $Z_1, Z_2, \ldots, Z_n$. The asymptotic properties of extremes (maxima or minima), intermediate order statistics and exceedances over (or below) prespecified thresholds are determined by the upper and lower tails of the underlying distribution.

Although extreme value theory seems to have originated mainly from the needs of astronomers in accepting or rejecting outlying observations, after its significant theoretical developments during 1920–1950, a substantial number of articles appeared dealing with practical applications of extreme value statistics in the stochastic analysis of meteorological phenomena (rainfalls, floods), strengths of materials, seismic activity, insurance and actuarial models, radioactive emissions etc. For a comprehensive bibliography of literature on extreme value distributions and their applications, the interested reader is referred to the monographs by Kotz and Nadarajah (2000), Reiss and Thomas (1997), Coles (2001) and Embrechts *et al.* (1997).

In theory, the distribution of M_n can be derived exactly by the formula $P(M_n \leq x) = F(x)^n$; however this is not immediately helpful in practice, since the cumulative distribution function F of Z_i is usually unknown. A typical approach to overcome this is to look for approximate (asymptotic) families of models for F^n which can be estimated on the basis of extreme data only. This is analogous to the classical practice of approximating the distribution of sample means by the normal distribution as justified by the central limit theorem.

Since $\lim_{n\to\infty} P(M_n \leq x) = \lim_{n\to\infty} F(x)^n = 0$ for all x such that $F(x) < 1$, the asymptotic distribution of M_n does not provide any valuable information. However, it is conceivable that more insight into the magnitude of maxima would be given by the centered and normalized maxima (a fact that parallels the normalization used in the central limit theorem).

The fundamental Fisher–Tippett theorem states that, if there exist constants $c_n > 0$ and $d_n \in \Re$ such that

$$\lim_{n\to\infty} P\left(\frac{M_n - d_n}{c_n} \leq x\right) = H(x) \tag{3.10}$$

where H is a non-degenerate distribution function, then H belongs to the location-scale family of one of the following three distributions:

$$\textit{Fréchet}: \qquad \Phi_a(x) = \begin{cases} 0, & x \le 0 \\ exp(-x^{-a}), & x > 0 \end{cases}, \qquad a > 0$$

$$(\textit{Reversed})\ \textit{Weibull}: \quad \Psi_a(x) = \begin{cases} exp(-(-x)^a), & x \le 0 \\ 1, & x > 0 \end{cases}, \qquad a > 0$$

$$\textit{Gumbel}: \qquad \Lambda(x) = exp(-e^{-x}), \quad x \in \Re,$$

i.e. $H(x) = \Phi_a((x-\mu)/\sigma)$ or $H(x) = \Psi_a((x-\mu)/\sigma)$, or $H(x) = \Lambda((x-\mu)/\sigma)$, where $\mu \in \Re$ is a location parameter and $\sigma > 0$ a scale parameter. These families have been termed extreme value distributions, while the respective sequences c_n, d_n are called *norming constants*.

If (3.10) holds true for a cumulative distribution function F, then we shall say that $\overline{F}$ belongs to the maximum domain of attraction (MDA) of H (notation: $\overline{F} \in MDA(H)$). The next theorem characterizes the family of distributions that belong to an MDA (the reader may consult any of the monographs provided earlier for the proof).

Theorem 3.4.2 *The cumulative distribution function F belongs to the MDA of the extreme value distribution H with norming constants $c_n > 0, d_n \in \Re$ if and only if*

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\ln H(x), \qquad x \in \Re.$$

When $H(x) = 0$ the limit is interpreted as $+\infty$.

We shall now establish some non-degenerate convergence results for the threshold-based multiple scan statistic $W_{n,k,r}$, under the assumption that the cumulative distribution function F of Y_i belongs to the MDA of Φ_a, Ψ_a or Λ.

Theorem 3.4.3 *If $\overline{F} \in MDA(H)$ with norming constants $c_n > 0, d_n \in \Re$, then*

$$\lim_{n\to\infty} P\left(\frac{Y_{m:r:k}(n) - d_{a_n}}{c_{a_n}} \le x\right) = \lim_{n\to\infty} P(W_{n,k,r}(c_{a_n}x + d_{a_n}) < m)$$
$$= \sum_{i=0}^{m-1} f_{CP}(x; i),$$

where $a_n = n^{1/r}$ and $f_{CP}(x;\cdot)$ is the probability mass function of a compound Poisson distribution with parameters

$$\lambda(x) = \binom{k-1}{r-1}(-\ln H(x))^r \tag{3.11}$$

and compounding probability mass function given by (3.8).

PROOF. Since $\overline{F} \in MDA(H)$, we conclude by Theorem 3.4.2 that

$$\lim_{n\to\infty} n\overline{F}(c_n x + d_n) = -\ln H(x), \quad x \in \Re.$$

Next applying Theorem 3.4.1 for $u_n = c_n x + d_n$ and $\tau = -\ln H(x)$, we deduce that $W_{n,k,r}(u_{a_n})$ converges weakly to a compound Poisson distribution with parameter

$$\binom{k-1}{r-1}\tau^r = \binom{k-1}{r-1}(-\ln H(x))^r = \lambda(x) \tag{3.12}$$

and compounding probability mass function as described in (3.8). The assertion for the asymptotic distribution of the moving window order statistic $Y_{m:r:k}$ follows immediately by exploiting formula (3.1). ■

According to Theorem 3.4.3, the asymptotic distribution of $W_{n,k,r}(c_{a_n}x + d_{a_n})$ can be approximated for large values of n by a compound Poisson distribution with parameter $\lambda(x)$ given by (3.12). In view of the comments following Theorem 3.4.1, one may improve the quality of the approximation by replacing the parameter $\lambda(x)$ by (cf. (3.9))

$$\begin{aligned}\lambda^*(x) &= \lambda(x)\left(1 - \frac{\tau}{a_n}\right)^{k-r+1} \\ &= \binom{k-1}{r-1}(-\ln H(x))^r \left(1 + \frac{\ln H(x)}{n^{1/r}}\right)^{k-r+1}.\end{aligned}$$

Applying specifically to the three classical extreme value distributions, we mention in brief the following (we shall use the notation x_F for the right end point of F, i.e. $x_F = \sup\{x \in \Re : F(x) < 1\}$).

a. Maximum domain of attraction of Fréchet

If $\overline{F} \in MDA(\Phi_a)$ then $x_F = \infty$ and a possible choice of c_n, d_n is $c_n = F^{-1}(1 - n^{-1})$ and $d_n = 0$ (where F^{-1} denotes the generalized inverse function of F). Since $-\ln H(x) = -\ln \Phi_a(x) = x^{-a}$, the parameters $\lambda(x)$, $\lambda^*(x)$ reduce to

$$\lambda(x) = \binom{k-1}{r-1}x^{-ra}, \lambda^*(x) = \binom{k-1}{r-1}x^{-ra}\left(1 - \frac{x^{-a}}{n^{1/r}}\right)^{k-r+1}, \quad x > 0.$$

Typical members of this class are the classical heavy (right) tailed distributions, e.g. Cauchy, Pareto, log-gamma.

b. Maximum domain of attraction of (reversed) Weibull

If $\overline{F} \in MDA(\Psi_a)$ then x_F is finite and a feasible choice for c_n, d_n is $c_n = x_F - F^{-1}(1 - n^{-1})$ and $d_n = x_F$. The parameters of the approximating compound Poisson distributions are now

$$\lambda(x) = \binom{k-1}{r-1}(-x)^{ra}, \tag{3.13}$$

$$\lambda^*(x) = \binom{k-1}{r-1}(-x)^{ra}\left(1 - \frac{(-x)^a}{n^{1/r}}\right)^{k-r+1}, \quad x \le 0. \tag{3.14}$$

Typical members of this class are the uniform and beta distributions.

c. Maximum domain of attraction of Gumbel
If $\overline{F} \in MDA(\Lambda)$ then F admits a representation of the form (see e.g. Embrechts *et al.* (1997))

$$\overline{F}(x) = c(x)e^{-\int_z^x \frac{g(t)}{a(t)}dt}, z < x < x_F,$$

where z is a real number with $z < x_F$ and c, g are (measurable) functions such that $c(x) \to c_0 > 0$, and $g(x) \to 1$ when $x \uparrow x_F$. The function $a(.)$ is a positive, absolutely continuous function with density a' such that $a'(x) \to 0$ as $x \uparrow x_F$. A valid choice for the function a is

$$a(x) = \int_x^{x_F} \frac{\overline{F}(t)}{\overline{F}(x)}dt,$$

while the norming constants can be defined as $d_n = F^{-1}(1 - n^{-1})$ and $c_n = a(d_n)$. The parameters $\lambda(x)$, $\lambda^*(x)$ of the approximating compound Poisson distributions now read

$$\lambda(x) = \binom{k-1}{r-1}e^{-rx}, \tag{3.15}$$

$$\lambda^*(x) = \binom{k-1}{r-1}e^{-rx}\left(1 - \frac{e^{-x}}{n^{1/r}}\right)^{k-r+1}. \tag{3.16}$$

Typical members of this class are the normal, exponential and gamma distributions.

The special case $m = 1$ of Theorem 3.4.3 reveals the asymptotic distribution of the maximum of a specific order statistic evaluated on moving windows of fixed length k, namely

$$Y_{1:r:k} = \max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}).$$

In applications, Y_i usually represent values of a process measured on a regular time scale, e.g. hourly measurements of sea level, daily claim sizes in a specific portfolio, monthly mean temperatures. Then the statistic $Y_{r:\,k}^{(i)}$ is a location measure for fixed length periods (e.g. $Y_{r:\,k}$ might be the median for the k observations obtained over k consecutive days), and therefore $Y_{1:r:k}$ will correspond to the maximum of the location measures. Such a scenario arises in quite a natural way if the collected data are erased after a certain period (due to storage restrictions) or can be scanned by a scanner of restricted range. Another instance where $Y_{1:r:k}$ might be used is in a similar fashion with the MA (moving average) charts, if we replace the average by a more robust location measure such as the median (or any other order statistic of the moving sample).

The asymptotic distribution of $Y_{1:r:k}$ is described in the following corollary.

Corollary 3.4.1 *If $\overline{F} \in MDA(H)$ with norming constants $c_n > 0, d_n \in \Re$, then*

$$\lim_{n\to\infty} P\left(\frac{\max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}) - d_{a_n}}{c_{a_n}} \le x\right) = e^{-\lambda(x)},$$

where $\lambda(x)$ is given by (3.11).

Proof. We obtain the results immediately by using the obvious formula

$$\lim_{n\to\infty} P\left(\frac{\max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)}) - d_{a_n}}{c_{a_n}} \le x\right)$$
$$= \lim_{n\to\infty} P(W_{n,k,r}(c_{a_n}x + d_{a_n}) < 1) = f_{CP}(x;0) = e^{-\lambda(x)}.$$

■

It is noteworthy that the asymptotic distribution of the maximum of the moving window order statistics (properly centered and normalized) belongs to the same domain of attraction as the original distribution of Y_i; the only parameters that are affected are the location parameter μ and the scale parameter σ. To establish a formal proof of this assertion it suffices to observe the following.

a. Maximum domain of attraction of Fréchet
If $H(x) = \Phi_a(x)$ then

$$\lambda(x) = \binom{k-1}{r-1}(-\ln \Phi_a(x))^{-r} = \binom{k-1}{r-1}(x)^{-ra} = \left(\frac{x-\mu}{\sigma}\right)^{-a'}, x > 0,$$

where

$$a' = ar, \quad \mu = 0, \quad \sigma = \binom{k-1}{r-1}^{1/ra}$$

and therefore

$$e^{-\lambda(x)} = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-a'}} = \Phi_{a'}\left(\frac{x-\mu}{\sigma}\right).$$

b. Maximum domain of attraction of (reversed) Weibull
If $H(x) = \Psi_a(x)$ then

$$\lambda(x) = \binom{k-1}{r-1}(-\ln \Psi_a(x))^{-r} = \binom{k-1}{r-1}(-x)^{ra} = \left(-\frac{x-\mu}{\sigma}\right)^{a'}, x \le 0,$$

where

$$a' = ar, \quad \mu = 0, \quad \sigma = \binom{k-1}{r-1}^{-1/ra}$$

and therefore

$$e^{-\lambda(x)} = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-a'}} = \Psi_{a'}\left(\frac{x-\mu}{\sigma}\right).$$

c. Maximum domain of attraction of Gumbel
If $H(x) = \Lambda(x)$ then

$$\lambda(x) = \binom{k-1}{r-1}(-\ln \Lambda(x))^{-r} = \binom{k-1}{r-1}e^{-rx} = e^{-\frac{x-\mu}{\sigma}}, x \in \Re,$$

where

$$\mu = \frac{1}{r}\ln\binom{k-1}{r-1}, \sigma = \frac{1}{r}$$

and therefore

$$e^{-\lambda(x)} = e^{-e^{-\frac{x-\mu}{\sigma}}} = \Lambda\left(\frac{x-\mu}{\sigma}\right).$$

Finally, if Corollary 3.4.1 is to be used for approximating the distribution of $Y_{1:r:k}$ for small values of n, the quality of approximation will be improved by using $\lambda^*(x)$ instead of $\lambda(x)$.

3.4.3 Examples

In order to exemplify further the usefulness of extreme value theory as presented in the previous sections we consider some typical continuous distributions and illustrate the effectiveness of the approximations established by Theorem 3.4.3 and Corollary 3.4.1.

a. Pareto distribution
Let us assume that $Y_1, Y_2, \ldots$ follow a typical Pareto distribution with cumulative distribution function

$$F(x) = 1 - \left(\frac{c}{x}\right)^a, x \geq c,$$

where a and c are two positive parameters. This is perhaps the most popular heavy-tailed distribution with a lot of applications in socioeconomic and insurance/actuarial models, see e.g. Johnson *et al.* (1994) or the excellent text by Arnold (1985).

In this case, letting $c_n = F^{-1}(1 - n^{-1}) = cn^{1/a}, d_n = 0$, we get

$$n\overline{F}(c_n x + d_n) = n\left(\frac{c}{cn^{1/a}x}\right)^a = x^{-a}, x > 0,$$

which ascertains that $\overline{F} \in MDA(\Phi_a)$. Hence, by Theorem 3.4.3

$$\lim_{n\to\infty} P(\frac{Y_{m:r:k}(n)}{cn^{1/ra}} \leq x) = \lim_{n\to\infty} P(W_{n,k,r}(cn^{1/ra}x) < m)$$
$$= \sum_{i=0}^{m-1} f_{CP}(x;i), x > 0,$$

where $f_{CP}(x;\cdot)$ is the probability mass function of a compound Poisson distribution with parameter

$$\lambda(x) = \binom{k-1}{r-1} x^{-ra}, x > 0$$

and compounding distribution with the density (3.8). Moreover,

$$\lim_{n\to\infty} P\left(\frac{\max(Y_{r:k}^{(1)}, Y_{r:k}^{(2)}, \ldots, Y_{r:k}^{(n-k+1)})}{cn^{1/ra}} \leq x\right) = e^{-\lambda(x)}, x > 0,$$

while a better approximation could be achieved by using

$$\lambda^*(x) = \binom{k-1}{r-1} x^{-ra} \left(1 - \frac{x^{-a}}{n^{1/r}}\right)^{k-r+1}, x > 0$$

in place of $\lambda(x)$.

By way of example we mention also that, should one be interested in the distribution of the second largest among the moving order statistics $Y_{r:k}^{(i)}$, $i = 1, 2, \ldots, n-k+1$, he could resort to the approximate formula (apply Theorem 3.4.3 for $m = 2$ and recall (3.7))

$$P\left(Y_{2:r:k}(n) \leq x\right) \approx e^{-\lambda^*\left(\frac{x}{cn^{1/ra}}\right)} \left(1 + \frac{r-1}{k-1}\lambda^*\left(\frac{x}{cn^{1/ra}}\right)\right).$$

In Figure 3.1 we present the distribution of $Y_{m:r:k}(n)$ (adequately normalized) for $n = 50$ and $n = 500$ for four choices of the parameters m, k, r. The smooth curve displays the asymptotic distributions established above while the shaded histogram was obtained by the 100,000 simulated values of $Y_{m:r:k}(n)/\ (cn^{1/ra})$ (where Y_i, $i = 1, 2, \ldots, n$ follow a Pareto distribution with $c = 1, a = 2$).

b. Uniform distribution

As a second example let us consider a sequence of random variables $Y_1, Y_2, \ldots$ which follow the uniform distribution

$$F(x) = x, \quad 0 < x < 1.$$

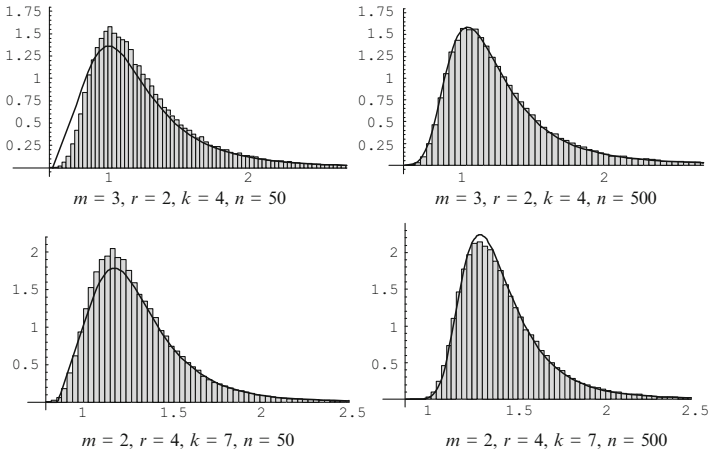


Figure 3.1. Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the Pareto distribution $F(x) = 1 - x^{-2},\ x \geq 1$.

Since $\overline{F} \in MDA(\Psi_a)$ with right end point $x_F = 1$, we may use the norming constants $c_n = x_F - F^{-1}(1 - n^{-1}) = n^{-1}, d_n = x_F = 1$ to gain the asymptotic results

$$\lim_{n\to\infty} P\left(\frac{Y_{m:r:k}(n) - 1}{n^{-1/r}} \leq x\right) = \lim_{n\to\infty} P(W_{n,k,r}(n^{-1/r}x + 1)) < m)$$
$$= \sum_{i=0}^{m-1} f_{CP}(x; i), \quad x > 0$$

with the parameter $\lambda(x)$ of the compound Poisson distribution given by (3.13).

In Figure 3.2 the exact distribution of $Y_{m:r:k}(n)$ (estimated by simulation) is compared to the approximate distribution gained by the last formula for the same set of choices for the parameters n, m, r, k as before (in order to achieve a better accuracy for the asymptotic results, formula (3.14) was used for the compound Poisson distribution parameters instead of (3.13).

c. Normal and exponential distribution

Two typical examples of the distributions belonging to the maximum domain of attraction of the Gumbel distribution are the exponential with mean $1/\beta$ and the standard normal distribution. A set of appropriate norming constants is offered by (see e.g. Table 3.4.2 in Embrechts *et al.* (1997))

$$c_n = \beta^{-1}, \quad d_n = \beta^{-1} \ln n,$$

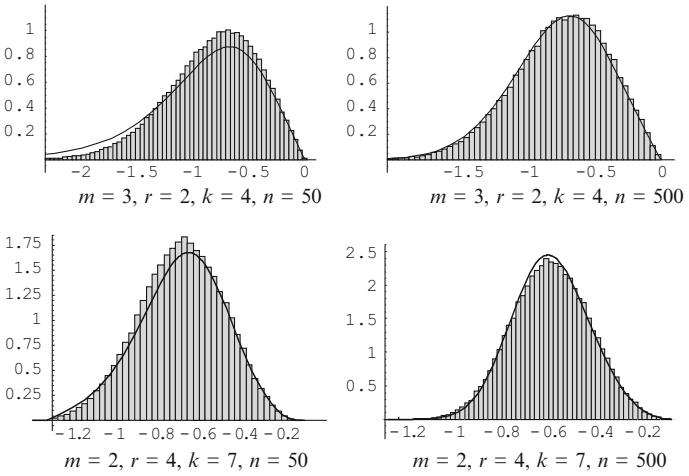


Figure 3.2. Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the uniform distribution $F(x) = x,\ 0 < x < 1$.

and

$$c_n = (2\ln n)^{-1/2}, \qquad d_n = (2\ln n)^{1/2} - \frac{\ln 4\pi + \ln\ln n}{2(2\ln n)^{1/2}},$$

respectively (the second pairs consists of a set of reasonable approximations for the norming constants). A direct application of Theorem 3.4.3 reveals that the asymptotic distribution of $Y_{m:r:k}(n)$, after carrying out a proper normalization, may be approximated by the aid of a compound Poisson with the parameter $\lambda(x)$ obtained from formula (3.15) (for a better approximation for finite n, we may use formula (3.16) instead the one given in (3.8)).

In Figures 3.3 and 3.4 a graphical comparison between the exact and asymptotic distributions of $Y_{m:r:k}(n)$ is carried out for exponential and normal sequences $Y_1, Y_2, \ldots$, respectively. The low quality of the approximation observed in Figure 3.4 should be attributed to the slow convergence of the distribution of the maximum of normal variables to the Gumbel distribution (which is due to the fact that the rate of convergence of $n\overline{\Phi}(c_n x + d_n)$ to e^{-x} is of order $O((\ln n)^{-1})$.

In closing, we mention that a series of results relating to minima (instead of maxima) and the asymptotic behavior under the assumption that the underlying distribution belongs to a minimum domain of attraction of the three extreme type distributions could also be established. Since these outcomes follow immediately from the corresponding results established here by using $-Y_i$ in the place of Y_i, we shall not pursue these topics here.

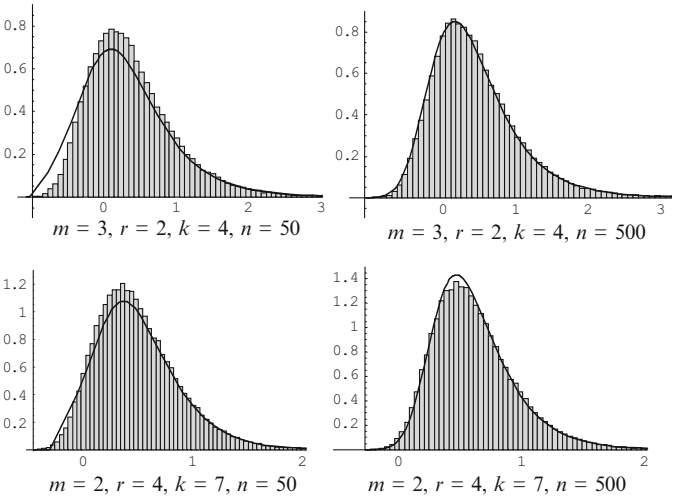


Figure 3.3. Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the exponential distribution $F(x) = 1 - e^{-x},\ x \geq 0$.

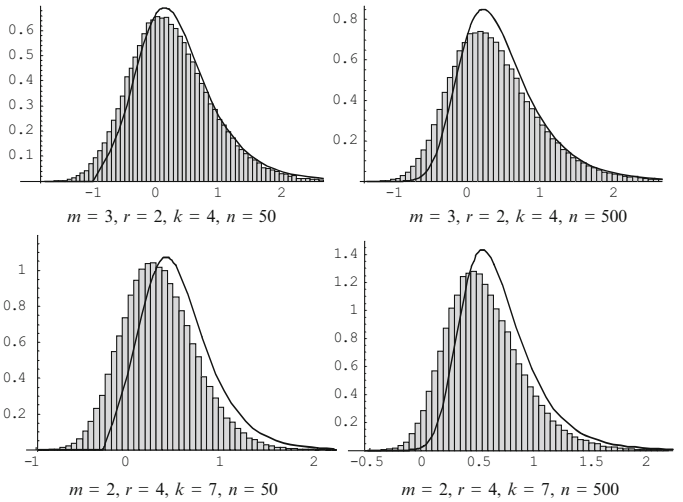


Figure 3.4. Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the (standard) normal distribution $\Phi(x),\ x \in \Re$.

Acknowledgment

The research of FSM was supported by the National Scholarship Foundation of Greece.

References

1. Arnold, B.C. (1985). Pareto distributions, In *Encyclopedia of Statistical Sciences*, S. Kotz, N.L. Johnson and C.B. Read (editors), 568–574, John Wiley & Sons, New York.
2. Arnold, B.C. and Balakrishnan, N. (1989). *Relations, Bounds, and Approximations for Order Statistics*, Springer, New York.
3. Arratia, R.L., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method, *Statistical Science*, **5**, 403–423.
4. Arratia, R.L., Gordon, L. and Waterman, M. (1990). The Erdős-Rényi Law in distribution, for coin tossing and sequence matching, *Annals of Statistics*, **18**, 539–570.
5. Balakrishnan, N. and Koutras, M.V. (2002). *Runs, Scans and Applications*, John Wiley & Sons, New York.
6. Barbour, A.D., Holst, L. and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, Oxford.
7. Boutsikas, M.V. and Koutras, M.V. (2001). Compound Poisson approximation for sums of dependent random variables, In Ch.A. Charalambides, M.V. Koutras, N. Balakrishnan (eds), *Probability and Statistical Models with Applications*, 63–86, Chapman & Hall, Boca Raton, FL.
8. Boutsikas, M.V. and Koutras, M.V. (2002). Modeling claim exceedances over thresholds, *Insurance: Mathematics and Economics*, **30**, 67–83.
9. Boutsikas, M.V. and Koutras, M.V. (2006). On the asymptotic distribution of the discrete scan statistic, *Journal of Applied Probability*, **43**, 1137–1154.
10. Bowers, N.L., Gerber, H.U., Hickman, J., Jones, D.A. and Nesbitt, C.J. (1997). *Actuarial Mathematics*, 2nd edition, The Society of Actuaries, Illinois.
11. Chen, J. and Glaz, J. (1999). Approximations for the distribution and the moments of discrete scan statistics, In *Scan Statistics and Applications*, J. Glaz and N. Balakrishnan, (eds), Birkhäuser, Boston, MA.
12. Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London.
13. David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, (3rd edition), John Wiley & Sons, New York.

14. Deheuvels, P. and Devroye, L. (1987). Limit laws of Erdős-Rényi-Shepp type, *The Annals of Probability*, **15**, 1363–1386.

15. Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes, *Annals of Applied Probability*, **2**, 329–357.

16. Dudkiewicz, J. (1998). Compound Poisson approximation for extremes for moving minima in arrays of independent random variables, *Applicationes Mathematicae*, **25**, 19–28.

17. Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.

18. Erdős, P. and Rényi, A. (1970). On a new law of large numbers, *Journal d'Analyse Mathematique*, **23**, 103–111.

19. Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bistate trials, *Journal of Applied Probability*, **38**, 908–916.

20. Fu, J.C. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and Its Applications*, Word Scientific Publishing, Singapore.

21. Glaz, J. and Balakrishnan, N. (eds.) (1999). *Scan Statistics and Applications*, Birkhäuser, Boston, MA.

22. Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *The Annals of Applied Probability*, **1**, 306–318.

23. Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer-Verlag, New York.

24. Goldstein, L. and Waterman, M. (1992). Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons, *Bulletin of Mathematical Biology*, **54**, 785–812.

25. Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, John Wiley & Sons, New York.

26. Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*, Imperial College Press, London.

27. Koutras, M.V. and Alexandrou, V.A. (1995). Runs, scans and run model distributions: a unified Markov chain approach, *Annals of the Institute of Statistical Mathematics*, **47**, 743–766.

28. Reiss, R.D. and Thomas, M. (1997). *Statistical Analysis of Extreme Values*, Birkhäuser, Basel.

4

Boundary Crossing Probability Computations in the Analysis of Scan Statistics

Hock Peng Chan,[1] I-Ping Tu,[2] and Nancy Ruonan Zhang[3]

[1] *Department of Statistics and Applied Probability, National University of Singapore, Singapore, Republic of Singapore*
[2] *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*
[3] *Department of Statistics, Stanford University, Stanford, CA, USA*

Abstract: The theory of boundary crossing probabilities in the study of repeated likelihood ratio tests was developed by Lai, Siegmund and Woodroofe in a series of articles and monographs appearing in the late 1970s and early to mid 1980s. This work formed part of the foundation for subsequent developments in the analysis of maxima of Gaussian and Poisson random fields used to provide accurate tail probability approximations of scan statistics. In this chapter, we (i) track these theoretical developments, (ii) study their applications on spatial scan statistics in astronomy and epidemiological studies and (iii) relate these theoretical developments to scan statistics used recently in genomics.

Keywords and phrases: Astronomy, boundary crossing probability, DNA copy number, epidemiology, genomics, maxima of random fields, neuroscience, scan statistic

4.1 Introduction

The study of scan statistics to detect either a signal at an unknown location or the presence of spatial clustering in a compact domain is a very active area of research, and the areas of applications are diverse, including astronomy, epidemiology, genomics, neuroscience, botany and ecology. The basic idea is as follows. A list of spatial or space-time vectors $\mathbf{x}_1, \ldots, \mathbf{x}_J$ associated with the occurrence of certain events of interest is observed in a domain D. In addition, there may also be a random variable or vector X_j that provides additional information on the jth occurrence for each $1 \leq j \leq J$. If there is a source of a cluster at an unknown location $\mathbf{t}$ (or a signal centered at $\mathbf{t}$), it may result either in an unusually large number of occurrences near $\mathbf{t}$ or the distribution of X_j

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_4,

might be different when $\mathbf{x}_j$ is near $\mathbf{t}$. For example, in case-control datasets in epidemiological studies, $X_j = 1$ denotes the occurrence of a case and $X_j = 0$ the occurrence of a control. When there is a source of a cluster of cases at $\mathbf{t}$, the probability that $X_j = 1$ will be higher when $\mathbf{x}_j$ is near $\mathbf{t}$. A score $S(\mathbf{t})$ is computed from $\{(\mathbf{x}_j, X_j) : 1 \le j \le J\}$ and a high score is expected when the source of the cluster is at $\mathbf{t}$. Since $\mathbf{t}$ is unknown, the scan statistic $M := \sup_{\mathbf{t}\in D} S(\mathbf{t})$ is the summary score for the presence of a cluster in D.

Lai and Siegmund (1977, 1979), Woodroofe (1978, 1982) and Siegmund (1985) developed a set of techniques to study boundary crossing probabilities of generalized likelihood ratio (GLR) sequential test statistics. These techniques were subsequently refined and extended by many researchers so that they can be applied on a wide variety of settings. We track these developments in Section 4.2 and elaborate on their applications in scan statistics in astronomy and epidemiology in Section 4.3 and genomics in Section 4.4. We conclude the paper with a few brief remarks in Section 4.5.

4.2 Theoretical Developments

Throughout this paper, $\mathbf{I}$ shall denote the indicator function, $|\cdot|$ the Lebesgue measure of a set or the determinant of a square matrix and $\|\cdot\|$ the L_2 norm. In addition, $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ and $\Phi(y) = \int_{-\infty}^{y} \varphi(x)\, dx$ are the density and cumulative distribution, respectively, of the standard normal. We write $a_n \sim b_n$ if $\lim_{n\to\infty}(a_n/b_n) = 1$. If $\mathbf{t} = (t_1, \ldots, t_d) \in \mathbf{R}^d$ and A is a subset of $\mathbf{R}^d$, then for any $w > 0$, $\mathbf{t} + wA = \{\mathbf{t} + w\mathbf{u} : \mathbf{u} \in A\}$. Before proceeding to the analytical techniques, we give a few examples to illustrate how the scores $S(\mathbf{t})$ are defined in different settings.

Example 4.2.1 Let J be either a fixed positive integer or a Poisson random variable. Assume that under the null hypothesis of no clustering, $\mathbf{x}_1, \ldots, \mathbf{x}_J$ are independent and identically distributed (i.i.d.) random variables uniformly distributed on a compact domain D. Let A be a nice compact set, for example, the box kernel $A = \{\mathbf{u} : \max_i |u_i| \le w/2\}$ or the spherical kernel $A = \{\mathbf{u} : \|\mathbf{u}\| \le w\}$ for some $w > 0$. Let $S(\mathbf{t})$ be the number of occurrences $\mathbf{x}_j$ lying inside $\mathbf{t} + A$ and M the corresponding scan statistic. Naus (1965, 1966, 1982), Huntington and Naus (1975) and Glaz (1989) provided approximate and exact p-value calculations of M when A is the box kernel. See Glaz, Naus and Wallenstein (2001) for comparisons against competing p-value approximations and bounds and also for a good overview of recent developments in scan statistics.

Example 4.2.2 Let $\mathbf{x}_1, \ldots, \mathbf{x}_J$ be the points on a lattice grid in a compact domain D. The detection of a signal is of interest here. Under the null hypothesis

of no signal, $X_1, \ldots, X_J$ are i.i.d. random variables from a baseline distribution F with log moment generating function $\psi(\theta) := \log Ee^{\theta X_1}$. Assume that $\Theta := \{\theta : \psi(\theta) < \infty\}$ is finite in a neighborhood of 0. Then the rate function of F is given by $\phi(\mu) = \sup_{\theta\in\Theta}[\theta\mu - \psi(\theta)]$ and F can be embedded in an exponential family $\{F_\theta, \theta \in \Theta\}$, with $dF_\theta(x) = e^{\theta x - \psi(\theta)}dF(x)$. Let A be a given signal shape and consider the alternative hypothesis

H_1: there exists $\theta \neq 0$ and $\mathbf{t} \in D$ such that $X_1, \ldots, X_J$ are independent with $X_j \sim F_\theta$ if $\mathbf{x}_j \in \mathbf{t} + A$ and $X_j \sim F$ otherwise,

indicating that a signal of shape A is centered at some unknown $\mathbf{t} \in D$. The log GLR score for testing the null hypothesis against the alternative hypothesis is $S(\mathbf{t}) = n_\mathbf{t}\phi(\bar{X}_\mathbf{t})$, where $n_\mathbf{t}$ is the number of points $\mathbf{x}_j$ lying in $\mathbf{t} + A$ and $\bar{X}_\mathbf{t} = n_\mathbf{t}^{-1}\sum_{\mathbf{x}_j\in\mathbf{t}+A} X_j$. Tail probabilities for the maxima of $S(\mathbf{t})$ were computed in Siegmund and Yakir (2000) via a change of measure argument.

Example 4.2.3 Researchers in neuroscience are interested in knowing if a neural spike time pattern, for example the pattern observed when a bird is learning a new song while awake, is repeated when the bird is sleeping. See Dave and Margoliasch (2000) for a more elaborate introduction to the problem. Let $T > 0$ and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ be a given template spike time pattern with $0 \leq y_n \leq T$ for all n and $\mathcal{X} = \{x_1, \ldots, x_J\}$ the neural spike times when the bird is sleeping, with $0 \leq x_j \leq U$ for all j, U large compared to T. We want to check if the spike time pattern $\mathcal{Y}$ is repeated inside $\mathcal{X}$; in other words, if there exists a time t such that $t + \mathcal{Y}$ and $\mathcal{X} \cap [t, t + T]$ are similar.

In Chi, Rauske and Margoliasch (2003), a pattern-filtering algorithm was used to match the spike time patterns. Let f be a nonincreasing kernel scoring function on $[0, \infty)$ with $f(0) > 0$ and $\lim_{u\to\infty} f(u) < 0$. Common examples include the continuous Hamming window kernel

$$f(u) = \begin{cases} \frac{1}{2}(1-\beta) + \frac{1}{2}(1+\beta)\cos\left(\frac{\pi u}{\epsilon}\right) & \text{if } u < \epsilon \\ -\beta & \text{if } u \geq \epsilon, \end{cases}$$

or the box kernel

$$f(u) = \begin{cases} 1 & \text{if } u < \epsilon \\ -\beta & \text{if } u \geq \epsilon. \end{cases}$$

The score

$$S(t) = \sum_{x_j\in[t,t+T]} \max_{1\leq n\leq N} f(|x_j - t - y_n|)$$

provides the value of a match between $t + \mathcal{Y}$ and $\mathcal{X} \cap [t, t + T]$. In Chi (2004), under the assumption that x_1 and $x_{i+1} - x_i$, $i \geq 1$, are i.i.d. exponential random variables, the exponent of the tail probability of $M = \sup_t S(t)$ was obtained using large deviation theory. Using the theory of boundary crossing probabilities, Chan and Loh (2007) obtained a more precise estimate, an approximation of the tail probability of M.

We shall illustrate the techniques behind the computation of boundary crossing probabilities with the signal detection problem described in Example 4.2.2. Let $d = 1$ and $\bar{X}_{i,j} = (j-i)^{-1}\sum_{i+1}^{j} X_k$ when $i < j$. Let $X_1, \ldots, X_J$ be i.i.d. random variables with distribution F under the null hypothesis and let the score

$$S(i,j) = (j-i)\phi(\bar{X}_{i,j}),$$

where ϕ is defined in Example 4.2.2. Let the scan statistic

$$M = \sup_{0 \le i < j \le J, w_0 \le (j-i) \le w_1} S(i,j).$$

We shall consider here the computation of $P\{M \ge c\}$ when $\log J = o(c)$, $J/c \to \infty$ and $w_k \sim \alpha_k c$ for some $0 < \alpha_0 < \alpha_1$ as $c \to \infty$. The problem has applications in sequential change-point detection, and is solved, for normal X_j when $w_0 = 0$ and $w_1 = \infty$, in Siegmund and Ventrakaman (1995) and extended to Markovian X_j satisfying minorization and drift conditions and ϕ replaced by a general function in Chan and Lai (2002, 2003).

Large deviation approximations

Let $v_\mu = \frac{d^2}{d\theta^2}\psi(\theta)|_{\theta=\theta_\mu}$ and $\Lambda = \{\mu : \alpha_1^{-1} \le \phi(\mu) \le \alpha_0^{-1}\}$. Assume for convenience that F has a continuous bounded density and Λ is a compact set lying in the interior of the support of F. Then the saddlepoint approximation

$$P\{\bar{X}_{i_0,j_0} \in d\mu\} \sim \left(\frac{j_0 - i_0}{2\pi v_\mu}\right)^{1/2} e^{-(j_0-i_0)\phi(\mu)}\, d\mu \tag{4.1}$$

holds uniformly over $\mu \in \Lambda$. Our interest is focused on μ satisfying $(j_0 - i_0)$ $\phi(\mu) = c + x$ for some x either of order 1 or small compared to c.

Local random walk

The next step involves an analysis of the local behavior of $S(i,j)$ for (i,j) close to (i_0, j_0) when $S(i_0, j_0) = c + x$. Let $\mu = \bar{X}_{i_0,j_0}$ and let $\theta_\mu \in \Theta$ satisfy $\phi(\mu) = \theta_\mu\mu - \psi(\theta_\mu)$. Since $\frac{d}{d\mu}\phi(\mu) = \theta_\mu$, it follows from a Taylor series expansion that

$$\begin{aligned} S(i,j) &= (j-i)\phi((\bar{X}_{i,j} - \mu) + \mu) \doteq (j-i)[\phi(\mu) + (\bar{X}_{i,j} - \mu)\theta_\mu] \\ &= S(i_0,j_0) + \sum_{k=1}^{J} (\mathbf{I}_{\{k \in [i,j]\}} - \mathbf{I}_{\{k \in [i_0,j_0]\}})[\theta_\mu X_k - \psi(\theta_\mu)]. \end{aligned} \tag{4.2}$$

Clearly, X_k follows distribution F for $k \le i_0$ and $k > j_0$ irregardless of the conditioning on $\bar{X}_{i_0,j_0}$. In addition, by Siegmund (1988), X_k is asymptotically of distribution F_μ (that is F_{θ_μ}) and asymptotically independent (for a fixed number of random variables) for $i_0 < k \le j_0$, when we condition on $\bar{X}_{i_0,j_0} = \mu$. Hence, under the conditioning,

$$\sum_{k=1}^{J} (\mathbf{I}_{\{k \in [i,j]\}} - \mathbf{I}_{\{k \in [i_0,j_0]\}})[\theta_\mu X_k - \psi(\theta_\mu)] \Rightarrow W_{i-i_0} + \widetilde{W}_{j-j_0}, \tag{4.3}$$

where W and $\widetilde{W}$ are independent random walks with independent increments $[\theta_\mu X_n - \psi(\theta_\mu)]$ and $[\theta_\mu \widetilde{X}_n - \psi(\theta_\mu)]$, respectively, with $X_n \sim F_\mu$ for $n \geq 1$, $X_n \sim F$ for $n \leq 0$, $\widetilde{X}_n \sim F$ for $n \geq 1$ and $\widetilde{X}_n \sim F_\mu$ for $n \leq 0$. We shall denote by P_μ the probability when W and $\widetilde{W}$ have increments with these joint distributions.

We are now left with the task of combining these large deviation approximations and local random walks, and we shall highlight three approaches here.

(I) Conditioning on the last-exit (or first-passage) time. This is the method most closely identified with the techniques developed to analyze sequential GLR test statistics. Unlike in sequential analysis where only one index is involved and what the last time is needs no explanation, here we need to deal with two indices i and j. We handle this by defining an ordering $\succ$ with $(i, j) \succ (i_0, j_0)$ if either $i > i_0$ and $j = j_0$ both occur or if $j > j_0$ occurs. By (4.1)–(4.3), if $(j_0 - i_0)\phi(\mu) = c + x$, then

$$\begin{aligned} &P\{\bar{X}_{i_0,j_0} \in d\mu, (j-i)\phi(\bar{X}_{i,j}) < c \text{ for all } (i,j) \succ (i_0, j_0)\} \\ &\sim \left(\frac{c+x}{2\pi\phi(\mu)v_\mu}\right)^{1/2} e^{-c-x} P_\mu\left\{\max_{k\geq 1} W_k \leq -x\right\} \\ &\quad \times P_\mu\left\{\max_{k\leq 0} W_k + \max_{\ell \geq 1} \widetilde{W}_\ell \leq -x\right\} d\mu. \end{aligned} \tag{4.4}$$

We sum (4.4) over $j_0 \geq i_0 + c/\phi(\mu)$ for a fixed i_0, noting that x increases by $\phi(\mu)$ for each increase of j_0 by 1, integrate over $\mu \in \Lambda$ and sum over $1 \leq i_0 \leq J$ to obtain

$$P\{M \geq c\} \sim J\left(\frac{c}{2\pi}\right)^{1/2} e^{-c} \int_\Lambda \gamma(\mu)(\phi(\mu))^{-3/2} v_\mu^{-1/2}\, d\mu, \tag{4.5}$$

where

$$\gamma(\mu) = \int_0^\infty e^{-x} P_\mu\left\{\max_{k\geq 1} W_k \leq -x\right\} P_\mu\left\{\max_{k\leq 0} W_k + \max_{\ell\geq 1} \widetilde{W}_\ell \leq -x\right\} dx.$$

A rigorous justification of (4.5) is more involved, as given in Siegmund and Venkatraman (1995) for the case of normal X_i. They also provided a simplification, relating γ to the overshoot constant

$$\nu(x) = 2x^{-2} \exp\Big\{-2\sum_{n=1}^\infty n^{-1/2}\Phi\Big(-\frac{x\sqrt{n}}{2}\Big)\Big\} \quad (x > 0), \tag{4.6}$$

in the normal case. This is achieved via an identity in Siegmund (1992). Analogous overshoot constant expressions for general X_i, relevant to both p-value and sample size calculations, can be found in Woodroofe (1979), Tu and Siegmund (1999), Storey and Siegmund (2001) and Tu (2009).

(II) Conditioning on local or global maxima. Let (i_0, j_0) be the indices at which the maximal value $M = S(i_0, j_0) \geq c$ is attained. By (4.1)–(4.3), we obtain (4.5) with the alternative representation

$$\gamma(\mu) = P_\mu \left\{ \max_{k \neq 0} W_k < 0 \right\} P_\mu \left\{ \max_{\ell \neq 0} \widetilde{W}_\ell < 0 \right\}.$$

This approach is more commonly used when the score is obtained via a continuous kernel function. A good reference is Rabinowitz and Siegmund (1997), which considers signal detection on a homogeneous Poisson process. This work is discussed in more detail in Section 4.3.1.

(III) Conditioning below a high crossing. The first two approaches involve conditioning above a high level c. There is yet another approach, adapted by Hogan and Siegmund (1986) from tail probability approximations of Gaussian random fields developed in Pickands (1969), Bickel and Rosenblatt (1973) and Qualls and Watanabe (1973). Fix i_0 and j_0 and let them be multiples of n for some large n. We condition on $S(i_0, j_0) < c$, compute the conditional probability that $S(i, j)$ exceeds c for some (i, j) lying in the domain $[i_0, i_0 + n] \times [j_0, j_0 + n]$, then add up these probabilities over different $i_0 < j_0$. By (4.1)–(4.3), if $(j_0 - i_0)\phi(\mu) = c - x$, then

$$P\{\bar{X}_{i_0,j_0} \in d\mu, (j-i)\phi(\bar{X}_{i,j}) \geq c \text{ for some } (i,j) \in [i_0, i_0+n] \times [j_0, j_0+n]\}$$

$$\sim \left(\frac{c-x}{2\pi\phi(\mu)v_\mu} \right)^{1/2} e^{-c+x} P_\mu \left\{ \max_{0 \leq k \leq n} W_k + \max_{0 \leq \ell \leq n} \widetilde{W}_\ell \geq x \right\} d\mu. \tag{4.7}$$

We sum (4.7) over $i_0 \leq j_0 \leq i_0 + c/\phi(\mu)$ with j_0 a multiple of n and i_0 fixed, integrate over $\mu \in \Lambda$, then sum over $1 \leq i_0 \leq J$ with i_0 a multiple of n, while choosing n large, to obtain (4.5) with

$$\gamma(\mu) = \lim_{n \to \infty} n^{-2} \int_{-\infty}^{\infty} e^x P_\mu \left\{ \max_{0 \leq k \leq n} W_k + \max_{0 \leq \ell \leq n} \widetilde{W}_\ell \geq x \right\} dx.$$

Again, additional technical arguments are needed here for a rigorous justification of these calculations. This approach was used in Chan and Zhang (2007) to compute tail probabilities of weighted scan statistics and in Chan and Loh (2007) to compute tail probabilities of template scoring scan statistics. The first problem will be elaborated further in Section 4.4.1.

4.3 Applications in Spatial Scan Statistics

We focus here on two examples to illustrate how the theory of boundary crossing probabilities can be used to obtain analytical p-values for spatial or space-time scan statistics. We start off on a problem with motivations in astronomy. Note

that the calculations for continuous kernel functions [Rabinowitz and Siegmund (1997)] and kernels containing discontinuities [Loader (1991)] are different. We then consider the problem of detecting clusters in a nonhomogeneous population using a case-control dataset.

4.3.1 Searching for a source of muon particles in the sky

Continuous kernel functions

Consider a background of homogeneous random cosmic rays with known intensity λ. By taking D sufficiently large, we may assume that edge effects are absent and that the particles are observed on $\mathbf{R}^d$. We shall denote the set of observed particle locations by $\{\mathbf{x}_j\}_{j=1}^{\infty}$. Let f be a non-negative kernel function on $\mathbf{R}^d$ that satisfies $\int f^2(\mathbf{x})d\mathbf{x} = 1$, is smooth and symmetric in each argument and vanishes rapidly at infinity. One concrete example is the Gaussian kernel $f(\mathbf{x}) = \pi^{-d/4}e^{-\|\mathbf{x}\|^2/2}$. Let $\mu = \int f(\mathbf{x})d\mathbf{x}$ and let the score

$$S(\mathbf{t}) = \lambda^{-1/2}\Big[\sum_{j=1}^{\infty} f(\mathbf{x}_j - \mathbf{t}) - \lambda\mu]. \tag{4.8}$$

Let $P_{\theta,\mathbf{t}}$ ($E_{\theta,\mathbf{t}}$) denote the probability measure (expectation) under which $\{\mathbf{x}_j\}_{j=1}^{\infty}$ is generated from a nonhomogeneous Poisson process with intensity

$$\lambda_{\theta,\mathbf{t}}(\mathbf{x}) := \lambda \exp[\theta f(\mathbf{x} - \mathbf{t})], \tag{4.9}$$

and let $P_{\theta,\mathbf{0}}$ be denoted more simply by P_θ. The nonhomogeneous Poisson process motivates $S(\mathbf{t})$ as the efficient score statistics as we let $\theta \to 0$ and also provides the change of measure for computing the tail probabilities of the scan statistic $M = \sup_{t\in D} S(\mathbf{t})$.

We provide an outline of the calculations and arguments given in Rabinowitz and Siegmund (1997) and refer the reader to the article itself for the details. Fix $c > 0$ and let $b = c\lambda^{1/2}$. By the Poisson clumping heuristic, see, for example, Siegmund (1988) or Aldous (1989),

$$P_0\{M \geq b\} \approx 1 - e^{-E_0 K},$$

where K is the number of local maxima in D exceeding the threshold b. Since f is smooth, $\nabla S(\mathbf{t})$ and $\nabla^2 S(\mathbf{t})$, the gradient and Hessian, respectively, of S at $\mathbf{t}$, are both well defined and continuous. It follows from Theorem 6.1 of Adler (1981), using a local maxima conditioning argument, that

$$E_0 K = |D| E_\theta \left[\left(\frac{dP_0}{dP_\theta}\right) |\nabla^2 S(\mathbf{0})| \mathbf{I}_{\{S(\mathbf{0})\geq b, \nabla S(\mathbf{0})=0, \nabla^2 S(\mathbf{0})<0\}}\right], \tag{4.10}$$

where the statement "$\nabla^2 S(\mathbf{0}) < 0$" means $\nabla^2 S(\mathbf{0})$ is a negative definite matrix, and the expectation on the right-hand side of (4.10) is defined with respect to a joint probability-density. Let

$$\psi(\theta) = \log E_0[e^{\theta\lambda^{1/2}S(\mathbf{0})}] = \lambda \int [e^{\theta f(\mathbf{x})} - 1 - \theta f(\mathbf{x})]\, d\mathbf{x}.$$

Then

$$E_\theta(\lambda^{1/2}S(\mathbf{0})) = \psi'(\theta) = \lambda \int f(\mathbf{x})[e^{\theta f(\mathbf{x})} - 1]d\mathbf{x},$$
$$\mathrm{Var}_\theta(\lambda^{1/2}S(\mathbf{0})) = \psi''(\theta) = \lambda \int f^2(\mathbf{x})e^{\theta f(\mathbf{x})}d\mathbf{x}.$$

Let the rate function $I(\theta) = \theta\psi'(\theta) - \psi(\theta)$ and select θ to satisfy $\psi'(\theta) = c\lambda$. By a Gaussian approximation on the process $S(\mathbf{t})$ under P_θ, and making use of the relations

$$E_\theta[\nabla S(\mathbf{0})] = 0, \qquad \mathrm{Cov}_\theta(S(\mathbf{0}), \nabla S(\mathbf{0})) = \mathbf{0},$$

$$E_\theta[\lambda^{1/2}\nabla^2 S(\mathbf{0})] = -\theta \mathrm{Cov}_\theta(\lambda^{1/2}\nabla S(\mathbf{0})), \qquad E_\theta(\nabla^2 S(\mathbf{0}), \nabla S(\mathbf{0})) = \mathbf{0},$$

$$\mathrm{Cov}_\theta\left(\frac{\partial}{\partial t_i}S(\mathbf{0}), \frac{\partial}{\partial t_j}S(\mathbf{0})\right) = \mathbf{I}_{\{i=j\}} \int \left[\frac{\partial}{\partial x_i}f(\mathbf{x})\right]^2 e^{\theta f(\mathbf{x})}d\mathbf{x},$$
$$\mathrm{Cov}_\theta(S(\mathbf{0}), \nabla^2 S(\mathbf{0})) = \int f(\mathbf{x})\nabla^2 f(\mathbf{x})e^{\theta f(\mathbf{x})}d\mathbf{x},$$

Rabinowitz and Siegmund obtained the approximation

$$E_0 K \sim \theta^{d-1}e^{-I(\theta)}(2\pi)^{-(d+1)/2}|D| \left\{ \frac{\prod_{i=1}^d \mathrm{Var}_\theta\left(\lambda^{1/2}\frac{\partial}{\partial t_i}S(\mathbf{0})\right)}{\mathrm{Var}_\theta(\lambda^{1/2}S(\mathbf{0}))} \right\}^{1/2}.$$

Rabinowitz and Siegmund (1997) also considered scaling of f by an unknown σ to capture clusters of different sizes. Consider the more general score function

$$S(\mathbf{t}, \sigma) = \lambda^{-1/2}\left[\sigma^{-d/2}\sum_{j=1}^{\infty} f\left(\frac{\mathbf{x}_j - \mathbf{t}}{\sigma}\right) - \sigma^{d/2}\lambda\mu\right],$$

and let the scan statistic $M_{\sigma_0,\sigma_1} = \sup_{\mathbf{t}\in D, \sigma_0\le\sigma\le\sigma_1} S(\mathbf{t}, \sigma)$, where $0<\sigma_0<\sigma_1<\infty$. We refer the reader to Rabinowitz and Siegmund (1997) pp. 175–179 for the tail approximation of M_{σ_0,σ_1}, which involves a more complicated derivation.

Kernel functions containing discontinuities

When f is not continuous, then $S(\mathbf{t})$ is also not continuous, and the approach given above does not work. We illustrate the general approach with the box-shaped kernel

$$f = \mathbf{I}_{A_\Delta}, \text{ where } A_\Delta = \{(x_1, x_2) : 0 \le x_1 \le \Delta_1, 0 \le x_2 \le \Delta_2\},$$

considered in Loader (1991). Let $N(\mathbf{t}, \Delta)$ denote the number of points $\mathbf{x}_j$ lying inside $\mathbf{t} + A_\Delta$. Let $D = [0,1]^2$ and consider $(\mathbf{t}, \Delta)$ such that $\mathbf{t} + A_\Delta \subset D$. We shall use as our score function at $(\mathbf{t}, \Delta)$, the log GLR test statistic for testing

H_0: intensity of Poisson process is λ at all $\mathbf{t} \in D$,

vs. H_1: intensity at $\mathbf{x}$ is $\lambda(\mathbf{x}) = \lambda \exp(\theta \mathbf{I}_{\{\mathbf{x} \in \mathbf{t} + A_\Delta\}})$ for some $\theta > 0$.

Let $\mathbf{t} \prec \mathbf{u}$ if $t_i < u_i$ for all i. Then

$$S(\mathbf{t}, \Delta) = \Big\{ N(\mathbf{t}, \Delta) \log \left(\frac{N(\mathbf{t}, \Delta)}{n \Delta_1 \Delta_2} \right) + [n - N(\mathbf{t}, \Delta)] \log \left(\frac{n - N(\mathbf{t}, \Delta)}{n(1 - \Delta_1 \Delta_2)} \right) \Big\} \mathbf{I}_{\{N(\mathbf{t}, \Delta) \geq n \Delta_1 \Delta_2\}}, \quad (4.11)$$

where n is the total number of points in D, and we consider the scan statistic

$$M_{\mathbf{w}_1, \mathbf{w}_2} = \sup_{\mathbf{w}_1 \prec \Delta \prec \mathbf{w}_2} \left[\sup_{\mathbf{t} + A_\Delta \subset D} S(\mathbf{t}, \Delta) \right], \quad (4.12)$$

for some $\mathbf{0} \prec \mathbf{w}_1 \prec \mathbf{w}_2$.

Loader (1991) first considered the case of fixed Δ and n. Let $D' = [0, 1 - \Delta_1] \times [0, 1 - \Delta_2]$ and consider the lattice grid $D'_\delta = D' \cap (\delta \mathbf{Z})^2$. Let $M = \sup_{\mathbf{t} \in D'} N(\mathbf{t}, \Delta)$ and $M_\delta = \sup_{\mathbf{t} \in D'_\delta} N(\mathbf{t}, \Delta)$. Let $P^{(n)}$ denote probability conditioned on n. Using the first-passage time approach given in (I) of Section 4.2, the tail approximations of $M_\delta := \sup_{\mathbf{t} \in D'_\delta} N(\mathbf{t}, \Delta)$ is first obtained. By using a good bound of $P^{(n)}\{M - M_\delta > 0\}$ for small $\delta > 0$, Loader (1991) showed that for any $\epsilon > 0$ with $\Delta_1 \Delta_2 (1 + \epsilon)$ rational,

$$P^{(n)}\{M \geq m\} \sim \frac{n^2 \Delta_1 \Delta_2 (1 - \Delta_1)(1 - \Delta_2) \epsilon^3}{(1 - \Delta_1 \Delta_2)^3 (1 + \epsilon)} P^{(n)}\{N(\mathbf{0}, \Delta) = m\},$$

as $m \to \infty$ with $m = n \Delta_1 \Delta_2 (1 + \epsilon)$ a positive integer.

We shall now proceed to the tail probabilities of $M_{\mathbf{w}_1, \mathbf{w}_2}$. For given $\eta > 0$, let $h(\Delta)$ be defined implicitly as a solution to the equation

$$h(\Delta) \log \left(\frac{h(\Delta)}{\Delta} \right) + [1 - h(\Delta)] \log \left(\frac{1 - h(\Delta)}{1 - \Delta} \right) = \frac{\eta^2}{2}, \quad (4.13)$$

subject to the constraint $h(\Delta) > \Delta$. Let $c = \eta \sqrt{n}$. Then by (4.11) and (4.12),

$$\{M_{\mathbf{w}_1, \mathbf{w}_2} \geq c^2/2\} = \left\{ \sup_{\mathbf{w}_1 \prec \Delta \prec \mathbf{w}_2} \sup_{\mathbf{t} + A_\Delta \subset D} [N(\mathbf{t}, \Delta) - n h(\Delta_1 \Delta_2)] \geq 0 \right\}. \quad (4.14)$$

The local random walk analysis of $S(\mathbf{t}, \Delta)$ involves both a tangent approximation

$$h(\Delta') \doteq h(\Delta) + (\Delta' - \Delta) h'(\Delta)$$

and a decomposition

$$N(\mathbf{t}', \Delta') - N(\mathbf{t}, \Delta) \doteq Z_1(t_1' - t_1) + Z_2(t_2' - t_2) + Z_3(t_1' - t_1 + \Delta_1' - \Delta_1) + Z_4(t_2' - t_2 + \Delta_2' - \Delta_2),$$

where $Z_1, \ldots, Z_4$ are independent two-sided Poisson processes. Then

$$P^{(n)}\{M_{\mathbf{w}_1,\mathbf{w}_2} \geq c^2/2\} \sim c^7 \phi(c) \int_{u_0}^{u_1} \frac{u^2}{\eta^7 [h'(u)]^3} \left(h'(u) - \frac{1-h(u)}{1-u} \right)^4 \times \left(\frac{1-h(u)}{1-u} - \frac{h(u)}{u} \right)^3 \left(\frac{-(1+u)\log u - 2(1-u)}{\sqrt{h(u)(1-h(u))}} \right) du, \tag{4.15}$$

where $u_0 = w_{10}w_{20}$ and $u_1 = w_{11}w_{21}$ are the areas of the smallest and largest windows, respectively. A simulation study conducted in Loader (1991) shows (4.15) to be more accurate than the approximation obtained using an asymptotic Gaussian process argument.

4.3.2 Case-control epidemiological studies

In the detection of disease clusters, we have to adjust for the nonhomogeneity of the underlying population, both in terms of the population, density and the distribution of disease risk factors like gender, age or ethnic group. One way to achieve this is through a case-control epidemiological study; see, for example, Whittemore *et al.* (1987), Cuzick and Edwards (1990), Diggle (1990) and Kulldorff (1997).

Assume we have a dataset of locations of disease cases and a corresponding dataset of locations of healthy controls. We merge the two datasets into one and denote it by $\{(\mathbf{x}_j, X_j) : 1 \leq j \leq J\}$, $\mathbf{x}_j$ denoting the location of the jth subject with $X_j = 1$ if it corresponds to a case and $X_j = 0$ if it corresponds to a control.

We focus here on the model proposed in Diggle (1990) to test if there exists a location risk factor that increases the occurrence rate of cases. Let $\lambda(\mathbf{x})$ be the rate of generating controls at position $\mathbf{x}$ and let $\rho\lambda(\mathbf{x})e^{\theta g(\mathbf{x},\mathbf{t})}$ be the rate of generating cases at position $\mathbf{x}$ with $\theta > 0$ when there is a risk factor at $\mathbf{t}$ and $\theta = 0$ when there is no risk factor. The semi-parametric likelihood is proportional to

$$\prod_{j=1}^{J} \{[\lambda(\mathbf{x}_j)\rho e^{\theta g(\mathbf{x}_j,\mathbf{t})}]^{X_j} [\lambda(\mathbf{x}_j)]^{1-X_j}\}$$

while the conditional likelihood for given $\mathbf{x}_1, \ldots, \mathbf{x}_J$ and $I = \sum_{j=1}^{J} X_i$ is

$$\frac{\prod_{j=1}^{J} e^{X_j \theta g(\mathbf{x}_j,\mathbf{t})}}{\sum_{\alpha \in U} \prod_{j=1}^{J} e^{\mathbf{I}_{\{j\in\alpha\}} \theta g(\mathbf{x}_j,\mathbf{t})}},$$

where U is the class of all $\binom{J}{I}$ subsets of $\{1,\ldots,J\}$ of size I. Let $\widehat{p}_0 = I/J$ and $\bar{g}(\mathbf{t}) = J^{-1}\sum_{j=1}^{J} g(\mathbf{x}_j,\mathbf{t})$. Then the efficient score statistic for testing the presence of a localized risk factor at $\mathbf{t}$ is

$$T_{\mathbf{t}} = \sum_{j=1}^{J}(X_j - \widehat{p}_0)[g(\mathbf{x}_j,\mathbf{t}) - \bar{g}(\mathbf{t})]. \tag{4.16}$$

Let the normalized score $S(\mathbf{t}) = T_{\mathbf{t}}/\sqrt{\mathrm{Var}(T_{\mathbf{t}})}$, where $\mathrm{Var}(T_{\mathbf{t}}) = \widehat{p}_0(1-\widehat{p}_0)$ $(J-2)\sum_{j=1}^{J}[g(\mathbf{x}_j,\mathbf{t}) - \bar{g}(\mathbf{t})]^2/(J-1)$. Rabinowitz (1994) obtained p-value estimates of $M = \sup_{t\in D} S(\mathbf{t})$ by applying the tail probability approximation of a Gaussian process having the same covariance structure as $S(\mathbf{t})$. Let $\sigma_{\mathbf{t},\mathbf{u}} = \mathrm{Cov}(S(\mathbf{t}), S(\mathbf{u}))$, $\Lambda_{\mathbf{t}}$ a matrix with (i,j)th element $-\left(\frac{\partial^2\sigma(\mathbf{s},\mathbf{u})}{\partial s_i \partial s_j}\right)\Big|_{\mathbf{s}=\mathbf{u}}$ and $\Lambda'_{\mathbf{t}} = P_{\mathbf{t}}^T \Lambda_{\mathbf{t}} P_{\mathbf{t}}$, where $P_{\mathbf{t}}$ is a $d \times (d-1)$ matrix comprising of orthonormal vectors of the tangent space of the boundary ∂D at $\mathbf{t}$. Then by Knowles and Siegmund (1989), Corollary 2,

$$P\{M > b\} \approx (2\pi)^{-d/2} b^{d-1}\varphi(b)\left(\int_D |\Lambda_{\mathbf{t}}|^{1/2} d\mathbf{t} + (\pi/2)^{1/2} b^{-1}\int_{\partial D} |\Lambda'_{\mathbf{t}}|^{1/2} d\mathbf{t}\right). \tag{4.17}$$

The SaTScan software developed by Kulldorff (2006) and Information Management Services, Inc. considers $g(\mathbf{x},\mathbf{t}) = \mathbf{I}_{\{\mathbf{x}-\mathbf{t}\|\le w\}}$ for some $w > 0$. Let $m_{\mathbf{t},w}$ and $n_{\mathbf{t},w}$ be the total number of cases and the total number of occurrences (=cases+controls), respectively, in $\{\mathbf{u} : \|\mathbf{u}-\mathbf{t}\| \le w\}$. Instead of the efficient score statistic, they consider the log GLR score

$$S(\mathbf{t},w) = [n_{\mathbf{t},w}\phi(m_{\mathbf{t},w}/n_{\mathbf{t},w}) + (I-n_{\mathbf{t},w})\phi((J-m_{\mathbf{t},w})/(I-n_{\mathbf{t},w}))]\mathbf{I}_{\{m_{\mathbf{t},w}/n_{\mathbf{t},w} > \hat{p}_0\}},$$

where $\phi(p) = p\log(p/\widehat{p}_0) + (1-p)\log[(1-p)/(1-\widehat{p}_0)]$. In the SaTScan software, p-values of the scan statistics, including scan statistics involving other types of data, are computed using permutation tests.

4.4 Recent Applications in Genomics

Scan statistics are useful for interpreting genomes in the post-sequencing phase. They play an exploratory role, with the goal of locating genomic regions exhibiting properties of extreme deviation to be singled out for further testing. There is a rich source of statistical problems here, many still relatively unexplored. Due to space constraints, we focus only on two examples because the description and solution of each category of problems require a different set of

domain knowledge. The first problem is on the scanning of a DNA sequence for predefined word patterns. The second is on the analysis of genomic profiling data, in particular DNA copy number profiling.

4.4.1 Biomolecular sequence analysis

DNA and protein sequences can be modeled as a linear sequence drawn from a stationary distribution on an alphabet representing either the 21 amino acids in the case of protein sequences, or the bases A,C,G and T in the case of DNA sequences. Over the years, researchers have identified specific word patterns that are associated with either the encouragement or suppression of certain biological activity.

Transcription factors are proteins that bind to specific parts of DNA, known as transcription factor binding sites (TFBSs), to control the timing and rate of transcription of DNA into RNA. The TFBSs are identified by scoring with respect to certain scoring matrices, and the presence of a cluster of these sites indicates that genes regulated by the associated transcription factors may be located nearby. Lifanov *et al.* (2003) successfully used scan statistics to locate clusters of binding sites in DNA sequences by counting the number of TFBS located in a sliding window, while Rajewsky *et al.* (2002) weighed the TFBS by the scores obtained from the scoring matrices.

A more classical application of scan statistics in counting word patterns is in the identification of origins of replication in viruses, cf. Masse *et al.* (1992). The four letters in the DNA alphabet can be divided into two complementary pairs with A–T one pair and C–G the second pair. In DNA sequences, a palindrome is a DNA word which, when read backwards, has the complementary spelling of the original word. For example, the word ACGCGCGT is a palindrome because its letter-wise complementary spelling is TGCGCGCA. In bacterial and viral genomes, palindromes occur with unusually high frequency near locations associated with the initiation of replication, known as origins of replication.

Karlin and Brendel (1992) formulated the r-scan statistic to detect anomalies in the spacing between occurrences of word patterns. Let n be the length of the genomic sequence and $x_1 < \cdots < x_J$ the locations of the patterns. Let $d_j = x_{j+1} - x_j$ be the inter-feature distances, $A_i^{(r)} = \sum_{k=i}^{i+r-1} d_k$ the r-scan process and $A^{(r)} = \min_{1 \le i \le J-r} A_i^{(r)}$ the minimal r-scan. Let $N_u(t)$ be the number of word patterns in the interval $(t, t+u]$ and $M_u = \sup_{0 \le t \le n-u} N_u(t)$ the maximal scan statistic. Then we have the duality

$$\{M_u \ge r+1\} = \{A^{(r)} \le u\},$$

and the two scan statistics can be used interchangeably. P-value approximations for the significance of r-scans were obtained by Arratia, Goldstein and Gordon

(1989) and Glaz *et al.* (1994) using Poisson and compound Poisson approximations, respectively. See also Leung and Yamashita (1999) for the applications of these p-value approximations on palindrome counting scan statistics.

In addition to Rajewsky *et al.* (2002), weighted scan statistics was also considered in Chew, Choi and Leung (2005) for scoring palindromic patterns, which we consider here to be palindromes having a length of at least ten DNA letters. Since the length of a palindrome must be even, Chew *et al.* let $X_j = \ell_j/2$, where ℓ_j is the length of the jth palindromic pattern. Let $S_u(t) = \sum_{\mathbf{x}_j \in (t,t+u]} X_j$ and let the weighted scan statistic $M_{n,u} = \sup_{0 \le t \le n-u} S_u(t)$. Chan and Zhang (2007) used a marked Poisson process approximation of $S_u(t)$ to obtain an approximation of the p-value of $M_{n,u}$. Let F be the distribution of X_j, which we assume to have positive mean μ. Let λ be the probability of observing a palindromic pattern at any one location. Let $K(\theta) = E(e^{\theta X_1})$ and for given $x > \lambda\mu$, define $\theta_x (> 0)$ and $\alpha_x (> \lambda)$ to be the unique constants satisfying

$$K'(\theta_x) = x/\lambda, \qquad \alpha_x = \lambda K(\theta_x). \tag{4.18}$$

Let the large deviation rate function $I(x) = -(\alpha_x - \lambda) + \theta_x x$ and define F_θ to be the tilted distribution of F satisfying $F_\theta(dx) = e^{\theta x} F(dx)/K(\theta)$, with probability mass function (density) f_θ when F is discrete (continuous). Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with the mixture probability mass function (density)

$$g(y) = \Big(\frac{\alpha_x}{\lambda + \alpha_x}\Big) f_{\theta_x}(y) + \Big(\frac{\lambda}{\lambda + \alpha_x}\Big) f(-y), \tag{4.19}$$

and let $R_k = Y_1 + \cdots + Y_k$. Define the overshoot constant

$$\nu_x = \lim_{b \to \infty} E[e^{-\theta_x (R_{\tau_b} - b)}], \text{ where } \tau_b = \inf\{k \ge 1 : R_k \ge b\}, \tag{4.20}$$

with b a multiple of η if F is arithmetic with span η, in other words, if F has support on the grid $\{0, \pm\eta, \pm 2\eta, \ldots\}$ but not on a coarser lattice grid containing 0. By the approach of conditioning below a high crossing, see (III) in Section 4.2, Chan and Zhang (2007) showed that

$$P\{M_{n,u} \ge ux\} \sim 1 - \exp\left\{ -\frac{(n-u)\nu_x e^{-uI(x)}(x - \lambda\mu)}{\sqrt{2\pi u \lambda K''(\theta_x)}} \right\}, \tag{4.21}$$

if $u \to \infty$ and $(n-u) \to \infty$ as $n \to \infty$.

In Figure 4.1, we use (4.21) to obtain threshold levels corresponding to a p-value of 0.001 in the search for clusters of palindromic patterns with window size u equal to 0.5 % of the genome length. For the unweighted case, $X_j = 1$ for all palindromic patterns, while for the weighted case, we choose $X_j = (\ell_j/2) - 4$.

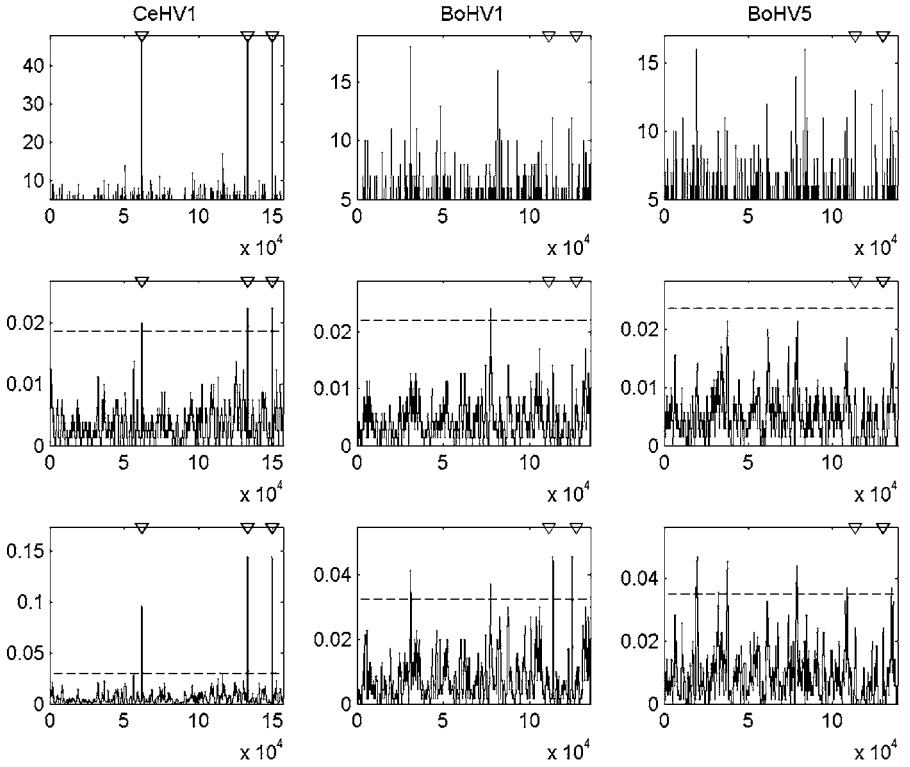


Figure 4.1. The x coordinate represents the locations of three well-known virus genomes. The y coordinate represents either half the length of the palindromic patterns (top plots), $u^{-1}N_u(t - u/2)$ for the unweighted case (middle plots) or $u^{-1}S_u(t - u/2)$ for the weighted case (bottom plots). The dotted lines are threshold levels corresponding to p-values of 0.001. The inverted triangles are experimentally validated origins of replication.

4.4.2 Detecting changes in DNA copy number

The DNA copy number is the number of copies of DNA at a region of a genome, the default being two for all human autosomes. The variation of this number, known as the DNA copy number variation (CNV), corresponds to gains and losses of specific chromosomal segments. These variations may be inherited [Redon *et al.* (2006)], or they may occur due to mutation and are then associated with certain diseases like cancer [Pinkel and Albertson (2005)]. In DNA copy number data, the quantity of homologous DNA present in a population of cells is measured by a set of probes, each mapping to a specific location in the genome.

Let X_j be the measured DNA quantity at probe j, relative to the expected value of two, at a fixed location x_j in the genome. We do not observe integer valued X_j due to inhomogeneity of the cell sample and substantial measurement error. Our objective is to partition the genome into segments of equal

copy number. We shall disregard irregularities in the spacing of the probe locations, a reasonable assumption for most experimental platforms and accepted in practice. Many different statistical methods have been applied to this problem; see Lai *et al.* (2005) for a broad survey of these methods. We shall focus here on the approach taken by Olshen *et al.* (2004). Consider a segment of the genome, containing J probes, which we would like to test for constant CNV. Define $\bar{X} = J^{-1}\sum_1^J X_j$ and $\hat{\sigma}^2 = J^{-1}\sum_1^J (X_j - \bar{X})^2$. Let

$$U(s,t) = \frac{\sum_{j=s+1}^t (X_j - \bar{X})}{\hat{\sigma}\sqrt{(t-s)[1-(t-s)/J]}}, \tag{4.22}$$

and

$$M = \max_{0\le s<t\le J, v_0<t-s<v_1} U^2(s,t). \tag{4.23}$$

When a significant p-value is obtained, for example by using the approximation in Siegmund (1986), we partition the segment and test each sub-segment further in the same manner.

Since most genomic profiling studies involve cohorts of individuals, it is of interest to pool samples together to gain power in detecting recurrent CNVs. This problem was first analyzed using hidden Markov models, cf. Shah *et al.* (2007), and has also been studied recently by Zhang *et al.* (2008) under the framework of a simultaneous scan of multiple aligned sequences for recurrent variant intervals of shared location. The formulation in Zhang *et al.* (2008) is as follows. For each sequence $i = 1, \ldots, N$ and position $j = 1, \ldots, J$, the random variables X_{ij} are mutually independent and normally distributed with mean values μ_{ij} and variances σ_i^2. Under the null hypothesis, $\mu_{i1} = \cdots = \mu_{iJ}$ for each sample i, and under the alternative hypothesis, there exists $\mathcal{J} \subset \{1, \ldots, N\}$ (with $\mathcal{J} \neq \emptyset$), and $\tau_1 < \tau_2$ with $v_0 \le (\tau_2 - \tau_1) \le v_1$ for some $1 \le v_0 \le v_1 < J$, such that for each $i \in \mathcal{J}$, $\mu_{ij} = \mu_{i0} + \delta_i \mathbf{I}_{\{\tau_1 < j \le \tau_2\}}$ for some $\delta_i \neq 0$. The GLR test in this setting yields the scan statistic

$$M = \max_{0\le s<t\le J, v_0\le t-s\le v_1} Z_{s,t}, \quad \text{where } Z_{s,t} = \sum_{i=1}^N \frac{[U_i^2(s,t) - 1]}{\sqrt{2N}}, \tag{4.24}$$

and $U_i(s,t)$ is defined as in (4.22) relative to the ith sequence.

The sum of chi-squares statistic in (4.24) pools signals from all samples, however weak. Zhang *et al.* (2008) also proposed a weighted sum of chi-squares statistic that requires individual sequences to show some evidence of a signal before it is allowed to contribute significantly to the pooled scan. Let $Q_i(s,t) = \mathbf{I}_{\{i\in\mathcal{J}\}}$ (the presence of (s,t) in the notation will be clear later). If $\mathcal{J}$ is known, then the log likelihood ratio statistic is

$$\max_{s<t} \sum_{i=1}^N \log\{[1 - Q_i(s,t)] + Q_i(s,t)e^{U_i^2(s,t)/2}\} = \max_{s<t} \sum_{i=1}^N Q_i(s,t)U_i^2(s,t)/2. \tag{4.25}$$

Since $Q_i(s,t)$ is not observable, a plug-in estimate is derived by using a Bayesian formulation. Let p denote the prior probability that $Q_i(s,t)=1$. Then the posterior mean of $Q_i(s,t)$, after maximizing with respect to the unknown parameters, is

$$\widehat{Q}_i(s,t) = \frac{e^{U_i^2(s,t)/2}}{r_p + e^{U_i^2(s,t)/2}}, \tag{4.26}$$

where $r_p = (1-p)/p$. Replacing Q_i by $\widehat{Q}_i$ in (4.25) and standardizing leads to the weighted sum of chi-squares statistic

$$Z^{(p)}(s,t) = \frac{\sum_{i=1}^N [w(U_i(s,t))U_i^2(s,t) - \mu_p]}{\sigma_p \sqrt{N}}, \tag{4.27}$$

where $w(u) = e^{u^2/2}/\{r_p + e^{u^2/2}\}$ and μ_p, σ_p^2 are the mean and variance, respectively, of $w(U)U^2$ when U is a standard normal random variable.

An approximation of the significance of scans using either (4.24) or (4.27) can be obtained via a last-exit time approach. Instead of the process $Z_{s,t}^{(p)}$, we consider more generally

$$Z_{s,t}^f = \frac{\sum_{i=1}^N [f(U_i(s,t)) - \mu]}{\sigma\sqrt{N}},$$

where f is a well-behaved function, $\mu = Ef(U)$ and $\sigma^2 = \mathrm{Var}(f(U))$. Under the assumption that the noise is independent between samples, $Z_{s,t}^f$ is a normalized sum of N i.i.d. processes, and thus for large N is approximately a mean zero Gaussian process on the two-dimensional indexing set $D = \{(s,t) : 0 \le s < t \le J, v_0 \le t - s \le v_1\}$ with covariance function

$$\rho(s,t,u,v) = \mathrm{Cov}(Z_{s,t}^f, Z_{u,v}^f) = \sigma^{-2}\mathrm{Cov}(f(U_1(s,t)), f(U_1(u,v))). \tag{4.28}$$

The function ρ is not differentiable, but its left and right partial derivatives exist and have the same magnitude. Hence, we may define

$$\rho'(s,t) = \lim_{a \uparrow 0} \left| \frac{\rho(s,t,s+a,t) - \rho(s,t,s,t)}{a} \right|. \tag{4.29}$$

By conditioning on the last-exit time, it follows from the calculations in Siegmund (1988) that

$$P\left\{ \max_{(s,t)\in D} Z_{s,t}^f > c \right\} \approx \frac{\varphi(c)}{c} \sum_{(s,t)\in D} \int_0^\infty e^{-x} P\left\{ \max_{n\ge 1} W_n^{(s,t)} \le -x \right\}$$

$$\times P\left\{ \min_{n\ge 0} W_n^{(s,t)} + \min_{n\ge 1} \widetilde{W}_n^{(s,t)} \ge x \right\} dx, \tag{4.30}$$

where $W_n^{(s,t)}$ is a random walk of i.i.d. normal random variables with mean $-c^2\rho'(s,t)$ and variance $2c^2\rho'(s,t)$, and $\widetilde{W}_n^{(s,t)}$ is an identically distributed random walk, independent of the first random walk. The formula in (4.30) uses a Gaussian approximation on $Z_{s,t}^f$, which is asymptotically a function of the chi-square random variables.

A more accurate approximation can be obtained by correcting for the skewness of $f(U)$. Let $\psi(\theta) = \log \exp\{\theta[f(U) - \mu]/\sigma\}$ and select θ to be the positive root of the equation $N^{1/2}\psi'(\theta) = c$. Replace $\varphi(c)/c$ in (4.30) with the saddlepoint approximation $[2\pi\psi''(\theta)]^{-1/2} \exp\{-N[\theta\psi'(\theta) - \psi(\theta)]\}$ and use Lemma 21 of Siegmund (1992) to evaluate the integral to obtain

$$P\left\{\max_{(s,t)\in D} Z_{s,t}^f > c\right\} \approx [2\pi\psi''(\theta)]^{-1/2} e^{-N[\theta\psi'(\theta)-\psi(\theta)]} c^3 \times \sum_{(s,t)\in D} [\rho'(s,t)]^2 \nu^2 \left(c_0[2\rho_1'(s,t)]^{1/2}\right), \quad (4.31)$$

where $c_0 = c/\sqrt{N}$ and ν is the overshoot constant given in (4.6).

The computation of the partial derivatives ρ' can be simplified by using the expression

$$\rho'(s,t) = (2\sigma^2)^{-1}\{E[f(U_{s,t})f'(U_{s,t})U_{s,t}] - E[f(U_{s,t})f''(U_{s,t})]\}\kappa(t-s), \quad (4.32)$$

where $\kappa(r) = [r(1 - r/J)]^{-1}$. For example, $f(x) = x$ corresponds to the simple one sample case and by (4.32), $\rho'(s,t) = \kappa(t-s)/2$. Substituting this in (4.31) provides us with the significance level approximation of Siegmund (1992). The sum of chi-squares statistic (4.24) corresponds to $f(x) = x^2$ and $\rho'(s,t) = \kappa(t-s)$.

4.5 Concluding Remarks

In addition to DNA copy number, scan statistics can be applied on many other types of genomic profiling data. Recent technological advancements have allowed the measurement of many types of genomic activity, all of which produce enormous quantities of data, where the primary goal is to locate regions of change from baseline in a linear sequence. There is a common theme of scanning for signals of unknown width and scanning for simultaneous signals in multiple sequences. Hoh and Ott (2000), Ji and Wong (2005) and Keles *et al.* (2006) are recent articles that apply scan statistics on the DNA genome. These advancements and advancements in other applied fields like neuroscience have resulted in the collection of a huge amount of data, and scan statistics have been useful in identifying meaningful signals and patterns.

In more traditional areas of scan statistics applications, for example in astronomy and epidemiological studies, there are still many important issues that can occupy the attention and time of researchers. Current scan statistics are geared towards the detection of one cluster of a predetermined shape. It will be interesting to study how scan statistics can be modified so that they can detect multiple clusters or signals with irregular shapes more effectively.

References

1. Adler, R.J. (1981). *The Geometry of Random Fields*, Wiley, New York.

2. Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*, Springer-Verlag, New York.

3. Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: The Chen-Stein method, *Annals of Probability*, **17**, 9–25.

4. Bickel, P. and Rosenblatt, M. (1973). Two-dimensional random fields, In *Multivariate Analysis–III* (Ed., P.K. Krishnaiah), pp. 3–15, Academic Press, New York.

5. Chan, H.P. and Lai, T.L. (2002). Boundary crossing probabilities for scan statistics and their applications to change-point detection, *Methodology and Computing in Applied Probability*, **4**, 317–336.

6. Chan, H.P. and Lai, T.L. (2003). Saddlepoint approximations and nonlinear boundary crossing probabilities of Markov random walks, *Annals of Applied Probability*, **13**, 395–429.

7. Chan, H.P. and Loh, W.L. (2007). Some theoretical results on neural spike train probability models, *Annals of Statistics*, **35**, 2691–2722.

8. Chan, H.P. and Zhang, N.R. (2007) Scan statistics with weighted observations, *Journal of the American Statistical Association*, **102**, 595–602.

9. Chew, D., Choi, K. and Leung, M. (2005). Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpes viruses, *Nucleic Acids Research*, **33**, e134.

10. Chi, Z. (2004). Large deviations for template matching between point processes, *Annals of Applied Probability*, **15**, 153–174.

11. Chi, Z., Rauske, P.L. and Margoliasch, D. (2003). Pattern filtering for detection of neural activity, with example from HVc activity during sleep in zebra finches, *Neural Computing*, **15**, 2307–2337.

12. Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations, *Journal of the Royal Statistical Society, Series B*, **52**, 73–104.

13. Dave, A.S. and Margoliasch, D. (2000). Song replay during sleep and computational rules for sensorimotor vocal learning, *Science*, **290**, 812–816.

14. Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society, Series A*, **153**, 349–362.

15. Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic, *Journal of the American Statistical Association*, **84**, 560–566.

16. Glaz, J., Naus, J., Roos, M. and Wallenstein, S. (1994). Poisson approximations for the distribution and moments of ordered m-spacings. *Journal of Applied Probability*, **31**, 271–281.

17. Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer-Verlag, New York.

18. Hogan, M.L. and Siegmund, D. (1986). Large deviations for the maxima of some random fields, *Advances in Applied Mathematics*, **7**, 2–22.

19. Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes, *Proceedings of the National Academy of Sciences*, **97**, 9615–9617.

20. Huntington, R. and Naus, J. (1975). A simple expression for kth nearest neighbor coincidence probabilities, *Annals of Probability*, **3**, 894–896.

21. Ji, H. and Wong, W.H. (2005). TileMap: create chromosomal map of tiling array hybridizations, *Bioinformatics*, **21**, 3629–3636.

22. Karlin, S. and Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis, *Science*, **257**, 39–49.

23. Keles, S., van der Laan, M., Dudoit, S. and Cawley, S.E. (2006). Multiple testing methods for ChIP-Chip high density oligonucleotide array data, *Journal of Computational Biology*, **13**, 579–613.

24. Knowles, M. and Siegmund, D. (1989). On Hotelling's approach to testing for a nonlinear parameter in regression, *International Statistical Review*, **57**, 205–220.

25. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

26. Kulldorff, M. (2006). *SaTScan User Guide*, http://www.satscan.org/ techdoc.html.

27. Lai, T.L. and Siegmund, D. (1977, 1979). A nonlinear renewal theorem with applications to sequential analysis I, *Annals of Statistics*, **5**, 946–955, II, *Annals of Statistics*, **7**, 60–76.

28. Lai, W.R., Johnson, M.D., Kucherlapati, R. and Park, P.J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data, *Bioinformatics*, **21**, 3763–3770.

29. Leung, M.Y. and Yamashita, T.E. (1999). Applications of the scan statistic in DNA sequence analysis, In *Scan Statistics and Applications*. (Ed., J. Glaz and N. Balakrishnan), pp. 269–286, Birkhäuser, Boston, MA.

30. Lifanov, A., Makeev, V., Nazina, A. and Papatsenko, D. (2003). Homotypic regulatory clusters in *Drosophila*, *Genome Research*, **13**, 579–588.

31. Loader, C. (1991). Large-deviation approximations to the distribution of scan statistics, *Advances in Applied Probability*, **23**, 751–771.

32. Masse, M.J.O., Karlin, S., Schachtel, G.A. and Mocarski, E.S. (1992). Human cytomegalovirus origin of DNA replication (oriLyt) residues with a highly complex repetitive region, *Proceedings of the National Academy of Science*, **89**, 5246–5250.

33. Naus, J. (1965). Clustering of random points in two dimensions, *Biometrika*, **52**, 263–267.

34. Naus, J. (1966). Some probabilities, expectations, and variances for the size of largest clusters and smallest intervals, *Journal of the American Statistical Association*, **61**, 1191–1199.

35. Naus, J. (1982). Applications for distributions of scan statistics, *Journal of the American Statistical Association*, **77**, 177–183.

36. Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, **5**, 557–572.

37. Pickands, J. (1969). Upcrossing probabilities for stationary Gaussian processes, *Transactions of the American Mathematical Society*, **145**, 51–73.

38. Pinkel, D. and Albertson, D.G. (2005). Array comparative genomic hybridization and its applications in cancer, *Nature Genetics*, **37**, Suppl 11–17.

39. Qualls, C. and Watanabe, H. (1973). Asymptotic properties of Gaussian random fields, *Transactions of the American Mathematical Society*, **177**, 155–171.

40. Rabinowitz, D. (1994). Detecting clusters in disease incidence, In *Change-points Problems* (Ed., E. G. Carlstein, H.-G. Müller and D. Siegmund), 255–275, IMS, Hayward, CA.

41. Rabinowitz, D. and Siegmund, D. (1997). The approximate distribution of the maximum of a smoothed Poisson random field, *Statistica Sinica*, **7**, 167–180.

42. Rajewsky, N., Vergassola, M., Gaul, U. and Siggia, E. (2002). Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo, *BMC Bioinformatics*, **3**, e30.

43. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H. *et al.* (2006). Global variation in copy number in the human genome, *Nature*, **444**, 444–454.

44. Shah, S.P., Lam, W.L., Ng, R.T. and Murphy, K.P. (2007). Modeling recurrent DNA copy number alterations in array CGH data, *Bioinformatics*, **23**, 450–458.

45. Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, Springer-Verlag, New York.

46. Siegmund, D. (1986). Boundary crossing probabilities and statistical applications, *Annals of Statistics*, **14**, 361–404.

47. Siegmund, D. (1988). Tail probabilities for the maxima of some random fields, *Annals of Probability*, **16**, 487–501.

48. Siegmund, D. (1992). Tail approximations for maxima of random fields, In *Probability Theory: Proceedings of the 1989 Singapore Probability Conference* (Ed., L.H.Y. Chen, K.P. Choi, K. Hu and J.-H. Lou), pp. 147–158, Walter de Gruyter, Berlin.

49. Siegmund, D. and Venkatraman, E.S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point, *Annals of Statistics*, **23**, 255–271.

50. Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics, *Bernoulli*, **6**, 191–213.

51. Storey, J.D. and Siegmund, D. (2001). Approximate p-values for local sequence alignments: Numerical studies. *Journal of Computational Biology*, **8**, 549–556.

52. Tu, I. (2009). Asymptotic overshoots for arithmetic i.i.d. random variables, to appear in *Statistica Sinica*, **19**, 315–323.

53. Tu, I. and Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis, *Advances in Applied Probability*, **31**, 510–531.

54. Whittemore, A.S., Friend, N., Brown, B. and Holly, E. (1987). A test to detect clusters of diseases, *Biometrika*, **74**, 631–635.

55. Woodroofe, M. (1978). Large deviations of likelihood ratio statistics with applications to sequential testing, *Annals of Statistics*, **6**, 72–84.

56. Woodroofe, M. (1979). Repeated likelihood ratio tests, *Biometrika*, **66**, 454–463.

57. Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*, SIAM, Philadelphia, PA.

58. Zhang, N.R., Siegmund, D., Ji, H. and Li, J. (2008). Detecting simultaneous change-points in multiple sequences. Technical Report, Department of Statistics, Stanford University, Palo Alto, CA.

5

Approximations for Two-Dimensional Variable Window Scan Statistics

Jie Chen[1] **and Joseph Glaz**[2]

[1] *Department of Mathematics, University of Massachusetts, Boston, MA, USA*
[2] *Department of Statistics, University of Connecticut, Storrs, CT, USA*

Abstract: In this chapter, approximations for distributions of a two-dimensional maximum scan score-type statistic and a minimum p-value scan statistic are derived for independent and identically distributed binomial and Poisson observations. Both unconditional and conditional models are considered. For the conditional models, it is assumed that the total number of observations in the region is known. Numerical results are presented to evaluate the accuracy of the specified probability of Type I error and to compare the power of these variable window-type scan statistics with fixed single window scan statistics.

Keywords and phrases: Cluster detection, maximum scan score-type statistic, minimum p-value statistic, variable window

5.1 Introduction

In this chapter, we review the methods and approximations investigated in Glaz and Zhang (2004) and (2006) for two-dimensional variable window scan statistics. New approximations and algorithms are presented as well. The focus in this article is on approximations for scan statistics based on observed data in a rectangular region via approximations that have been derived for fixed window scan statistics [Glaz, Naus and Wallenstein (2001) and Chen and Glaz (2002)]. Generalized likelihood ratio tests for variable window scan statistics, implemented via simulations, have been extensively investigated in the statistical literature, including: Duczmal and Buckeridge (2006), Kulldorff (1997, 2006), Kulldorff and Nagarwalla (1995), Kulldorff, Tango and Park (2003), Song and

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_5,

Kulldorff (2003), Modarres and Patil (2007), Neill and Moore (2004, 2006), Neill and Lingwall (2007), Patil and Taillie (2004), Tango (2007) and Tango and Takahashi (2005). In this chapter, we will not discuss this important approach for implementing variable window scan statistics. Several chapters in this volume will present recent developments using the likelihood ratio approach. Recently, methods of false discovery control have been employed for variable window scan statistics in two dimensions [Perone-Pacifico, Genovese, Verdinelli, Wasserman (2004, 2007)]. We will not discuss these methods here either. A review chapter on this topic is included in this volume.

The chapter is organized as follows. In Section 5.2, we present a new approximation for a two-dimensional scan statistic for independent and identically distributed (i.i.d.) observations modeled by a binomial or a Poisson distribution that has been discussed for the Poisson model in Guerriero, Willett and Glaz (2009). In Section 5.3, we investigate two variable window scan statistics, for the unconditional case and for the conditional case when the total number of observed events in the region is known. Algorithms and numerical results for evaluating the performance of these variable window scan statistics as discussed in Section 5.3 are given in Section 5.4. A brief summary and directions for future research are presented in Section 5.5.

5.2 Two-Dimensional Discrete Scan Statistics

Let $[0, T_1] \times [0, T_2]$ be a given rectangular region. Let $h_i = T_i/n_i > 0$, where n_i are positive integers, $i = 1, 2$. In many applications the exact locations of the observed events in the region are unknown. What is usually available are the counts in small rectangular subregions. For $1 \le i \le n_1$ and $1 \le j \le n_2$, let X_{ij} be the number of events that have been observed in the rectangular subregion $[(i-1)h_1, ih_1] \times [(j-1)h_2, jh_2]$. We are interested in detecting unusual clustering of these events under the null hypothesis that X_{ij} are i.i.d. nonnegative integer-valued random variables from a specified distribution. For $2 \le m_1 \le n_1 - 1$, $2 \le m_2 \le n_2 - 1$, $1 \le i_1 \le n_1 - m_1 + 1$ and $1 \le i_2 \le n_2 - m_2 + 1$, define

$$Y_{i_1,i_2} = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} X_{ij} \tag{5.1}$$

to be the number of events in a rectangular region comprising m_1 by m_2 adjacent rectangular subregions with area h_1h_2 and the southwest corner located at the point $((i_1 - 1)h_1, (i_2 - 1)h_2)$. If Y_{i_1,i_2} exceeds a preassigned value of k, we will say that k events are clustered within the inspected region. A *two-dimensional discrete scan statistic* [Chen and Glaz (1999)] is defined as the largest number

of events in any m_1 by m_2 adjacent rectangular subregions with area h_1h_2 and the southwest corner located at the point $((i_1-1)h_1,(i_2-1)h_2)$:

$$S_{m_1,m_2}(N_1,N_2)=\max\{Y_{i_1,i_2};1\le i_1\le N_1-m_1+1,1\le i_2\le N_2-m_2+1\}. \tag{5.2}$$

When the size of the rectangular region is fixed throughout the presentation of the results, we abbreviate $S_{m_1,m_2}(N_1,N_2)$ to S_{m_1,m_2}. S_{m_1,m_2} can be viewed as an extension of the one-dimensional discrete scan statistic discussed in Glaz, Naus and Wallenstein (2001, Chapter 13). It has been used in testing the null hypothesis, H_0, of randomness that assumes the X_{ij}'s are i.i.d. binomial random variables with parameters L and $0<p<1$ or i.i.d. Poisson random variables with mean $\mu>0$, respectively. For the alternative hypothesis, H_1, of clustering one often specifies a rectangular subregion

$$R(i_1,i_2)=[(i_1-1)h_1,(i_1+m_1-1)h_1]\times[(i_2-1)h_2,(i_2+m_2-1)h_2]$$

such that for any $i_1\le i\le i_1+m_1-1$ and $i_2\le j\le i_2+m_2-1$, X_{ij} has a binomial distribution with parameters L and p_1, where $p_1>p$ or a Poisson distribution with mean μ_1 where $\mu_1>\mu$, respectively. For $ij\notin[i_1,i_1+m_1-1]\times[i_2,i_2+m_2-1]$, X_{ij} is distributed according to the distribution specified by the null hypothesis. It is well known that the generalized likelihood ratio test rejects the null hypothesis in favor of the alternative hypothesis whenever S_{m_1,m_2} exceeds the value k, where k is determined from a specified significance level of the testing procedure. Approximations for $P(S_{m_1,m_2}\ge k)$ are discussed in Glaz, Naus and Wallenstein (2001, Chapter 16.1).

The use of S_{m_1,m_2} for testing the null hypothesis of randomness specified above is of interest in many areas of science and technology, including astronomy [Darling and Waterman (1986)], computer science [Pfaltz (1983)], ecology [Cressie (1991) and Koen (1991)], epidemiology [Cressie (1991) and Kulldorff (1997)], image analysis [Rosenfeld (1978)], pattern recognition [Panayirci and Dubes (1983)], minefield detection via remote sensing [Glaz (1996)] and reliability theory [Barbour, Chryssaphinou and Roos (1996), Boutsikas and Koutras (2000), Fu and Koutras (1994), Koutras, Papadopoulus and Papastavridis (1993), Malinowski and Preuss (1995), and Salvia and Lasher (1990)].

In this chapter, we consider testing the null hypothesis specified above against the following clustering type alternatives, in the special case when $N_1=N_2=N$ and $m_1=m_2=m$. Let $1\le i_0,j_0\le N-m+1$ and $2\le m\le N/4$ be unknown parameters. We assume that for $i_0\le i\le i_0+m-1$ and $j_0\le j\le j_0+m-1$, $X_{i,j}$ are i.i.d. binomial random variables with parameters L and p_1, $0<p<p_1<1$, or i.i.d. Poisson random variables with mean μ_1, $\mu_0<\mu_1$, respectively, while in the rest of the region $X_{i,j}$ are i.i.d. binomial random variables with parameters L and $0<p<1$ or i.i.d. Poisson random variables with mean μ_0, respectively. Since the size of the rectangular window

m is unknown, we propose in Section 5.3 two variable window scan statistics based on a sequence of fixed window size scan statistics: $S_{m_1\times m_1}, ..., S_{m_n\times m_n}$, where $2 \le m_j < m_{j+1} \le N/4$, $1 \le j \le n-1$.

In retrospective investigations, the total number of observed events in the entire region

$$\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} X_{ij} = a$$

is known. For a fixed rectangular scanning window of size m, a *conditional* two-dimensional discrete scan statistic and its upper tail probabilities are denoted by

$$S_{m_1,m_2}(N_1, N_2; a) \tag{5.3}$$

and

$$P(S_{m_1,m_2}(N_1, N_2; a) \ge k) = P\left\{ S_{m_1,m_2}(N_1, N_2) \ge k | \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} X_{ij} = a \right\}, \tag{5.4}$$

respectively. When the size of the inspected region is clearly understood, we abbreviate $S_{m_1,m_1}(N_1, N_2; a)$ to $S_{m_1,m_2}(a)$.

Let $X_{i,j}$ be i.i.d. binomial random variables with parameters L and $0 < p < 1$ or i.i.d. Poisson random variables with mean $\mu > 0$, respectively. In this chapter, to simplify the presentation of the results, for the unconditional and conditional models, we will assume that $N_1 = N_2 = N$ and $m_1 = m_2 = m$.

In the conditional model, under the null hypothesis of an i.i.d. binomial model with parameters L and $0 < p < 1$, the conditional distribution of $\{X_{ij}, 1 \le i, j \le N\}$ given that $\sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij} = a$, is multivariate hypergeometric with a discrete density function given by

$$P\left\{ X_{ij} = x_{ij}, 1 \le i, j \le N | \sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij} = a \right\} = \frac{\left[\prod_{i=1}^{N}\prod_{j=1}^{N}\binom{L}{x_{ij}}\right]}{\binom{N^2L}{a}}, \tag{5.5}$$

where $\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} = a$ and $0 \le x_{ij} \le L, 1 \le i, j \le N$.

Under the null hypothesis of an i.i.d. Poisson model with mean μ, the conditional distribution of $\{X_{ij}, 1 \le i, j \le N\}$, given that $\sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij} = a$, is multinomial with a discrete density function given by

$$P\left\{ X_{ij} = x_{ij}, 1 \le i, j \le N | \sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij} = a \right\} = \binom{a}{x_{12},....,x_{NN}} \left(\frac{1}{N^2}\right)^a, \tag{5.6}$$

where $\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} = a$ and $0 \le x_{ij}, 1 \le i, j \le N$. Approximations for the distribution and moments of the conditional scan statistic $S_{m_1,m_1}(N_1, N_1; a)$ for binomial and Poisson models are discussed in Glaz, Naus and Wallenstein (2001, Chapter 16.2) and Chen and Glaz (2002).

In this chapter, we consider testing the null hypothesis specified above against clustering-type alternatives described below. One class of alternative hypotheses has the following representation. Let $1 \le i_0, j_0 \le N - m + 1$ and $2 \le m \le N/4$ be unknown parameters. We assume that for $i_0 \le i \le i_0 + m - 1$ and $j_0 \le j \le j_0 + m - 1$, $X_{i,j}$ are i.i.d. binomial random variables with parameters L and p_1, $0 < p < p_1 < 1$, or i.i.d. Poisson random variables with mean μ_1, $\mu_0 < \mu_1$, respectively, and a_1 events have been observed, while in the rest of the region $X_{i,j}$ are i.i.d. binomial random variables with parameters L and $0 < p < 1$ or i.i.d. Poisson random variables with mean μ_0, respectively, and $a - a_1$ events have been observed, where a_1 is large in comparison to $a - a_1$. Therefore, the alternative hypotheses in the binomial or Poisson models comprise a class of a product of two multivariate hypergeometric distributions or a product of two multinomial distributions, respectively, with parameters i_0, j_0, m, a, and a_1. More general classes of alternative hypotheses can be considered as well. One can divide the locations $\{1 \le i, j \le N\} \backslash \{i_0 \le i \le i_0 + m - 1\} \times \{j_0 \le j \le j_0 + m - 1\}$ into M disjoint rectangular location regions and assume that in each of these regions $X_{i,j}$ are i.i.d. binomial random variables with parameters L and $0 < p_{0j} < 1$ or i.i.d. Poisson random variables with mean μ_{0j}, $1 \le j \le M$. Moreover, one assumes that a_{0j} events have been observed in the jth region, $1 \le j \le M$, where $a_1 + \sum_{j=1}^{m} a_{0j} = a$ and a_1 is large in comparison to a_{0j}. Therefore, the alternative hypotheses here are a product of several multivariate hypergeometric distributions or a product of several multinomial distributions, respectively, with parameters i_0, j_0, m, a, a_1, and $a_{0j}, 1 \le j \le M$.

Since the size of the rectangular window m is unknown, in Section 5.3 we propose two variable window scan statistics based on a sequence of fixed window size scan statistics: $S_{m_1 \times m_1}(a), ..., S_{m_n \times m_n}(a)$, where $2 \le m_j < m_{j+1} \le N/4$, $1 \le j \le n - 1$.

For the unconditional case, we present below a new approximation for $P(S_{m,m} \ge k)$. Note that the general formula is valid for any i.i.d. nonnegative integer-valued observations. For the special case of i.i.d. Bernoulli observations, this approximation has the same structure as an approximation in Glaz, Naus and Wallenstein (2001, Equation 16.40) and is a member of a general class of approximations in Boutsikas and Koutras (2000, Equation 16.33).

We first derive an approximation for $P(S_{m,m} \ge k)$ for i.i.d. binomial observations with parameters L and $0 < p < 1$. An approach in Glaz, Naus and Wallenstein (2001, Section 16.1.6) leading to an approximation in Glaz, Naus and Wallenstein (2001, Equation 16.40) yields for $1 \le k < m^2 L$:

$$P(S_{m,m} \ge k) \approx 1 - \frac{[P(S_{m,m}(m+1, m+1) \le k-1)]^{(N-m)^2}}{[P(S_{m,m}(m, m+1) \le k-1)]^{(N-m-1)(N-m)}} \times \frac{(q_{m,2m-1})^{(N-2m)(N-m-1)}}{(q_{m,2m})^{(N-2m+1)(N-m-1)}}, \quad (5.7)$$

where

$$q_{m,m+l-1} = P(A_{1,1}^c \bigcap A_{1,2}^c \cdots \bigcap A_{1,l}^c), \tag{5.8}$$

$1 \le l \le N - m + 1$, and

$$A_{i_1,i_2} = \sum_{i=i_1}^{i_1+m-1} \sum_{j=i_2}^{i_2+m-1} X_{i,j} \ge k, \tag{5.9}$$

$1 \le i_1, i_2 \le N - m + 1$. The quantities $q_{m,2m-1}$ and $q_{m,2m}$ are evaluated via an algorithm of Saperstein (1976) extended in Glaz and Naus (1991) or Karwe and Naus (1997).

To evaluate $P(S_{m,m}(m, m+1) \le k-1)$ and $P(S_{m,m}(m+1, m+1) \le k-1)$ one employs a conditioning approach. Let $\{X_{i,j}; 1 \le i \le m+1, 1 \le j \le m+1\}$ be i.i.d. binomial random variables with parameters L and p. Let $Y_{i_1,j_1}^{i_2 j_2} = \sum_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} X_{i,j}$. To evaluate $P(S_{m,m}(m, m+1) \le k-1)$, we condition on $Y_{1,2}^{m,m} = y$, to get

$$\begin{aligned} &P(S_{m,m}(m, m+1) \le k-1) \\ &= \sum_{y=0}^{(k-1)(m-1)mL} P(S_{m,m}(m, m+1) \le k-1 | Y_{1,2}^{m,m} = y) P(Y_{1,2}^{m,m} = y). \end{aligned} \tag{5.10}$$

Now,

$$\begin{aligned} &P(S_{m,m}(m, m+1) \le k-1 | Y_{1,2}^{m,m} = y) \\ &= P(Y_{1,1}^{m,m} \le k-1, Y_{1,2}^{m,m+1} \le k-1 | Y_{1,2}^{m,m} = y) \\ &= P(Y_{1,1}^{m,1} \le k-1-y, Y_{1,m+1}^{m,m+1} \le k-1-y) = \left[P(Y_{1,1}^{m,1} \le k-1-y)\right]^2. \end{aligned}$$

Therefore,

$$\begin{aligned} &P(S_{m,m}(m, m+1) \le k-1) \\ &= \sum_{y=0}^{(k-1)\wedge(m-1)mL} \left[P(Y_{1,1}^{m,1} \le k-1-y)\right]^2 P(Y_{1,2}^{m,m} = y), \end{aligned}$$

where $Y_{1,1}^{m,1}$ is a binomial random variable with parameters mL and p and $Y_{1,2}^{m,m}$ is a binomial random variable with parameters $(m-1)mL$ and p. We can express the event

$$\begin{aligned} &(S_{m,m}(m+1, m+1) \le k-1) \\ &= (Y_{1,1}^{m,m} \le k-1, Y_{1,2}^{m,m+1} \le k-1, Y_{2,1}^{m+1,m} \le k-1, Y_{2,2}^{m+1,m+1} \le k-1). \end{aligned} \tag{5.11}$$

Also,

$$\begin{aligned} Y_{1,1}^{m,m} &= Y_{2,2}^{m,m} + Y_{1,1}^{1,1} + Y_{2,1}^{m,1} + Y_{1,2}^{1,m}, \\ Y_{1,2}^{m,m+1} &= Y_{2,2}^{m,m} + Y_{1,m+1}^{1,m+1} + Y_{1,2}^{1,m} + Y_{2,m+1}^{m,m+1}, \\ Y_{2,1}^{m+1,m} &= Y_{2,2}^{m,m} + Y_{m+1,1}^{m+1,1} + Y_{m+1,2}^{m+1,m} + Y_{2,1}^{m,1}, \end{aligned}$$

and

$$Y_{2,2}^{m+1,m+1} = Y_{2,2}^{m,m} + Y_{m+1,m+1}^{m+1,m+1} + Y_{m+1,2}^{m+1,m} + Y_{2,m+1}^{m,m+1}.$$

Therefore, to evaluate $P(S_{m,m}(m+1,m+1) \le k-1)$ we need to condition on $Y_{2,2}^{m,m}$, $Y_{2,1}^{m,1}$, $Y_{1,2}^{1,m}$, $Y_{m+1,2}^{m+1,m}$, and $Y_{2,m+1}^{m,m+1}$. The condition is done in the following order: $Y_{2,2}^{m,m} = y_1, Y_{1,2}^{1,m} = y_2, Y_{m+1,2}^{m+1,m} = y_3, Y_{2,1}^{m,1} = y_4, Y_{2,m+1}^{m,m+1} = y_5$. We get

$$\begin{aligned} P(S_{m,m}(m+1,m+1) \le k-1) = & \sum_{y_1=0}^{(k-1)\wedge(m-1)^2 L} \sum_{y_2=0}^{(k-1-y_1)\wedge(m-1)L} \\ & \sum_{y_3=0}^{(k-1-y_1)\wedge(m-1)L} \sum_{y_4=0}^{(k-1-y_1-(y_2\vee y_3))\wedge(m-1)L} \sum_{y_5=0}^{(k-1-y_1-(y_2\vee y_3))\wedge(m-1)L} \\ & \Big[P(S_{m,m}(m+1,m+1) \le k-1 | Y_{2,2}^{m,m} = y_1, Y_{1,2}^{1,m} = y_2, Y_{m+1,2}^{m+1,m} = y_3, \\ & Y_{2,1}^{m,1} = y_4, Y_{2,m+1}^{m,m+1} = y_5)\Big] \times P\left(Y_{2,2}^{m,m} = y_1\right) P\left(Y_{1,2}^{1,m} = y_2\right) \\ & P\left(Y_{m+1,2}^{m+1,m} = y_3\right) P\left(Y_{2,1}^{m,1} = y_4\right) P\left(Y_{2,m+1}^{m,m+1} = y_5\right). \end{aligned} \tag{5.12}$$

Then

$$\begin{aligned} & P\left(S_{m,m}(m+1,m+1) \le k-1 | Y_{2,2}^{m,m} = y_1, Y_{1,2}^{1,m} = y_2, Y_{m+1,2}^{m+1,m} = y_3, \right. \\ & \left. Y_{2,1}^{m,1} = y_4, Y_{2,m+1}^{m,m+1} = y_5\right) \\ = & P\left(Y_{1,1}^{1,1} \le a_1, Y_{1,m+1}^{1,m+1} \le a_2, Y_{m+1,1}^{m+1,1} \le a_3, Y_{m+1,m+1}^{m+1,m+1} \le a_4\right) \\ = & P\left(Y_{1,1}^{1,1} \le a_1\right) P\left(Y_{1,m+1}^{1,m+1} \le a_2\right) P\left(Y_{m+1,1}^{m+1,1} \le a_3\right) P\left(Y_{m+1,m+1}^{m+1,m+1} \le a_4\right), \end{aligned}$$

where

$$a_1 = k-1-y_1-y_2-y_4, \;\; a_2 = k-1-y_1-y_2-y_5$$

and

$$a_3 = k-1-y_1-y_3-y_4, \;\; a_4 = k-1-y_1-y_3-y_5.$$

Therefore,

$$P(S_{m,m}(m+1,m+1)\le k-1)=\sum_{y_1=0}^{(k-1)\wedge(m-1)^2L}\ \sum_{y_2=0}^{(k-1-y_1)\wedge(m-1)L}\ \sum_{y_3=0}^{(k-1-y_1)\wedge(m-1)L}\ \sum_{y_4=0}^{(k-1-y_1-(y_2\vee y_3))\wedge(m-1)L}\ \sum_{y_5=0}^{(k-1-y_1-(y_2\vee y_3))\wedge(m-1)L} \tag{5.13}$$

$$\left[P\left(Y_{1,1}^{1,1}\le a_1\right)P\left(Y_{1,m+1}^{1,m+1}\le a_2\right)P\left(Y_{m+1,1}^{m+1,1}\le a_3\right)P\left(Y_{m+1,m+1}^{m+1,m+1}\le a_4\right)\right.$$

$$\times P\left(Y_{2,2}^{m,m}=y_1\right)P\left(Y_{1,2}^{1,m}=y_2\right)P\left(Y_{m+1,2}^{m+1,m}=y_3\right)P\left(Y_{2,1}^{m,1}=y_4\right)$$

$$\left.P\left(Y_{2,m+1}^{m,m+1}=y_5\right)\right],$$

where $Y_{1,1}^{1,1}, Y_{1,m+1}^{1,m+1}, Y_{m+1,1}^{m+1,1}$ and $Y_{m+1,m+1}^{m+1,m+1}$ are i.i.d. binomial random variables with parameters L and p, $Y_{2,2}^{m,m}$ has a binomial distribution with parameters $(m-1)^2L$ and p, and $Y_{1,2}^{1,m}$, $Y_{m+1,2}^{m+1,m}$, $Y_{2,1}^{m,1}$, $Y_{2,m+1}^{m,m+1}$ are i.i.d. binomial random variables with parameters $(m-1)L$ and p. All nine binomial random variables are independent of each other.

A similar method is used to derive an approximation for the i.i.d. Poisson model with mean μ. Here too the quantities $q_{m,2m-1}$ and $q_{m,2m}$ are evaluated via an algorithm of Saperstein (1976) extended in Glaz and Naus (1991) or Karwe and Naus (1997). For the Poisson model,

$$P(S_{m,m}(m,m+1)\le k-1)=\sum_{y=0}^{k-1}\left[P(Y_{1,1}^{m,1}\le k-1-y)\right]^2P(Y_{1,2}^{m,m}=y), \tag{5.14}$$

where $Y_{1,1}^{m,1}$ is a Poisson random variable with mean equal to $m\mu$ and $Y_{1,2}^{m,m}$ is a Poisson random variable with mean equal to $m(m-1)\mu$ and

$$P(S_{m,m}(m+1,m+1)\le k-1)=\sum_{y_1=0}^{k-1}\ \sum_{y_2=0}^{k-1-y_1}\ \sum_{y_3=0}^{k-1-y_1}\ \sum_{y_4=0}^{k-1-y_1-(y_2\vee y_3)}\ \sum_{y_5=0}^{k-1-y_1-(y_2\vee y_3)} \tag{5.15}$$

$$\left[P\left(Y_{1,1}^{1,1}\le a_1\right)P\left(Y_{1,m+1}^{1,m+1}\le a_2\right)P\left(Y_{m+1,1}^{m+1,1}\le a_3\right)P\left(Y_{m+1,m+1}^{m+1,m+1}\le a_4\right)\right.$$

$$\times P\left(Y_{2,2}^{m,m}=y_1\right)P\left(Y_{1,2}^{1,m}=y_2\right)P\left(Y_{m+1,2}^{m+1,m}=y_3\right)P\left(Y_{2,1}^{m,1}=y_4\right)$$

$$\left.P\left(Y_{2,m+1}^{m,m+1}=y_5\right)\right],$$

where

$$a_1 = k - 1 - y_1 - y_2 - y_4,\ \ a_2 = k - 1 - y_1 - y_2 - y_5,$$

$$a_3 = k - 1 - y_1 - y_3 - y_4,\ \ a_4 = k - 1 - y_1 - y_3 - y_5,$$

and $Y_{1,1}^{1,1}, Y_{1,m+1}^{1,m+1}, Y_{m+1,1}^{m+1,1}$ and $Y_{m+1,m+1}^{m+1,m+1}$ are i.i.d. Poisson random variables with mean μ, $Y_{2,2}^{m,m}$ has a Poisson distribution with mean equal to $(m-1)^2\mu$ and $Y_{1,2}^{1,m}$, $Y_{m+1,2}^{m+1,m}$, $Y_{2,1}^{m,1}$, $Y_{2,m+1}^{m,m+1}$ are i.i.d. Poisson random variables with mean $(m-1)\mu$.

5.3 Variable Window Discrete-Type Scan Statistics

Let $X_{i,j}$ be i.i.d. binomial random variables with parameters L and $0 < p < 1$ or i.i.d. Poisson random variables with mean $\mu > 0$, respectively. We are interested in the case where the window size $2 \leq m \leq N/4$ is unknown. Glaz and Zhang (2004) investigated the performance of multiple scan statistics in a two-dimensional case for the i.i.d. Bernoulli model. The algorithms for implementing the testing procedures based on multiple window scan statistics, which are multivariate statistics, are quite complex. Therefore, we will not discuss them here. We will investigate two variable window scan statistics: the maximum scan score-type scan statistic [Glaz and Zhang (2006)] and the minimum p-value scan statistic [Zhang and Glaz (2008)].

5.3.1 Unconditional case

In the unconditional case, the total number of observed events in the region targeted for inspection is unknown. We first present a two-dimensional maximum scan score-type statistic. In Section 5.4, an algorithm for implementing this scan statistic is discussed. This extends Glaz and Zhang (2006), where a maximum scan score-type statistic is discussed for i.i.d. Bernoulli trials in a one-dimensional case.

Let $2 \leq m_1 < m_2 < ..... < m_n \leq N/4$ be a given sequence of window lengths with associated discrete scan statistics $S_{m_1,m_1}, S_{m_2,m_2}, ..., S_{m_n,m_n}$, respectively. We define a two-dimensional maximum scan score-type statistic

$$T_n = \max_{1 \leq j \leq n} \left\{ \frac{S_{m_j,m_j} - \mu_j}{\sigma_j} \right\}, \tag{5.16}$$

where $\mu_j = E\left(S_{m_j,m_j}\right)$ and $\sigma_j = SD\left(S_{m_j,m_j}\right)$ are the expected value and the standard deviation of S_{m_j,m_j}, respectively. For a given significance level α, we reject the null hypothesis that $X_{i,j}$ are i.i.d. binomial random variables with parameters L and $0 < p < 1$ or i.i.d. Poisson random variables with mean

$\mu > 0$, if $T_n \geq t$, where $\alpha = P(T_n \geq t)$. To obtain the critical value t, an accurate approximation for $P(T_n \geq t)$ is needed. Following the approach in Glaz and Zhang (2006), we get

$$P(T_n \geq t) = P\left\{\bigcup_{j=1}^{n}\left(S_{m_j,m_j} \geq t\sigma_j + \mu_j\right)\right\} = P\left\{\bigcup_{j=1}^{n}\left(S_{m_j,m_j} \geq k_j\right)\right\}, \tag{5.17}$$

where

$$k_j = \begin{cases} [t\sigma_j + \mu_j], & t\sigma_j + \mu_j \text{ is an integer} \\ [t\sigma_j + \mu_j] + 1, & \text{otherwise} \end{cases} \tag{5.18}$$

and $[x]$ denotes the integer part of x. For $n \geq 2$, it follows from Equation (5.17) and Glaz and Zhang (2004) that

$$\begin{aligned} P(T_n \geq t) &\leq \sum_{j=1}^{n} P\left(S_{m_j,m_j} \geq k_j\right) \\ &- \sum_{j=1}^{n-1} P\left\{\left(S_{m_j,m_j} \geq k_j\right) \cap \left(S_{m_{j+1},m_{j+1}} \geq k_{j+1}\right)\right\}, \end{aligned} \tag{5.19}$$

where

$$\begin{aligned} &P\{(S_{m_j,m_j} \geq k_j) \cap (S_{m_{j+1},m_{j+1}} \geq k_{j+1})\} \\ &= P\{S_{m_j,m_j} \geq k_j | S_{m_{j+1},m_{j+1}} \geq k_{j+1}\} P(S_{m_{j+1},m_{j+1}} \geq k_{j+1}). \end{aligned} \tag{5.20}$$

One can approximate

$$\begin{aligned} &P\{S_{m_j,m_j} \geq k_j | S_{m_{j+1},m_{j+1}} \geq k_{j+1}\} \\ &\approx P\{S_{m_j,m_j}(m_{j+1}, m_{j+1}; k_{j+1}) \geq k_j\} + P(S_{m_j,m_j} \geq k_j). \end{aligned} \tag{5.21}$$

Approximation (5.21) is based on the fact that, conditional on $S_{m_{j+1},m_{j+1}} \geq k_{j+1}$, the event $S_{m_j,m_j} \geq k_j$ will either occur within a rectangle of size $m_{j+1} \times m_{j+1}$, where k_{j+1} 1's have been observed, or outside that rectangular region. The latter event is approximated by $P(S_{m_j,m_j} \geq k_j)$. If the event $S_{m_j,m_j} \geq k_j$ occurs within a rectangle of size $m_{j+1} \times m_{j+1}$, where k_{j+1} 1's have been observed, we approximate that probability by $P\{S_{m_j,m_j}(m_{j+1}, m_{j+1}; k_{j+1}) \geq k_j\}$, the tail probability of a two-dimensional conditional scan statistic, defined in Equation (5.4). To implement this two-dimensional scan score-type statistic, we approximate μ_j and σ_j using Chen and Glaz (1999) and the new approximation for the distribution of the unconditional scan statistic given in Section 5.2. $P(S_{m_j,m_j} \geq k_j)$ and $P(S_{m_{j+1},m_{j+1}} \geq k_{j+1})$ are approximated as well via the new approximation in Section 5.2. $P\{S_{m_j,m_j}(m_{j+1}, m_{j+1}; k_{j+1}) \geq k_j\}$ is approximated using Chen and Glaz (2002).

The advantage of T_n over the multiple scan statistic, $(S_{m_1,m_1}, S_{m_2,m_2}, ..., S_{m_n,m_n})$, investigated in Glaz and Zhang (2004), is derived from T_n being a univariate statistic. Therefore, its range can be expressed as a union of disjoint intervals on which $S_{m_j,m_j} = k_j$, $1 \leq j \leq n$, are fixed. This decomposition of the range of T_n leads to a natural ordering of the values of $\{k_j; 1 \leq j \leq n\}$, with respect to $P(T_n \geq t)$, via Equation (5.18). This ordering yields an algorithm for obtaining a unique rejection region for a given significance level α. In Section 5.4, we discuss this algorithm and present numerical results to evaluate the performance of T_n.

Now, let $2 \leq m_1 < m_1 < \cdots < m_n \leq N/4$ be the sizes of n rectangular windows. For $1 \leq j \leq n$, let k_j be the observed value of S_{m_j,m_j} and $p_j = P(S_{m_j,m_j} \geq k_j \mid H_0)$ the associated p-value. To test H_0 vs. H_1 we propose the following test statistic:

$$P_{\min} = \min\{p_j; 1 \leq j \leq k\}, \tag{5.22}$$

to be called the *minimum P-value statistic*. In the context of variable window scan statistics, $P_{\min}$ has been introduced in Hoh and Ott (2000) for a one-dimensional $0-1$ i.i.d. Bernoulli model. It has been extended to the two-dimensional case in Zhang and Glaz (2008) for implementing a variable window Bayesian scan statistic.

5.3.2 Conditional case

Let $2 \leq m_1 < m_2 < \cdots\cdots < m_n \leq N/4$ be a given sequence of window lengths with associated conditional discrete scan statistics $S_{m_1,m_1}(a), S_{m_2,m_2}(a), ..., S_{m_n,m_n}(a)$, respectively. We define a two-dimensional maximum scan score-type statistic

$$T_n(a) = \max_{1\leq j\leq n} \left\{ \frac{S_{m_j,m_j}(a) - \mu_j(a)}{\sigma_j(a)} \right\}, \tag{5.23}$$

where $\mu_j(a) = E\left(S_{m_j,m_j}(a)\right)$ and $\sigma_j(a) = SD\left(S_{m_j,m_j}(a)\right)$ are the expected value and the standard deviation of $S_{m_j,m_j}(a)$, respectively. For a given significance level α, we reject the null hypothesis that $X_{i,j}$ are i.i.d. binomial random variables with parameters L and $0 < p < 1$ or i.i.d. Poisson random variables with mean $\mu > 0$, if $T_n(a) \geq t$, where $\alpha = P(T_n(a) \geq t)$. To obtain the critical value t, an accurate approximation for $P(T_n(a) \geq t)$ is needed. Following the approach in Glaz and Zhang (2006), we get

$$\begin{aligned} P(T_n(a) \geq t) &= P\left\{ \bigcup_{j=1}^{n} \left(S_{m_j,m_j}(a) \geq t\sigma_j(a) + \mu_j(a)\right) \right\} \\ &= P\left\{ \bigcup_{j=1}^{n} \left(S_{m_j,m_j}(a) \geq k_j\right) \right\}, \end{aligned} \tag{5.24}$$

where

$$k_j = \begin{cases} [t\sigma_j(a) + \mu_j(a)], & t\sigma_j(a) + \mu_j(a) \text{ is an integer} \\ [t\sigma_j(a) + \mu_j(a)] + 1, & \text{otherwise} \end{cases} \tag{5.25}$$

and $[x]$ denotes the integer part of x. For $n \geq 2$, it follows from Equation (5.24) and Glaz and Zhang (2004) that

$$\begin{aligned} P(T_n(a) \geq t) &\leq \sum_{j=1}^{n} P\left(S_{m_j,m_j}(a) \geq k_j\right) \\ &- \sum_{j=1}^{n-1} P\left\{\left(S_{m_j,m_j}(a) \geq k_j\right) \cap \left(S_{m_{j+1},m_{j+1}}(a) \geq k_{j+1}\right)\right\}, \end{aligned} \tag{5.26}$$

where

$$\begin{aligned} &P\{(S_{m_j,m_j}(a) \geq k_j) \cap (S_{m_{j+1},m_{j+1}}(a) \geq k_{j+1})\} \\ &= P\{S_{m_j,m_j}(a) \geq k_j | S_{m_{j+1},m_{j+1}}(a) \geq k_{j+1}\} P(S_{m_{j+1},m_{j+1}}(a) \geq k_{j+1}). \end{aligned} \tag{5.27}$$

One can approximate

$$\begin{aligned} &P\{S_{m_j,m_j}(a) \geq k_j | S_{m_{j+1},m_{j+1}}(a) \geq k_{j+1}\} \\ &\approx P\{S_{m_j,m_j}(m_{j+1}, m_{j+1}; k_{j+1}) \geq k_j\} + P(S_{m_j,m_j}(a - k_{j+1}) \geq k_j). \end{aligned} \tag{5.28}$$

Approximation (5.28) is based on the fact that conditional on $S_{m_{j+1},m_{j+1}}(a) \geq k_{j+1}$, the event $S_{m_j,m_j}(a) \geq k_j$ will either occur within a rectangle of size $m_{j+1} \times m_{j+1}$, where k_{j+1} 1's have been observed, or outside that rectangular region. The latter event is approximated by $P(S_{m_j,m_j}(a - k_{j+1}) \geq k_j)$. If the event $S_{m_j,m_j}(a) \geq k_j$ occurs within a rectangle of size $m_{j+1} \times m_{j+1}$, where k_{j+1} 1's have been observed, we approximate that probability by $P\{S_{m_j,m_j}(m_{j+1}, m_{j+1}; k_{j+1}) \geq k_j\}$, the tail probability of a two-dimensional conditional scan statistic, defined in Equation (5.4). To implement this two-dimensional scan score-type statistic, we approximate $P(S_{m_j,m_j}(a) \geq k_j)$, $P(S_{m_j,m_j}(a - k_{j+1}) \geq k_j)$, $P\{S_{m_j,m_j}(m_{j+1}, m_{j+1}; k_{j+1}) \geq k_j\}$, $\mu_j(a)$ and $\sigma_j(a)$ using the methods in Chen and Glaz (2002).

Let $2 \leq m_1 < m_1 < \cdots < m_n \leq N/4$ be the sizes of n rectangular windows. For $1 \leq j \leq n$, let k_j be the observed value of $S_{m_j,m_j}(a)$ and $p_j(a) = P(S_{m_j,m_j}(a) \geq k_j \mid H_0)$ the associated p-value. To test H_0 vs. H_1 we propose the following test statistic:

$$P_{\min}(a) = \min\{p_j(a); 1 \leq j \leq k\}, \tag{5.29}$$

to be called the *conditional minimum P-value statistic*. In Section 5.4 an algorithm for implementing this test statistic is presented along with numerical results to evaluate its performance.

5.4 Numerical Results

5.4.1 Unconditional case

In this section, numerical results are presented to evaluate the accuracy of achieving a specified probability of Type I error for the variable window scan statistics T_n and P_{min}. The probability of Type I error for T_n is evaluated using the new approximations derived in Section 5.2. Since there is no exact distribution available for the P_{min} statistic, for a given significant level α, the critical value p_α given by

$$P_{H_0}(P_{\min} \leq p_\alpha) = \alpha,$$

is evaluated via simulation with 10,000 trials.

For selected values of the parameters, we simulate the power of T_n and P_{min} and compare it with the power of fixed window scan statistics $S_{m_j,m_j}, 1 \leq j \leq n$, for unconditional Bernoulli, binomial and Poisson models. The power of these scan statistics is evaluated via simulation based on 10,000 trials for classes of alternative hypotheses and listed in Tables 5.1–5.5, for the Bernoulli, binomial and Poisson models, respectively.

5.4.2 Conditional case

In this section, the power and the accuracy of the probability of Type I error of $P_{min}(a)$ and $T_n(a)$ are compared with the individual fixed window scan statistics $S_{m_1,m_1}(a), S_{m_2,m_2}(a), \ldots, S_{m_n,m_n}(a)$, respectively. For a specified significance level, the power of $S_{m_j,m_j}(a)$, $T_n(a)$ and $P_{min}(a)$ is evaluated for the following

Table 5.1. Comparison of power for i.i.d. Bernoulli distribution with $p_0 = .001$.

n_1	μ_1	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	0.01	0.0254	0.0378	0.0516	0.0472	0.0355
	0.05	0.1431	0.1585	0.1863	0.1319	0.0636
10	0.01	0.0569	0.0825	0.0821	0.1026	0.0512
	0.05	0.7693	0.8119	0.6511	0.8171	0.6497
20	0.01	0.5587	0.3601	0.1967	0.3752	0.4567
	0.05	1.0000	1.0000	0.9919	0.9995	1.0000
Type I error		0.0597	0.0486	0.0446	0.0556	0.0636

Table 5.2. Comparison of power for i.i.d. Bernoulli distribution with $p_0 = .0025$.

n_1	μ_1	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	0.01	0.0503	0.0493	0.0500	0.0423	0.0301
	0.05	0.1273	0.1050	0.1385	0.0690	0.0373
10	0.01	0.0648	0.0606	0.0654	0.0531	0.0320
	0.05	0.6434	0.6261	0.5123	0.6457	0.3835
20	0.01	0.2112	0.1672	0.1161	0.1440	0.1722
	0.05	0.9998	1.0000	0.9300	0.9950	0.9999
Type I error		0.0348	0.0486	0.0431	0.0480	0.0349

Table 5.3. Comparison of power for i.i.d. Bernoulli distribution with $p_0 = .005$.

n_1	μ_1	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	0.01	0.0577	0.0459	0.0474	0.0390	0.0247
	0.05	0.0924	0.0806	0.0846	0.0481	0.0267
10	0.01	0.0607	0.0488	0.0511	0.0421	0.0234
	0.05	0.4748	0.4634	0.3218	0.4499	0.1582
20	0.01	0.0882	0.0716	0.0623	0.0617	0.0400
	0.05	0.9941	0.9945	0.8345	0.9713	0.9945
Type I error		0.0512	0.0496	0.0596	0.0543	0.0313

Table 5.4. Comparison of power for i.i.d. binomial distribution with $L = 5$ and $p_0 = .001$.

n_1	p_1	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	0.010	0.1005	0.1029	0.0957	0.0506	0.0296
	0.050	0.8844	0.8852	0.8862	0.6934	0.3281
10	0.010	0.4804	0.4703	0.3327	0.4525	0.1605
	0.020	0.9587	0.9537	0.8564	0.9477	0.7466
20	0.005	0.6375	0.6594	0.2690	0.4457	0.6166
	0.008	0.9659	0.9716	0.6550	0.8787	0.9664
	0.010	0.9962	0.9967	0.8409	0.9696	0.9962
Type I error		0.0564	0.0570	0.0529	0.0434	0.0288

Table 5.5. Comparison of power for i.i.d. Poisson distribution with $\mu_0 = .001$.

n_1	μ_1	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	0.10	0.4730	0.4780	0.4774	0.4939	0.3739
	0.20	0.8677	0.8678	0.8691	0.8741	0.7970
10	0.05	0.8190	0.8304	0.8234	0.6523	0.8094
	0.10	0.9942	0.9939	0.9948	0.9688	0.9937
20	0.01	0.4907	0.4744	0.5003	0.2000	0.3686
	0.02	0.9298	0.9147	0.9316	0.5577	0.8174
Type I error		0.0493	0.0554	0.0542	0.0505	0.0460

Table 5.6. Comparison of power for $a = 10$ for i.i.d. Bernoulli model.

n_1	a_1	a_2	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	2	3	0.2202	0.2180	0.1994	0.1950	0.1061
	2	5	0.4142	0.4179	0.2652	0.3668	0.2516
10	2	5	0.3161	0.3142	0.1570	0.2829	0.2174
	3	1	0.8016	0.8010	0.3119	0.7222	0.1678
	3	5	0.8370	0.8452	0.3447	0.7441	0.4286
20	4	1	0.4446	0.5557	0.1580	0.4044	0.4773
	4	5	0.5825	0.6839	0.1760	0.4681	0.5819
	5	1	1.0000	1.0000	0.2374	0.6258	1.0000
25	4	5	0.3957	0.3967	0.1406	0.3330	0.4836
	5	1	0.7004	0.7027	0.1741	0.4477	0.7622
	5	3	0.7138	0.7201	0.1745	0.4592	0.7788
Type I error			0.0504	0.0522	0.0500	0.0478	0.525

classes of alternative hypothesis that lead to a larger number of occurrences of events in the $n_1 \times n_1$ and $n_1 \times N - n_1$ rectangular subregions:

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} X_{i,j} = a_1, \sum_{i=1}^{n_1}\sum_{j=1}^{N-n_1} X_{i,j} = a_2 \quad \text{and} \quad \sum_{i=1}^{N-n_1}\sum_{j=1}^{N} X_{i,j} = a - a_1 - a_2,$$

where $n_1 < N$ and $a_1 + a_2 \leq a$. The selected values of a_1, a_2 and n_1 are listed in Tables 5.6–5.11 for the Bernoulli, binomial and Poisson models, respectively.

Table 5.7. Comparison of power for $a = 25$ for i.i.d. Bernoulli model.

n_1	a_1	a_2	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	2	10	0.3940	0.4109	0.4033	0.3339	0.2574
	2	15	0.7790	0.7850	0.7106	0.7003	0.6639
10	4	5	0.4118	0.4229	0.4069	0.5783	0.1789
	4	10	0.5976	0.6050	0.4812	0.6499	0.4251
	4	15	0.8730	0.8752	0.6400	0.8025	0.7875
25	5	5	0.0752	0.0738	0.0733	0.0799	0.0642
	8	10	0.7996	0.7981	0.2569	0.4293	0.8401
	8	15	0.9184	0.9174	0.3555	0.5570	0.9306
Type I error			0.0475	0.0499	0.0434	0.0435	0.0517

Table 5.8. Comparison of power for $a = 50$ for i.i.d. Bernoulli model.

n_1	a_1	a_2	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	3	5	0.1092	0.1067	0.1500	0.0737	0.0365
	3	10	0.2344	0.2290	0.2291	0.1614	0.0889
	5	5	1.0000	1.0000	1.0000	0.5905	0.1332
10	5	5	0.3507	0.3482	0.2804	0.3848	0.0801
	5	10	0.4318	0.4248	0.2931	0.4268	0.1578
	5	20	0.8198	0.8154	0.4847	0.7248	0.6348
25	10	10	0.3572	0.3309	0.1126	0.1958	0.4233
	10	15	0.4122	0.3879	0.1266	0.2203	0.4756
	10	25	0.8220	0.8086	0.2328	0.4515	0.8680
Type I error			0.0568	0.0521	0.0561	0.0479	0.0454

Table 5.9. Comparison of power for $L = 5$ and $a = 50$ for i.i.d. binomial model.

n_1	a_1	a_2	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	3	8	0.1740	0.1738	0.1678	0.0882	0.0435
		10	0.2310	0.2306	0.2064	0.1213	0.0642
10	5	10	0.3941	0.3940	0.2664	0.2994	0.1129
		20	0.8034	0.8026	0.4784	0.6681	0.5313
20	8	20	0.5643	0.5624	0.1782	0.3047	0.5429
		30	0.9957	0.9957	0.4108	0.7199	0.9953
Type I error			0.0501	0.0499	0.0441	0.0307	0.0290

Table 5.10. Comparison of power for $a = 100$ for i.i.d. Poisson model.

n_1	a_1	a_2	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	4	10	0.1293	0.1259	0.1695	0.0889	0.0401
	5	20	1.0000	1.0000	1.0000	0.4287	0.2153
10	5	10	0.0665	0.0717	0.0753	0.0629	0.0427
	10	20	1.0000	1.0000	0.7258	1.0000	0.6466
20	10	20	0.0999	0.0940	0.0674	0.0963	0.0730
	10	25	0.1652	0.1518	0.0804	0.1329	0.1520
	15	25	1.0000	1.0000	0.2193	0.5387	1.0000
Type I error			0.0496	0.0556	0.0470	0.0481	0.0483

Table 5.11. Comparison of power for $a = 300$ for i.i.d. Poisson model.

n_1	a_1	a_2	T_3	P_{min}	$S_{5,5}$	$S_{10,10}$	$S_{20,20}$
5	6	50	0.5838	0.5840	0.5940	0.5010	0.3372
		60	0.8737	0.8737	0.8411	0.8054	0.6222
10	12	50	0.5109	0.5116	0.2875	0.5497	0.2324
		60	0.7383	0.7383	0.3941	0.7365	0.4597
10		70	0.9413	0.9415	0.5459	0.9286	0.7528
20	25	50	0.2088	0.2091	0.0906	0.1791	0.2380
		75	0.4831	0.4848	0.1490	0.3151	0.6037
		100	0.9987	0.9985	0.3639	0.8044	0.9999
Type I error			0.0489	0.0478	0.0480	0.0457	0.0461

5.5 Summary

In this chapter, we investigated the performance of a maximum scan score-type statistic and a minimum p-value statistic, as two-dimensional variable window-type statistics, for binomial and Poisson models, and for unconditional and conditional models. From the numerical results it is evident that both statistics perform quite well in comparison with fixed window scan statistics, when the size of the scanning window is not known. From a computational point of view, it is easier to implement the minimum p-value statistic. We intend to investigate further the performance of this statistic for continuous-type data in two-dimensional regions.

References

1. Barbour, A. Chryssaphinou, O. and Roos, M. (1996). Compound Poisson approximation is system reliability, *Naval Research Logistics*, **43**, 251–264.

2. Boutsikas, M.V. and Koutras, M.V. (2000). Reliability approximations for Markov chain imbeddable systems, *Methodology and Computing in Applied Probability* **2**, 393–412.

3. Chen, J. and Glaz, J. (1999). Approximations for discrete scan statistics, In: *Recent Advances on Scan Statistics*, J. Glaz and N. Balakrishnan (Eds.), p. 27–66. Birkhäuser, Boston, MA.

4. Chen, J. and Glaz, J. (2002). Approximations for a conditional two-dimensional scan statistic, *Statistics & Probability Letters*, **58**, 287–296.

5. Cressie, N.A.C. (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

6. Darling, R.W.R. and Waterman, M.S. (1986). Extreme value distribution for the largest cube in random lattice, *SIAM J. Appl. Math.* **46**, 118–132.

7. Duczmal, L. and Buckeridge, D.L. (2006). A workflow spatial scan statistic, *Statistics in Medicine*, **25**, 743–754.

8. Fu, J.C. and Koutras, M.V. (1994). Poisson approximation for 2-dimensional patterns, *Ann. Inst. Statist. Math.* **46**, 179–192.

9. Glaz, J. (1996). Discrete scan statistics with applications to minefield detection, In: *Detection and Remediation Technologies for Mines and Mine-Like Target*, (A. C. Dubey, R. L. Barnard, C. J. Lowe and J. E. McFee Eds.), 420–429. Proceedings SPIE, 10th Annual International Aero Sense Symposium, Orlando, Florida.

10. Glaz, J. and Naus, J. (1991). Tight bounds for scan statistics probabilities for discrete data, *Annals of Applied Probability*, **1**, 306–318.

11. Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics,* Springer-Verlag, New York.

12. Glaz, J. and Zhang, Z. (2004). Multiple window discrete scan statistics, *Journal of Applied Statistics*, **31**, 979–992.

13. Glaz, J. and Zhang, Z. (2006). Maximum scan score-type statistics, *Statistics & Probability Letters*, **76**, 1316–1322.

14. Guerriero, M., Willett, P. and Glaz, J. (2009). Distributed target detection in a sensor network using scan statistics, *IEEE Transactions on Signal Processing (*submitted in revised form*)*.

15. Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes, *Proceedings of the National Academy of Sciences*, **97** (17), 9615–9617.

16. Karwe, V. V. and Naus, J. (1997). New recursive methods for scan statistics probabilities, *Computational Statistics and Data Analysis*, **23**, 389–404.

17. Koen, C. (1991). A computer program package for the statistical analysis of spatial point processes in square, *Biometric Journal*, **33**, 493–503.

18. Koutras, M.V., Papadopoulus, G.K. and Papastavridis, S.G. (1993). Reliability of 2-dimensional consecutive-k-out-of n: F systems, *IEEE Transactions on Reliability,* **42**, 658–661.

19. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

20. Kulldorff, M. (2006). Tests for spatial randomness adjusting for an underlying inhomogeneity: A general framework, *Journal of the American Statistical Association*, **101**, 1289–1305.

21. Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference, *Statistics in Medicine*, **14**, 799–810.

22. Kulldorff, M., Tango, T. and Park, P.J. (2003). Power comparisons for disease clustering tests, *Computational Statistics & Data Analysis*, **42**, 665–684.

23. Malinowski, J. and Preuss, W. (1995). Reliability of circular consecutively-connected systems with multistate components, *IEEE Transactions on Reliability*, **44**, 532–534.

24. Modarres, R. and Patil, G.P. (2007). Hotspot detection with bivariate data, *Journal of Statistical Planning and Inference*, **137**, 3643–3654.

25. Neill, D.B. and Moore, A.W. (2004). A fast multi-resolution method for detection of significant spatial disease clusters, *Advances in Neural Information Processing Systems*, **16**, 651–658.

26. Neill, D.B. and Moore, A.W. (2006). Methods for detecting spatial and spatio-temporal clusters, *Handbook of Biosurveillance*, 243–254, Elsevier, Amsterdam.

27. Neill, D.B. and Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance, *Advances in Disease Surveillance*, **4**, 106.

28. Panayirci, E. and Dubes, R.C. (1983). A test for multidimensional clustering tendency, *Pattern Recognition* **16**, 433–444.

29. Patil, G.P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

30. Perone-Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False discovery control for random fields, *Journal of the American Statistical Association*, **99**, 1002–1014.

31. Perone-Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2007). Scan clustering: A false discovery approach, *Journal of Multivariate Analysis*, **98**, 1441–1469.

32. Pfaltz, J.L. (1983). Convex clusters in a discrete m-dimensional space, *SIAM Journal Computing*, **12**, 746–750.

33. Rosenfeld, A. (1978). Clusters in digital pictures, *Information and Control,* **39**, 19–34.

34. Salvia, A.A and Lasher, W.C. (1990). 2-dimensional consecutive-k-out-of n: F models, *IEEE Transactions on Reliability,* **39**, 382–385.

35. Saperstein, B. (1976). The analysis of attribute moving averages: MIL-STD-105D reduced inspection plans, *Proceedings of Sixth Conference on Stochastic Processes and Applications*, Tel-Aviv University.

36. Song, C. and Kulldorff, M. (2003). Power evaluation of disease clustering tests, *International Journal of Health Geographics*, **2**, 9.

37. Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11.

38. Tango, T. (2007). A spatial scan statistic scanning only the regions with elevated risk, *Advances in Disease Surveillance*, **4**, 117.

39. Zhang, Z. and Glaz, J. (2008). Bayesian variable window scan statistics, *Journal of Statistical Planning and Inference*, **138**, (11), 3561–3567.

6

Applications of Spatial Scan Statistics: A Review

Marcelo Azevedo Costa[1] **and Martin Kulldorff**[2]

[1] *Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*
[2] *Department of Ambulatory Care and Prevention, Harvard University and Harvard Pilgrim Health Care, Boston, USA*

Abstract: In 1965, Joseph Naus published his now classical paper on spatial scan statistics, entitled 'Clustering of random points in two dimensions'. This paper set in motion an important statistical theory of spatial scan statistics and an avalanche of spatial scan statistics applications in a wide variety of fields, including archaeology, astronomy, brain imaging, criminology, demography, early detection of disease outbreaks, ecology, epidemiology, forestry, geology, history, psychology and veterinary medicine. In this chapter, we survey this wide variety of applications.

Keywords and phrases: Scan statistic, spatial, geography, applications

6.1 Introduction

Suppose we observe a number of points located within a geographical or spatial region. These points may, for example, reflect the locations of trees, ant nests, diseased individuals or post offices. The general aim of the spatial scan statistic is to detect and evaluate the statistical significance of a spatial cluster of events that cannot be explained by an underlying probability model defined by a null hypothesis of spatial randomness. There are spatial scan statistics for two, three and more dimensions. If the scanning is done over a three-dimensional area defined by both space and time, we have a space-time scan statistic, which is an important special case of the three-dimensional spatial scan statistic.

Since first presented by Naus in 1965, spatial scan statistics have been applied in many different fields such as infectious diseases, cancer, cardiology, pediatrics, rheumatology, auto-immune diseases, neurological diseases, liver diseases, diabetes, geriatrics, parasitology, alcohol and drugs, accidents, veterinary medicine, demography, forestry, toxicology, psychology, medical imaging, history, criminology, astronomy and geology. The aim of this chapter is to present

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_6,

a review of the areas in which the spatial scan statistic has been applied, providing a broad sense of how it is being used across the globe and across scientific disciplines. After a brief methodological review, we present examples of applications by field of study. A final discussion presents a brief summary of the main findings.

Although this chapter emphasizes the use of spatial scan statistics, there are many other important spatial statistical methods. From a user perspective, the spatial scan statistic is best viewed as one of several important tools for the successful analysis of geographical and spatial data. Other important methods include visualization techniques, descriptive statistics of rates and proportions, spatial smoothing methods, kriging, global clustering tests, regression for spatially correlated data and so on.

6.2 Brief Methodological Overview

Suppose we have a square region with a number of points. In its original form, studied by Naus (1965), the spatial scan statistic consists of a rectangular scanning window with a fixed size and shape. This window is continuously moved over the predefined square study region, covering all possible locations, and the definition of the spatial scan statistic is the maximum number of points in the scanning window at any given time. The next step is to find the probability of observing at least that many points within the window, under the null hypothesis of randomly located points, generated by a homogeneous Poisson process. In mathematical language, we want to know the probability of finding at least one rectangle with dimensions u and v with at least n out of N points uniformly distributed in the unit square. While simple to state, the complexity of this problem lies in the multiple testing inherent in the many window locations and the overlapping nature of those windows, resulting in the maximum being taken from a set of highly dependent observations. Using some very beautiful and powerful mathematics, Naus (1965) developed theoretical formulas to obtain upper and lower bounds for those probabilities, showing that the bounds converge to the true probability.

Following the pioneering paper by Naus, there have been a number of further methodological developments of spatial scan statistics, in order to handle different types of data. The spatial region to be scanned may be of different shapes; the scanning window may be of different sizes and shapes; the analysis may or may not be conditional on the total number of points observed; the observations may be generated by a homogeneous Poisson process, an inhomogeneous Poisson process, or by a Bernoulli, multinomial, normal or exponential distribution function; there may be a need to adjust for covariates or temporal trends; and

so on. For each application, the scan statistic parameters and probabilistic models must be appropriately selected to fit the data and the scientific questions asked.

The study region is usually defined directly by the data and can be of a variety of shapes and sizes. The exact locations of the points may be known, so that we have a spatial point process. Alternatively, the data may be spatially aggregated, so that instead of points we have counts in a set of squares on a lattice or in a set of administrative geographical areas such as postal codes, census tracts or counties.

An important component of the spatial scan statistic is the shape and size of the scanning window. Naus (1965) used a rectangular window of any fixed shape and size, while Loader (1991) used a variable size rectangular window. Alm (1997, 1998) used circles, ellipses and triangles. Kulldorff (1997) defined a spatial scan statistic for any variably sized collection of windows, using a continuously variable size circle in his example. More recently, spatial scan statistics have also been defined using non-parametrically defined windows [Duczmal and Assunção (2004), Patil and Taillie (2003, 2004), Assunção *et al.* (2006), Tango and Takahashi (2005)], taking very irregular shapes. The shape of the window does not need to be the same as the shape of the study region.

Rather than defining the null hypothesis based on a homogeneous Poisson process, another assumption for the null hypothesis implies that intensity varies within the region, following an underlying known population defined by an inhomogeneous Poisson process [Turnbull *et al.* (1990)]. Areas with higher population are then expected to have more points under the null hypothesis, reflecting, for example, the fact that there are more cancer cases per geographical unit in urban compared to rural areas, simply because of the higher population density. Spatial scan statistics have also been developed for discrete 0/1 Bernoulli data [Chen and Glaz (1996), Kulldorff (1997)], as well as for multinomial [Jung, Kulldorff and Klassen (2007)], normal [Kulldorff, Huang and Konty (2008); Huang *et al.* (2009)] and survival type data [Huang, Kulldorff and Gregorio (2007); Cook, Gold and Li (2007)].

An important extension of the spatial scan statistic is to three or more dimensions [Alm (1998)]. The most common of these is the space-time scan statistic, where time is added as a third dimension [Kulldorff *et al.* (1998)]. The size and shape of the study region and scanning window can be defined as before for the purely spatial scan statistic, while time is added as a third dimension. Retrospective space-time scan statistics provide a mechanism to detect and evaluate past or present clusters that might have appeared anytime during the study period. Prospective space-time scan statistics only consider windows that touch the current date in order to only detect and evaluate the existence of clusters that are currently present. The latter method is used in early disease outbreak detection surveillance systems [Kulldorff (2001); Kulldorff *et al.* (2005)].

As we survey the application of spatial and space-time scan statistics, we will consider most of the above variants of the spatial scan statistic, as different versions are useful for different types of applications.

6.3 Applications in Medical Imaging

The spatial scan statistic has been applied to important problems in brain imaging. Naiman and Priebe (2001) have used it for positron emission tomography (PET) scan brain imagery data. Yoshida, Naya and Miyashita (2003) have applied it for neural response data in monkeys. Injections of retrograde tracers in a specific region (cases) and adjacent regions (control) in the brain generated maps with pixels associated to selective and non-selective neurons. Significant clusters of selective neurons were found, assuming a Bernoulli model.

Spatial scan statistics have also been used for breast cancer digital mammography data. The goal is the detection of clustered microcalcifications, which may be indicative of a cancerous tumor [Priebe, Olson and Healy (1997a); Naiman and Priebe (2001)]. Popescu and Lewitt (2006a, 2006b) mimic a cancer nodule detection system. A circular scanning window with fixed radius and variable center is used. The test statistic is the sum of the values of the pixels inside the window. The null distribution of the test statistic is generated by scanning background-only images.

6.4 Applications in Cancer Epidemiology

The incidence, prevalence or mortality rates of cancer may vary geographically for a number of reasons, including spatial variation of environmental or behavioral risk factors or the genetic make-up of the population. Spatial scan statistics have often been used to detect and/or evaluate the statistical significance of geographical cancer clusters, as cancer clusters will also occur simply by chance in some parts of the map. Leukemia was the first cancer that was observed using spatial scan statistics, with Turnbull *et al.* (1990) studying leukemia in upstate New York and Hjalmars *et al.* (1996) studying childhood leukemia in Sweden. Hjalmars *et al.* (1996) did not find any statistically significant clusters in their data, even though there had been one leukemia cluster alarm reported in the press a few years earlier. While the cluster was detected, it was not statistically significant and was not even among the three top clusters. In contrast, Viel *et al.* (2000) found a statistically significant cluster of soft-tissue sarcoma and

non-Hodgkins lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels in France.

Under the null hypothesis, the incidence or mortality of cancer is assumed to follow a Poisson distribution, where the expected number of cases in a particular location is proportional to the covariate-adjusted population in that location. Age and other covariates are adjusted for by using indirect standardization. Let $b_{(i,k)}$ be the population in age group k in location i and let $B_{(k)}$ be the total population in age group k. Let $C_{(k)}$ be the total number of cases in age group k. The indirectly age standardized expected number of cases in location i is then

$$\mu_i = \sum_k b_{(i,k)} \cdot C_{(k)}/B_{(k)} \tag{6.1}$$

Age must always be adjusted for in cancer incidence and mortality studies. If not, there will be significant clusters in areas with a predominately older population, since older people are at higher risk of being diagnosed with and dying from cancer. It is often interesting to also adjust for other known risk factors, including socio-economic variables such as ethnicity, educational levels or urbanicity, as well as biological variables such as skin color for skin cancer or parity for breast cancer studies [Kulldorff *et al.* (1997); Hsu *et al.* (2004); Klassen, Kulldorff and Curriero (2005)]. A very interesting approach is to reduce the number of socio-economic variables by only taking a few components from principal component analysis [Sheehan *et al.* (2004), Sheehan and DeChello (2005), Fukuda *et al.* (2005)], usually two independent components. After an adjustment, cancer clusters will disappear if they can be explained by the covariates that were adjusted for. However, the number of clusters can also increase, as a true cluster can be hidden in an unadjusted analysis.

The spatial scan statistic is able to detect and evaluate the statistical significance of individual clusters, but it won't provide an estimate of incidence or mortality rates throughout the map. For that, other statistical methods are needed as a complement, such as the mapping of smoothed rates using conditional autoregressive models [Thomas and Carlin (2003) and Buntinx *et al.* (2003)].

For most cancer sites, there may be a long time between exposure and diagnosis and an even longer time between exposure and death. Han *et al.* (2004) presents a notable approach for breast cancer clustering analysis by using place of residency at the (i) time of birth, (ii) time of menarche and (iii) time of birth of the first child, as alternative geographical coordinates in separate spatial analyses. In this way, the study provides an opportunity to examine geographical clustering of breast cancer at various points during life. Significant clusters were found for the time of birth and time of menarche analyses, with similar results. There were fewer clusters when the data was analyzed using place of residence at time of diagnosis.

Spatial scan statistics can also be used to study the geographical variation of a particular subtype, in order to determine if there are geographical clusters of late stage cancer or cancer of a particular type or grade [Roche, Skinner and Weinstein (2002); Gregorio *et al.* (2002); Sheehan and DeChello (2005); Klassen, Kulldorff and Curriero (2005)]. The detection of a geographical cluster with a high proportion of late stage breast cancer cases may indicate a need to improve breast cancer mammography screening in that geographical area. In these analyses, no census population data are used. Rather, the total number of diagnosed cancer cases is the 'population' while the 'cases' are those cancer cases that are of a certain type, such as late stage. A Bernoulli probability model is suitable for this type of data. These types of spatial scan statistics have also been used to study the geographical variation in cancer treatments [Gregorio *et al.* (2001)]. When there are more than two different stages or grades, it is possible to use a spatial scan statistic for ordinal data, which Jung, Kulldorff and Klassen (2007) did for prostate cancer stage in Maryland, United States.

There may also be an interest in the geographical variation in the survival time after a cancer diagnosis, to determine if there are geographical areas with exceptionally poor survival. This is a continuous outcome. Such analyses must be able to handle censored data and adjust for differences in prognostic factors such as the age of the patient and the stage or grade of the cancer. Using a spatial scan statistic for exponentially distributed data with censoring, Huang, Kulldorff and Gregorio (2007) studied prostate cancer survival in Connecticut, United States.

6.5 Applications in Infectious Disease Epidemiology

In infectious disease surveillance, the spatial and space-time scan statistics are used for two different purposes. The first is retrospective in nature, where historical data are used to detect geographical areas with many cases of the disease. Such clusters can either be temporary in nature, due to an outbreak, or long lasting, if the area or population is especially prone to infection. Different aspects of the infectious disease will influence the proper choice of spatial scan statistic parameters. The incubation time of the disease, for example, is a very important feature to incorporate in the selection of the scanning time window length.

Cousens *et al.* (2001) describe the spatial investigation of 84 cases of variant Creutzfeldt–Jakob disease (vCJD), a rare and fatal disease caused by the same transmissible agent as in bovine spongiform encephalopathy (mad cow disease) and therefore hypothetically associated with the consumption of beef products. With the spatial scan statistic, one statistically significant cluster with five

cases was detected. A subsequent investigation revealed a local butcher shop as a likely common source of infection.

Fevre *et al.* (2001) used the spatial scan statistic to study sleeping sickness in Uganda. Sleeping sickness is caused by a parasite that is transmitted to humans by the *tsetse* fly, which picks up the infection from domestic cattle. A purely spatial analysis was performed using the number of cases diagnosed over a 32-day period, from the time of the first recorded case to the time when vector control measures started to be implemented. A case control study was designed, where each case was matched with one control by age, gender and month of admission. Consequently, the spatial analysis was carried out assuming that cases and controls followed a Bernoulli distribution. One significant cluster was found around an important regional cattle market.

Chaput, Meek and Heimer (2002) provide some useful insights into exploring the data through evaluating separate data streams from just-confirmed and confirmed plus probable cases of human granulocytic ehrlichiosis. A spatial analysis in a 12-town area for tick-borne infections is presented using cases during four years of surveillance. The cluster analyses were conducted using either confirmed or both confirmed and probable disease cases obtained from active and passive surveillance system reports. Both datasets provided similar results.

A purely spatial analysis of a variation of vCJD in France is presented by Huillard d'Aignaux *et al.* (2002). In addition to the use of the spatial scan statistic for cluster detection and evaluation, exploratory analyses are also provided, including maps and tests for global spatial clustering. Due to evidence that the incubation period for the disease can be longer, the cluster analyses were done for both place of residency and place of birth.

Listeriosis is a bacterial food-borne pathogen that may be present in 1 to 5 percent of common ready-to-eat food products and which can cause a rare severe invasive disease manifestation and even death in humans [Sauders *et al.* (2003)]. In particular, since the spread of the bacteria is associated with contaminated food, the source of exposure might come from either global food distribution or local sources. As a consequence, spatial-temporal clustering might detect large or small clusters. A cluster analysis using the spatial scan statistic was conducted using different molecular subtyping strategies (ribotype) from sterile sites. Clusters with the same subtyping may represent clusters with a common source of exposure, potentially increasing the ability to detect outbreaks.

When studying sexually transmitted diseases, Wylie, Cabral and Jolly (2005) also used the spatial scan statistic by differentiating the cases by genotype. According to the authors, the underlying assumption behind genotyping is that two individuals infected by the same strain of an infectious agent are more likely to have an epidemiological link to each other than two individuals infected by a different strain.

Pearl *et al.* (2006) used the scan statistic to detect outbreaks of *Escherichia coli* O157. The study used a sequential cluster detection procedure, which starts with a purely temporal analysis followed by a purely spatial analysis for each year and finally by a spatio-temporal analysis.

The second purpose for using scan statistics for infectious disease data is prospective in nature, when continuously collected data are analyzed in real or near real time in order to quickly detect an emerging infectious disease outbreak. In most cases, a space-time scan statistic is then used. As soon as a new cluster is detected, specific actions to contain and eradicate the contaminant source of the disease or to stop the disease dynamics would be taken.

Mostashari *et al.* (2003) have proposed a surveillance system for West Nile virus through the daily reporting of dead birds by the public. The county-level density of dead birds and crows was strongly correlated with levels of West Nile virus activity in 2000, suggesting that dead bird surveillance could detect subsequent outbreaks in. Multiple dead bird reports for the same location on the same day were counted as one. Results show that in most cases, dead bird clusters not only preceded the time of collection of mosquitoes and birds that were tested positive for West Nile virus but also the reports of human cases near the cluster area.

Space-time scan statistics have also been used for syndromic surveillance, where a daily feed of automated medical health records is used for the early detection of infectious disease outbreaks [Kulldorff *et al.* (2005)].

6.6 Applications in Parasitology

Enemark *et al.* (2002), Washington *et al.* (2004), Odoi *et al.* (2004) and Reperant and Deplazes (2005) have all used spatial statistics in parasitology. A very nice subtype clustering analysis is presented by Enemark *et al.* (2002) for *Crypstosporidium parvum*, a protozoan parasite that infects the gastrointestinal tract and is recognized as a major cause of diarrhea. Washington *et al.* (2004) performed clustering analysis in sentinel sites before and after a public intervention program for the elimination of lymphatic filariasis in Haiti. After the intervention occurred, the most significant cluster was found in an area where drug coverage was low. Odoi *et al.* (2004) used the spatial scan statistic to study giardiasis in Canada and Reperant and Deplazes (2005) used it to study *Capillaria hepatica* infection in Switzerland.

6.7 Other Medical Applications

Hypoplastic left heart malformation is a congenital cardiovascular malformation. Parental exposure to various categories of solvents is correlated to the occurrence of cases in newborn children. Kuehl and Loffredo (2006) used the spatial scan statistic to search for disease clusters and evidence of industrial release of solvents in Baltimore, Maryland, United States. After geographical clusters were detected, the results were used to fit different multiple logistic regression models stratified by residence within or outside the clusters at the time of conception.

Several papers [Sankoh *et al.* (2001), George *et al.* (2001), Forand *et al.* (2002), Andrade *et al.* (2004), Ozdenerol *et al.* (2005), Ali *et al.* (2005)] have used spatial and space-time scan statistics for pediatric data. Sankoh *et al.* (2001) analyzed childhood mortality in northwest Burkina Faso (West Africa) in the 1993–1998 period. A purely spatial analysis was performed for each year of data, providing time-independent analyses that detect clusters for specific years. Their results show that a particular village was found as the most likely cluster in both the purely spatial and space-time analyses. When this village was omitted from the analysis, a new analysis was conducted which identified the previous secondary cluster as the most likely. Data exclusion is one way to focus spatial clustering away from an evident area.

Sabel *et al.* (2003) used the spatial scan statistic to detect and evaluate geographical clusters of amyotrophic lateral sclerosis in Finland. Separate analyses were done using place of birth and place of death as the geographical coordinates. The cluster found using the place of birth overlapped with the most significant cluster found using the place of death.

Using the spatial scan statistic, Ala *et al.* (2006) showed that the prevalence of primary biliary cirrhosis patients listed for transplantation was higher near a New York City superfund toxic waste site. In this particular analysis, a focused cluster analysis was also done by including the longitude and latitude of each New York City superfund site. This approach changed the center of the most significant cluster to a new location.

The spatial scan statistic has also been used for systemic sclerosis in the United States [Walsh and Fenster (1997)], lupus in the United States [Walsh and DeChello (2001)], diabetes in Canada [Green *et al.* (2003)], multiple sclerosis in Scotland [Donnan *et al.* (2005)] and asthma in the United States [Cook, Gold and Li (2007)], among many other diseases and locations. It has also been used to study the geography of alcohol and drug use [Hanson and Wieczorek (2002)] and pesticide exposure [Sudakin, Horowitz and Giffin (2002)].

6.8 Applications in Veterinary Medicine

In veterinary medicine, spatial scan statistics have been used for domestic animals as well as wildlife. Many different domestic animals have been studied, including cattle [Norström, Pfeiffer and Jarp (2000)], horses [USDA (2001)], sheep [Ward (2001); Falconi, Ochs and Deplazes (2002)], pigs [Berke and Grosse (2003)], chickens and turkeys [Guerin *et al.* (2005)], farmed salmon [Knuesel, Segner and Wahli (2003)] and dogs [Ward (2002)]. The spatial scan statistic has been especially popular for epidemiological investigations of bovine spongiform encephalopathy (mad cow disease), with studies in Switzerland [e.g. Schwermer *et al.* (2002)], France [Abrial *et al.* (2003)], Ireland [Sheridan *et al.* (2005)], Spain [Allepuz *et al.* (2007)] and the Netherlands [Heres, Brus and Hagenaars (2008)].

For wildlife data, the spatial scan statistic has been used to study various diseases among foxes in Germany [Berke *et al.* (2002)], sea otters in California [Miller *et al.* (2002)], coyotes in California [Hoar *et al.* 2003], deer in Wisconsin [Joly *et al.* (2003)] and badgers in Ireland [Olea-Popelka *et al.* (2003)]. When evaluating spatial clusters for wildlife data, a main challenge is the nonstationarity of many animals, and their ability to travel a long distance before being sampled [Hoar *et al.* 2003]. Miller *et al.* (2002) tried to detect spatial clusters of parasites in sea otters, but possibly due to high mobility, the spatial analysis did not detect any statistically significant clusters. An alternative is to sample static sources of isolates such as animal carcasses [Smith *et al.* (2000)].

In the geographical analysis of disease, it is often useful to use multiple spatial statistical methods to investigate different aspects or features of the spatial pattern. For example, in their study of acute respiratory disease in Norwegian cattle, Norström, Pfeiffer and Jarp (2000) also used the Knox test (1964) and Jacquez's k-nearest neighbor test (1996) to look at space-time interaction and a kernel-density interpolation for exploratory analysis. Sheridan *et al.* (2005) used the coordinates of major cattle feed suppliers to evaluate clusters around such prespecified locations by using a focused cluster test. Results provided evidence of association between significant clusters and feed sources.

6.9 Applications in Forestry

Coulston and Riitters (2003) and Riitters and Coulston (2005) have used the spatial scan statistic for forest data from the eastern United States. In a purely spatial analysis, the population size is the number of 0.009-ha units of forest land in a county and the number of cases are the number of units with perforated

forest, which is forest located near holes in an otherwise intact and continuous forest cover. So, counties with a higher population mean more forest land, and counties with a high ratio of cases to population mean a high proportion of perforated forest. In a posterior spatial analysis, they take a previously detected primary cluster as the new study region and apply the spatial scan statistic a second time to see if there are any new smaller clusters within the old larger cluster. In this way, they have found several small clusters arranged in a linear fashion along the I-95 highway. This result shows that the primary cluster had an irregular spatial component. In another analysis, using the space-time scan statistic and 10 years of data, they defined cases as the number of units with insects or pathogens.

Tuia *et al.* (2008) used the spatial scan statistic to detect and evaluate space-time clusters of forest fires. They conclude that the 'evaluation of the presence of spatial and temporal patterns in fire occurrence and their significance could have a great impact in forthcoming studies on fire occurrences prediction'.

6.10 Applications in Geology

Conover, Bement and Iman (1979) applied the spatial scan statistic to geology data, where the aim was to detect uranium deposits by using radiation measurement taken from an airplane. As the measurements contain a fair amount of random background noise, the goal was to detect clusters of high radiation readings.

6.11 Applications in Astronomy

Astronomy would seem like a natural area of application for the three-dimensional scan statistic, but we are not aware of any such application. However, the two-dimensional scan statistic has been applied in astronomy. In a study on star formation, Marcos and Marcos (2008) used the two-dimensional scan statistic to study the spatial clustering of 'open star clusters', which are physically related groups of stars held together by mutual gravitational attraction. The 'spatial' study regions were defined by galactic longitude as the first dimension and either radial velocity, proper motion or inclination as the second dimension, in three different analyses. A number of statistically significant clusters were found.

6.12 Applications in Psychology

Margai and Henry (2003) used the spatial scan statistic to detect geographical clusters of high prevalence of learning disabilities among children in Binghamton, New York, United States. They found a statistically significant cluster in the northwestern part of the city. As a complement to the spatial scan statistic, they used Moran's I to evaluate whether there was general evidence of global spatial clustering throughout the city. They also explored a set of socio-economic variables potentially correlated to the spatial occurrence of individuals with learning disabilities. They compared the means of these variables inside and outside the detected spatial cluster through t-tests. They also applied discriminant analysis using the cluster status as the dependent variable and significant variables obtained from previous t-test analyses. This last approach represents an alternative and indirect method to associate detected geographical clusters to a set of socio-economic variables.

6.13 Applications to Accidents

Nkhoma *et al.* (2004) applied spatial scan statistics for accidental poisoning mortality data. Cases were divided according to specific toxic agents. Both spatial and space-time scan statistics were used to evaluate the data with and without the influence of a time trend. Yiannakoulias *et al.* (2003) used the spatial scan statistic to study the geography of fall injuries in the elderly.

6.14 Applications in Criminology and Warfare

Beato *et al.* (2001) used both the spatial scan statistic and Bayesian smoothing techniques to study the geographical distribution of homicides in Belo Horizonte, Brazil. Statistically significant clusters were found in areas known for drug trafficking activities. Ceccato and Haining (2004) used the spatial scan statistic to compare the location of crime events during two distinct periods in Malmö, Sweden, before and after the building of the new Öresund bridge connecting Malmö with Copenhagen, Denmark. No significant clusters were found close to the vicinity of the bridge, but there were notable shifts in the geographical locations of some clusters as well as new clusters for some of the crimes.

Priebe, Olson and Healy (1997b) have used the spatial scan statistic for minefield detection using remote sensing data.

6.15 Applications in Demography

Callado Chavez (2003) has used the spatial scan statistic to evaluate the geography of fecundity, the potential for reproduction, in Costa Rica.

6.16 Applications in the Humanities

Spatial scan statistics are not widely used in the humanities, but there are some examples from anthropology, archaeology and history. In a very interesting study, Witham and Oppenheimer (2004) used the spatial scan statistic to study the geographical distribution of excess deaths in England due to the 1783 Laki Craters volcanic eruption in Iceland, which fumigated many parts of Europe with volcanic gases and particles. They found that the eastern part of England was the most affected region. In anthropology, Usher and Allen (2005) used the scan statistic for spatial genetic analysis to evaluate kinship clusters in cemeteries. Waller (2006) used the spatial scan statistic as well as many other spatial statistical techniques to compare the geographical distribution of early versus late period archaeological sites from the Anasazi culture in Black Mesa, Arizona.

6.17 Scan Statistic Software

Different versions of the spatial scan statistic have been included in a couple of statistical software packages. The freely available SaTScan^{TM} software (www.satscan.org) can be used to run the purely spatial and space-time scan statistics for Poisson, Bernoulli, multinomial, normal and exponentially distributed data. ClusterSeer (www.terraseer.com) is a commercial software that includes the purely spatial and space-time scan statistics together with a number of other spatial statistical methods.

6.18 Discussion

Different types of data require different forms of the spatial scan statistic, but the underlying principle is the same as in the pioneering paper by Naus in 1965. In this chapter, we have presented a partial sample of the applications for which the spatial scan statistic has been used. As can be seen from the literature review, the spatial scan statistic has been applied in a remarkable number of different subject areas, from the small spatial scale of medical imaging to the large spatial scale of astronomy. The method is most commonly used in cancer, infectious disease and veterinary epidemiology. These are areas with a long and strong interest in epidemiology in general. They are also areas with a long tradition of disease cluster and outbreak investigations, for which the spatial scan statistic is ideally suited.

The spatial scan statistic is increasingly being used for other diseases as well. The number of applications in non-medical areas is more limited, but we think that may change with time. With the increasing use of geographical information systems in many different disciplines, there will be an increase in the use of formal methods of statistical inference to complement the beautiful maps that are created. Areas for which we think that the spatial scan statistic will play an especially important role include archaeology, astronomy, criminology, demography, ecology, geography and medical imaging.

The spatial scan statistic is also used in ways that do not lead to publications in scientific journals. For example, many public health officials use it for routine disease surveillance on a daily, weekly or yearly basis to monitor the geographical distribution of disease. Likewise, spatial scan statistics are used by law enforcement agencies for the routine monitoring of crime activities.

Acknowledgment

This research was funded by grant #RO1CA095979 from the National Cancer Institute.

References

1. Abrial, D., Calavas, D., Lauvergne, N., Morignat, E. and Ducrot, C. (2003). Descriptive spatial analysis of BSE in western France, *Veterinary Research*, **34**, 749–760.

2. Ala, A., Stanca, C.M., Bu-Ghanim, M., Ahmado, I., Branch, A.D., Schiano, T.D., Odin, J.A. and Bach, N. (2006). Increased prevalence of primary biliary cirrhosis near superfund toxic waste sites, *Hepatology*, **43**, 525–531.

3. Ali, M., Asefaw, T., Byass, P., Beyene, H. and Karup Pedersen, F. (2005). Helping northern Ethiopian communities reduce childhood mortality: population-based intervention trial, *Bulletin of the World Health Organization*, **83**, 27–33.

4. Allepuz, A., López-Quílez, A., Forte, A., Fernández, G. and Casal, J. (2007). Spatial analysis of bovine spongiform encephalopathy in Galicia, Spain (2000-2005), *Preventive Veterinary Medicine*, **79**, 174–185.

5. Alm, S.E. (1997). On the distributions of scan statistics of a two dimensional Poisson process, *Adv. in Appl. Probab.*, **29**, 1–18.

6. Alm, S.E. (1998). Approximation and simulation of the distributions of scan statistics for Poisson processes in higher dimensions, *Extremes*, **1**, 111–126.

7. Andrade, A.L., Silva, S.A., Martelli, C.M., Oliveira, R.M., Morais Neto, O.L., Siqueira Junior, J.B., Melo, L.K. and Di Fabio, J.L. (2004). Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of central Brazil, *Cadernos de Saúde Pública*, **20**, 411–421.

8. Assunção, R., Costa, M.A., Tavares, A. and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters, *Statistics in Medicine*, **25**:5, 723–742.

9. Beato Filho, C.C., Assunção, R.M., Silva, B.F., Marinho, F.C., Reis, I.A. and Almeida, M.C. (2001). Homicide clusters and drug traffic in Belo Horizonte, Minas Gerais, Brazil from 1995 to 1999, *Cadernos de Saúde Pública*, **17**, 1163–1171.

10. Berke, O. and Grosse Beilage, E. (2003). Spatial relative risk mapping of pseudorabies-seropositive pig herds in an animal-dense region, *Journal of Veterinary Medicine*, **B50**:4, 322–325.

11. Berke, O., von Keyserlingk, M., Broll, S. and Kreienbrock, L. (2002). On the distribution of *Echinococcus multilocularis* in red foxes in Lower Saxony: identification of a high risk area by spatial epidemiological cluster analysis. *Berliner und Munchener Tierarztliche Wochenschrift*, **115**, 428–434.

12. Buntinx, F., Geys, H., Lousbergh, D., Broeders, G., Cloes, E., Dhollander, D., Op De Beeck, L., Vanden Brande, J., Van Waes, A. and Molenberghs, G. (2003). Geographical differences in cancer incidence in the Belgian province of Limburg, *European Journal of Cancer*, **39**, 2058–2072.

13. Callado Chavez, A. (2003). Fecundidad adolescente en el gran área metropolitana de Costa Rica, *Población y Salud en Mesoamérica*, **1**, 4.

14. Ceccato, V. and Haining, R. (2004). Crime in border regions: The Scandinavian case of Öresund, 1998-2001, *Annals of the Association of American Geographers*, **94**, 807–826.

15. Chaput, E.K., Meek, J.I. and Heimer, R. (2002). Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut, *Emerging Infectious Diseases*, **8**, 943–948.

16. Chen, J. and Glaz, J (1996). Two-dimensional discrete scan statistics, *Statist. Probab. Lett.*, **23**, 751–771.

17. Conover, W.J., Bement, T.R. and Iman, R.L. (1979). On a method for detecting clusters of possible uranium deposits, *Technometrics*, **21**, 277–282.

18. Cook, A.J., Gold, D.R. and Li, Y. (2007). Spatial cluster detection for censored outcome data, *Biometrics*, **63**, 540–549.

19. Coulston, J.W. and Riitters, K.H. (2003). Geographic analysis of forest health indicators using spatial scan statistics, *Environmental Management*, **31**, 764–773.

20. Cousens, S., Smith, P.G., Ward, H., Everington, D., Knight, R.S.G., Zeidler, M., Stewart, G., Smith-Bathgate, E.A.B., Macleod, M.A., Mackenzie, J. and Will, R.G. (2001). Geographical distribution of variant Creutzfeldt-Jakob disease in Great Britain, *The Lancet*, **357**, 1002–1007.

21. Donnan, P.T., Parratt, J.D.E., Wilson, S.V., Forbes, R.B., O'Riordan, J.I. and Swingler, R.J. (2005). Multiple sclerosis in Tayside, Scotland: detection of clusters using a spatial scan statistic, *Multiple Sclerosis*, **11**, 403–408.

22. Duczmal, L. and Assunção, R.A. (2004). Simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, *Computational Statistics and Data Analysis*, **45**, 269–286.

23. Enemark, H.L., Ahrens, P., Juel, C.D., Petersen, E., Petersen, R.F., Andersen, J.S., Lind, P. and Thamsborg, S.M. (2002). Molecular characterization of Danish *Cryptosporidium parvum* isolates, *Parasitology*, **125**, 331–341.

24. Falconi, F., Ochs, H. and Deplazes, P. (2002). Serological cross-sectional survey of psoroptic sheep scab in Switzerland, *Veterinary Parasitology*, **109**, 119–127.

25. Fevre, E.M., Coleman, P.G., Odiit, M., Magona, J.W., Welburn, S.C. and Woolhouse, M.E.J. (2001). The origins of a new *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in eastern Uganda, *The Lancet*, **358**, 625–628.

26. Forand, S.P., Talbot, T.O., Druschel, C. and Cross, P.K. (2002). Data quality and the spatial analysis of disease rates: congenital malformations in New York State, *Health and Place*, **8**, 191–199.

27. Fukuda, Y., Umezaki, M., Nakamura, K. and Takano, T. (2005). Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan, *International Journal of Health Geographics*, **4**, 16.

28. George, M., Wiklund, L., Aastrup, M., Pousette, J., Thunholm, B., Saldeen, T., Wernroth, L., Zaren, B. and Holmberg, L. (2001). Incidence and geographical distribution of sudden infant death syndrome in relation to content of nitrate in drinking water and groundwater levels, *European Journal of Clinical Investigation*, **31**, 1083–1094.

29. Green, C., Hoppa, R.D., Young, T.K. and Blanchard, J.F. (2003). Geographic analysis of diabetes prevalence in an urban area, *Social Science and Medicine*, **57**, 551–560.

30. Gregorio, D.I., Kulldorff, M., Barry, L., Samociuk, H. and Zarfos, K. (2001). Geographic differences in primary therapy for early stage breast cancer. *Annals of Surgical Oncology*, **8**, 844–849.

31. Gregorio, D.I., Kulldorff, M., Barry, L. and Samociuk, H. (2002). Geographic differences in invasive and in situ breast cancer incidence according to precise geographic coordinates, Connecticut, 1991–1995. *International Journal of Cancer*, **100**, 194–198.

32. Guerin, M.T., Martin, S.W., Darlington, G.A. and Rajic, A. (2005). A temporal study of *Salmonella* serovars in animals in Alberta between 1990 and 2001, *Canadian Journal of Veterinary Research*, **69**, 88–89.

33. Han, D.W., Rogerson, P.A., Nie, J., Bonner, M.R., Vena, J.E., Vito, D., Muti, P., Trevisan, M., Edge, S.B. and Freudenheim, J.L. (2004). Geographic clustering of residence in early life and subsequent risk of breast cancer (United States), *Cancer Causes and Control*, **15**, 921–929.

34. Hanson, C.E. and Wieczorek, W.F. (2002). Alcohol mortality: a comparison of spatial clustering methods, *Social Science and Medicine*, **55**, 791–802.

35. Heres, L., Brus, D.J. and Hagenaars, T.J. (2008). Spatial analysis of BSE cases in the Netherlands, *BMC Veterinary Research*, **4**:21.

36. Hjalmars, U., Kulldorff, M., Gustafsson, G. and Nagarwalla, N. (1996). Childhood leukemia in Sweden: using GIS and a spatial scan statistic for cluster detection, *Statistics in Medicine*, **15**, 707–715.

37. Hoar, B.R., Chomel, B.B., Rolfe, D.L., Chang, C.C., Fritz, C.L., Sacks, B.N. and Carpenter, T.E. (2003). Spatial analysis of *Yersinia pestis* and *Bartonella vinsonii* subsp. *berkhoffii* seroprevalence in California coyotes (*Canis latrans*), *Preventive Veterinary Medicine*, **56**, 299–311.

38. Hsu, C.E., Jacobson, H.E. and Soto Mas, F. (2004). Evaluating the disparity of female breast cancer mortality among racial groups - a spatiotemporal analysis, *International Journal of Health Geographics*, **3**:4.

39. Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data, *Biometrics*, **63**, 109–118.

40. Huang, L., Tiwari, R., Zuo, J., Kulldorff, M. and Feuer, E. (2009). Weighted normal spatial scan statistic for heterogenous population data, *Journal of the American Statistical Association*, in press.

41. Huillard d'Aignaux, J., Cousens, S.N., Delasnerie-Laupretre, N., Brandel, J.P., Salomon, D., Laplanche, J.L., Hauw, J.J. and Alperovitch, A. (2002). Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998, *International Journal of Epidemiology*, **31**, 490–495.

42. Jacquez, G.M. (1996). A k-nearest neighbour test for space-time interaction, *Statistics in Medicine*, **15**:18, 1935–1949.

43. Joly, D.O., Ribic, C.A., Langenberg, J.A., Beheler, K., Batha, C.A., Dhuey, B.J., Rolley, R.E., Bartelt, G., Van Deelen, T.R. and Samual, M.D. (2003). Chronic wasting disease in free-ranging Wisconsin white-tailed deer, *Emerging Infectious Disease*, **9**, 599–601.

44. Jung, I., Kulldorff, M. and Klassen, A. (2007). A spatial scan statistic for ordinal data, *Statistics in Medicine*, **26**, 1594–1607.

45. Klassen, A., Kulldorff, M. and Curriero, F. (2005). Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors, *International Journal of Health Geographics*, **4**, 1.

46. Kleinman K., Abrams A., Kulldorff M. and Platt R. (2005). A model-adjusted space-time scan statistic with an application to syndromic surveillance, *Epidemiology and Infection*, **133**, 409–419.

47. Knox, E.G. (1964). The detection of spacetime interactions, *J. Appl. Stat.*, **13**, 24–30.

48. Knuesel, R., Segner, H. and Wahli, T. (2003). A survey of viral diseases in farmed and feral salmonids in Switzerland., *Journal of Fish Diseases*, **26**:4, 167–182.

49. Kuehl, K.S. and Loffredo, C.A. (2006). A cluster of hypoplastic left heart malformation in Baltimore, Maryland, *Pediatric Cardiology*, **27**, 25–31.

50. Kulldorff M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

51. Kulldorff, M., Feuer, E.J., Miller, B.A. and Freedman, L.S. (1997). Breast cancer in northeastern United States: a geographical analysis, *American Journal of Epidemiology*, **146**, 161–170.

52. Kulldorff M., Athas W., Feuer E., Miller B. and Key C. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, *American Journal of Public Health*, **88**, 1377–1380.

53. Kulldorff M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society*, **A164**, 61–72.

54. Kulldorff M., Heffernan R., Hartman J., Assunção R.M. and Mostashari F. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks, *PLoS Medicine*, **2**, 216–224.

55. Kulldorff. M., Huang, L. and Konty, K. (2008). A spatial scan statistic for normally distributed data, *Manuscript*.

56. Loader, C.R. (1991). Large-deviation approximations to the distribution of scan statistics, *Adv. in Appl. Probab.*, **23**, 751–771.

57. Marcos, R.D.L.F. and Marcos, C.D.L.F. (2008). From star complexes to the field: open cluster families, *Astrophysical Journal*, **672**, 342–351.

58. Margai, F. and Henry, N. (2003). A community-based assessment of learning disabilities using environmental and contextual risk factors, *Social Science and Medicine*, **56**, 1073–1085.

59. Miller, M.A., Gardner, I.A., Kreuder, C., Paradies, D.M., Worcester, K.R., Jessup, D.A., Dodd, E., Harris, M.D., Ames, J.A., Packham, A.E. and Conrad, P.A. (2002). Coastal freshwater runoff is a risk factor for *Toxoplasma gondii* infection of southern sea otters (*Enhydra lutris nereis*), *International Journal for Parasitology*, **32**, 997–1006.

60. Mostashari, F., Kulldorff, M., Hartman, J.J., Miller, J.R. and Kulasekera, V. (2003). Dead bird clustering: a potential early warning system for West Nile virus activity, *Emerging Infectious Diseases*, **9**, 641–646.

61. Naiman, D.Q. and Priebe, C.E. (2001). Computing scan statistic p-values using importance sampling, with applications to genetics and medical image analysis, *Journal of Computational & Graphical Statistics*, **10**, 296–328.

62. Naus, J. I. (1965). Clustering of random points in two dimensions, *Biometrika*, **52**, 263–267.

63. Norström, M., Pfeiffer, D.U. and Jarp, J. (2000). A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds, *Preventive Veterinary Medicine*, **47**, 107–119.

64. Nkhoma, E.T., Hsu, C.E., Hunt, V.I. and Harris A.M. (2004). Detecting spatiotemporal clusters of accidental poisoning mortality among Texas counties, U.S., 1980 - 2001, *International Journal of Health Geographics*, **3**:25.

65. Odoi, A., Martin, S.W., Michel, P., Middleton, D., Holt, J. and Wilson, J. (2004). Investigation of clusters of giardiasis using GIS and a spatial scan statistic, *International Journal of Health Geographics*, **3**:11.

66. Olea-Popelka, F.J., Griffin, J.M., Collins, J.D., McGrath, G. and Martin, S.W. (2003). Bovine tuberculosis in badgers in four areas in Ireland: does tuberculosis cluster? *Preventive Veterinary Medicine*, **59**, 103–111.

67. Ozdenerol, E., Williams, B.L., Kang, S.Y. and Magsumbol, M.S. (2005). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters, *International Journal of Health Geographics*, **4**:19.

68. Patil, G.P. and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection, *Statistical Science*, **18**:4, 457–465.

69. Patil, G.P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

70. Pearl, D.L., Louie, M., Chui, L., Dore, K., Grimsrud, K.M., Leedell, D., Martin, S.W., Michel, P., Svenson, L.W. and McEwen, S.A. (2006). The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli* O157 in space and time in Alberta, Canada, 2000-2002, *Epidemiology and Infection*, **134**, 699–711.

71. Popescu, L.M. and Lewitt, R.M. (2006a). Comparison between TOF and non-TOF PET using a scan statistic numerical observer, In *2006 IEEE Nuclear Science Symposium Conference Record*, **3**, 1774–1780.

72. Popescu, L.M. and Lewitt, R.M. (2006b). Small nodule detectability evaluation using a generalized scan statistic model, *Physics in Medicine and Biology*, **51**, 6225–6244.

73. Priebe, C. E., Olson, T. and Healy D.M. Jr (1997a). A spatial scan statistic for stochastic scan partitions, *Journal of the American Statistical Association*, **92**, 1476–1484.

74. Priebe, C. E., Olson, T. and Healy D.M. Jr (1997b). Exploiting stochastic partitions for minefield detection. *Proceedings of SPIE, the International Society for Optical Engineering*, **3079**, 508.

75. Reperant, L.A. and Deplazes, P. (2005). Cluster of *Capillaria hepatica* infections in non-commensal rodents from the canton of Geneva, Switzerland, *Parasitology Research*, **96**, 340–342.

76. Riitters, K.H. and Coulston, J.W. (2005). Hot spots of perforated forest in the eastern United States, *Environmental Management*, **35**, 483–492.

77. Roche, L.M., Skinner, R. and Weinstein, R.B. (2002). Use of a geographic information system to identify and characterize areas with high proportions of distant stage breast cancer, *Journal of Public Health Management and Practice*, **8**, 26–32.

78. Sabel, C.E., Boyle, P.J., Lytnen, M., Gatrell, A.C., Jokelainen, M., Flowerdew, R. and Maasilta P. (2003). Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death, *American Journal of Epidemiology*, **157**, 898–905.

79. Sankoh, O.A., Ye, Y., Sauerborn, R., Muller, O. and Becher, H. (2001). Clustering of childhood mortality in rural Burkina Faso, *International Journal of Epidemiology*, **30**, 485–492.

80. Sauders, B.D., Fortes, E.D., Morse, D.L., Dumas, N., Kiehlbauch, J.A., Schukken, Y., Hibbs, J.R. and Wiedmann, M. (2003). Molecular subtyping to detect human listeriosis clusters, *Emerging Infectious Diseases*, **9**, 672–680.

81. Schwermer, H., Rufenacht, J., Doherr, M.G. and Heim, D. (2002). Geographic distribution of BSE in Switzerland, *Schweizer Archiv fur Tierheilkunde*, **144**, 701–708.

82. Sheehan, T.J., DeChello, L.M., Kulldorff, M., Gregorio, D.I., Gershman, S. and Mroszczyk, M. (2004). The geographic distribution of breast cancer incidence in Massachusetts 1988-1997, adjusted for covariates, *International Journal of Health Geographics*, **3**, 17.

83. Sheehan, T.J. and DeChello, L.M. (2005). A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997, *International Journal of Health Geographics*, **4**, 15.

84. Sheridan, H.A., McGrath, G., White, P., Fallon, R., Shoukri, M.M. and Martin, S.W. (2005). A temporal-spatial analysis of bovine spongiform encephalopathy in Irish cattle herds, from 1996 to 2000, *Canadian Journal of Veterinary Research*, **69**, 19–25.

85. Smith, K.L., DeVos, V., Bryden, H., Price, L.B., Hugh-Jones, M.E. and Keim, P. (2000). *Bacillus anthracis* diversity in Kruger National Park, *Journal of Clinical Microbiology*, **38**, 3780–3784.

86. Sudakin, D.L., Horowitz, Z. and Giffin, S. (2002). Regional variation in the incidence of symptomatic pesticide exposures: applications of geographic information systems, *Journal of Toxicology - Clinical Toxicology*, **40**, 767–773.

87. Tango, T. and Takahashi, K. (2005). A flexible shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11.

88. Thomas, A.J. and Carlin, B.P. (2003). Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering, *Statistics in Medicine*, **22**, 113–127.

89. Tuia, D., Ratle, F., Lasaponara, R., Telesca, L. and Kanevski, M. (2008). Scan statistics analysis of forest fire clusters. *Communications in Nonlinear Science and Numerical Simulation*, **13**, 1689–1694.

90. Turnbull, B., Iwano, E.J., Burnett, W.S., Howe, H.L. and Clark, L.C. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York, *Amer. J. Epidemiology*, **132**, 136–143.

91. United States Department of Agriculture. (2001). West Nile virus in equids in the Northeastern United States in 2000. USDA, APHIS, Veterinary Services.

92. Usher, B.M. and Allen, K.L. (2005). Identifying kinship clusters: SatScan for genetic spatial analysis, *American Journal of Physical Anthropology, Supplement*, **126**, S40, 210.

93. Viel, J.F., Arveux, P., Baverel, J. and Cahn, J.Y. (2000). Soft-tissue sarcoma and non-Hodgkins lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels, *American Journal of Epidemiology*, **152**, 13–19.

94. Waller, L.A. (2006). Detection of Clustering in Spatial Data. Emory University, Department of Biostatistics, Technical Report **06-12**.

95. Walsh, S.J. and Fenster, J.R. (1997). Geographical clustering of mortality from systemic sclerosis in the Southeastern United States, *Journal of Rheumatology*, **24**, 2348–2352.

96. Walsh, S.J. and DeChello, L.M. (2001). Geographical variation in mortality from systemic lupus erythematosus in the United States, *Lupus*, **10**, 637–646.

97. Ward, M.P. (2001). Blowfly strike in sheep flocks as an example of the use of a time-space scan statistic to control confounding, *Preventive Veterinary Medicine*, **49**, 61–69.

98. Ward, M.P. (2002). Clustering of reported cases of leptospirosis among dogs in the United States and Canada, *Preventive Veterinary Medicine*, **56**, 215–226.

99. Washington, C.H., Radday, J., Streit, T.G., Boyd, H.A., Beach, M.J., Addiss, D.G., Lovince, R., Lovegrove, M.C., Lafontant, J.G., Lammie, P.J. and Hightower, A.W. (2004). Spatial clustering of filarial transmission before and after a Mass Drug Administration in a setting of low infection prevalence, *Filaria Journal*, **3**, 3.

100. Witham, C.S. and Oppenheimer, C. (2004). Mortality in England during the 1783-4 Laki Craters eruption, *Bulletin of Volcanology*, **67**, 15–25.

101. Wylie, J.L., Cabral T. and Jolly, A.M. (2005). Identification of networks of sexually transmitted infection: a molecular, geographic, and social network analysis, *Journal of Infectious Diseases*, **191**, 899–906.

102. Yiannakoulias, N., Rowe, B.H., Svenson, L.W., Schopflocher, D.P., Kelly, K. and Voaklander, D.C. (2003). Zones of prevention: the geography of fall injuries in the elderly, *Social Science and Medicine*, **57**, 2065–2073.

103. Yoshida, M., Naya, Y. and Miyashita, Y. (2003). Anatomical organization of forward fiber projections from area TE to perirhinal neurons representing visual long-term memory in monkeys, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4257–4262.

7

Extensions of the Scan Statistic for the Detection and Inference of Spatial Clusters

Luiz Duczmal,[1] **Anderson Ribeiro Duarte,**[1] **and Ricardo Tavares**[2]

[1] *Statistics Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*
[2] *Department of Mathematics, Universidade Federal de Ouro Preto, Ouro Preto, Brazil*

Abstract: Naus's early 1965 paper [Naus (1965)] on spatial scan statistics paved the way for a considerable amount of research on geographic-based statistical analysis, inspiring intensive work in the most diverse contexts and applications, including epidemiology, syndromic surveillance, criminality and environmental sciences. Following one line of work, several methods for the detection of irregularly shaped clusters were developed. New tools were devised in order to account for the spatial mobility of individuals, the study of the spatial distribution of individuals according to their survival times and multiple data streams from different sources of syndromic counts. Several algorithms explored the utilization of parametric models other than the Poisson or Bernoulli distributions, and also non-parametric and learning models. We expect that this strong trend of application-driven methodologies in spatial scan statistics should continue in the foreseeable future.

Keywords and phrases: Spatial scan statistic, epidemiology, syndromic surveillance, disease cluster, irregularly shaped spatial cluster

7.1 Introduction

Algorithms for the detection and evaluation of the statistical significance of spatial clusters are important geographic tools in epidemiology, syndromic and disease surveillance, crime prevention and environmental sciences. The elucidation of the etiology of diseases, the availability of reliable alarms for detecting intentional and non-intentional outbreaks, the study of spatial patterns of criminal activities and the geographic monitoring of environmental changes are current topics of intense research. Methods for finding spatial clusters were reviewed in Elliott, Martuzzi and Shaddick (1995), Waller and Jacquez (2000),

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_7,

Kulldorff (1999), Lawson *et al.* (1999), Moore and Carpenter (1999), Glaz, Naus and Wallestein (2001), Lawson (2001), Balakrishnan and Koutras (2002) and Buckeridge *et al.* (2005).

A descendant of Naus's pioneering spatial scan statistic, Kulldorff's spatial scan statistic [Kulldorff (1997) and Kulldorff (1999)] is currently the most popular method for finding spatial clusters. The significance of the most likely cluster is estimated through a Monte Carlo simulation [Dwass (1957)]. It can be used for data with exact point locations or for aggregated data, where a study region is partitioned into cells. The circular scan [Kulldorff and Nargawalla (1995)], the most commonly used spatial scan statistic, completely sweeps the configuration space of circularly shaped clusters, but in many situations we would like to recognize spatial clusters in a much more general geometric setting. Several proposals for finding arbitrarily shaped spatial clusters are reviewed in Section 7.2. Section 7.3 examines a number of recent data-driven algorithms for cluster detection that have been developed to include spatial mobility, survival time, multiple data streams, alternative parametric models, and non-parametric and learning models. Finally, Section 7.4 presents a small list of recent interesting applications.

7.2 Irregularly Shaped Spatial Clusters

When searching for clusters with unlimited freedom of geometric shape, the power of detection is diminished. This happens because the collection of all connected zones, irrespective of shape, is very large; the maximum value of the objective function is likely to be associated with 'tree-shaped' clusters, which merely link the highest likelihood ratio cells of the map, without contributing to the discovery of geographically meaningful solutions that correctly delineate the 'true' cluster. In other words, there is much 'noise', against which the legitimate solutions cannot be distinguished. This problem occurs in every irregularly shaped cluster detector. In this section several proposed solutions for this issue are reviewed.

The upper level sets (ULS) scan statistic [Patil and Taillie (2004)] controls the excessive freedom of shape, exploring a very small collection of graph-connected candidate zones z, evaluated according to their rate (number of cases divided by the population at risk) in the study area of n regions. The ULS-tree is constructed such that selected zones with the highest rates consisting of only one individual region, which are local maxima for the rate, form the leaves of the ULS-tree. Neighboring regions in the study area are successively joined to the individual regions represented by the leaves, forming larger zones with lower rates which are then identified with the lower inner nodes of the ULS-tree.

Eventually, those aggregated zones coalesce, creating even larger, lower rated zones, represented as inner nodes closer to the root. The root itself represents the entire study area. The collection of zones represented by the ULS-tree nodes constitutes the ULS reduced parameter space, its cardinality being at most n. The ULS-tree needs to be calculated again for each new Monte Carlo replication. This procedure is fast, but it may possibly overlook many interesting clusters, due to the small cardinality of the ULS-tree. This issue is tackled in Patil *et al.* (2006), where an extension of the original ULS set is constructed. Modarres and Patil (2007) discussed an extension of the ULS scan statistic to bivariate data. The sensitivity of the joint hotspots to the degree of association between the variables is studied.

Duczmal and Assunção (2004) proposed a simulated annealing (SA) algorithm. The collection of connected irregularly shaped zones consists of all those zones for which the corresponding subgraphs are connected. This collection is very large, and it is impractical to calculate Kulldorff's log likelihood ratio (LLR) scan statistic for all of them. Instead the SA algorithm tries to visit only the most promising zones, as follows. Two zones are neighbors when they differ by a single region. For each individual region of the study area, the circular scan is used to define a starting cluster z_0. The algorithm chooses some neighbor z_1 among all the neighbors of z_0. In the next step, another neighbor z_2 is chosen among the neighbors of z_1, and so on, until a predefined threshold in the number of regions is attained. Thus, at each step a new zone is built, adding or excluding one cell from the zone in the previous step. Instead of always behaving like a greedy algorithm, always choosing the highest LR neighbor at every step, the SA algorithm evaluates if there has been little or no LR improvement during the latest steps; in that case, the SA algorithm opts for choosing a random neighbor. This is done while trying to avoid getting stuck at LR local maxima. The search is restarted many times, each time using each individual cell of the map as the initial zone. Thus, the effect of this strategy is to keep the program openly exploring the most promising zones in the configuration space and abandoning the directions that seem uninteresting. The best solution found by the program, which maximizes the LR is the most likely cluster. It is called a quasi-optimal solution, and is a compromise due to computer time restraints for the identification of the geographical location of the clusters.

The flexibly shaped (FS) spatial scan statistic [Tango and Takahashi (2005)] made an exhaustive search of all possible first-order connected clusters contained within a set encompassing the nearest K neighbors of a given region. For each region i, the FS scan considers K concentric circles plus all the sets of connected regions whose centroids are located within the Kth largest concentric circle. The procedure is repeated for each region of the map, enabling all connected clusters to be enumerated up to a size limit K. The set of potential clusters is stored in memory, so the runs under the null hypothesis are executed

without rebuilding them every time. For computational reasons, the search is restricted to relatively small clusters. The authors consider that a practical value for K is about 30—finding clusters larger than that should take more than one week of computation on a desktop PC. Compared to the SA without bounds on cluster size, the FS algorithm finds more compact clusters, but when the SA predefined number of regions threshold is set to the same size limit K, both algorithms give similar results. Takahashi *et al.* (2007) further extended the FS scan to detect space-time irregularly shaped clusters.

The static minimum spanning tree (SMST) proposed by Assunção *et al.* (2006) used a greedy algorithm to aggregate regions. Starting with a zone consisting of one individual region, the algorithm selects the adjacent region that maximizes the LR scan statistic and aggregates it to the zone successively until a maximum population proportion is attained, or all regions are used. The procedure is repeated for each region of the study area. The paper describes this algorithm as the growth of a minimum spanning tree; it minimizes the sum of edge weights, defined as the difference in rates between vertices within the tree. Each step of tree growth represents a new candidate cluster. The most likely cluster is defined as the cluster that maximizes the LR.

The density-equalizing Euclidean minimum spanning tree (DEEMST) method [Wieland *et al.* (2007)] was an improvement of the SMST idea. A study region is provided with n points in the data set of cases and controls. Neighboring points are connected through edges, forming the complete graph T of the whole study area. Initially, a Voronoi diagram of the control locations is built, subdividing the study area into regions, satisfying the property that the density, or the number of controls in each region divided by the region's area, is kept constant. This constitutes the density-equalizing cartogram, a distorted map in which the regions are magnified or demagnified according to their local density. Next, the method finds all the potential clusters, here defined as the subset of points S such that each subset of S is closer to at least one other point in S than to any other point outside of S. The authors prove that it is not necessary to consider all connected subgraphs of T: aside from the trivial n individual points, there are only $n - 1$ non-trivial potential clusters. They are found from the Euclidean minimum spanning tree (EMST) solution using a greedy edge deletion algorithm. This method does not use the LR statistic, but instead the sum of the Euclidean distances of the minimum spanning tree. This method was compared with the circular SaTScan. It was found that the EMST has more power to detect irregularly shaped clusters, but the circular scan has more power to detect large circular clusters. Compared with the circular SaTScan, EMST obtained higher average fraction of true cluster detected for noncircular clusters, accompanied by a diminished average fraction of the most likely cluster coinciding with the true cluster. That suggests that the EMST method reports fewer false negatives, but more false positives, than SaTScan.

Dematteï, Molinari and Daurès (2007) proposed a method based on the construction of a trajectory for multiple cluster detection using the spatial scan statistic in point data sets. It begins by determining a certain trajectory linking the data set points. The general idea of the method is based on the assumption that the consecutive points inside a cluster have lower associated distances than those of points outside the cluster, because the density of points is higher within the cluster. Potential clusters are located by modelling the multiple structural changes of the distances on the selection order, and the best model (containing one or several potential clusters) is selected. Finally a p-value is obtained for each potential cluster. The authors discuss the possibility that the trajectory leaves the cluster before going through all the cluster points. They conclude that the remaining cluster points will be detected as a second component cluster and that the proximity analysis of these two component clusters by specialists could allow them to build a new bigger cluster as the union of the two clusters detected. It is not clear, however, how a fast automatic procedure could be devised to construct these unions, particularly when there are more than just a few components.

Kulldorff *et al.* (2006) presented an elliptic version of the spatial scan statistic, generalizing the circular shape of the scanning window. It uses an elliptic scanning window of variable location, shape (eccentricity), angle and size, with and without an eccentricity penalty. The elliptic scan has more power to detect elongated clusters, compared to the circular scan statistic.

Duczmal, Kulldorff and Huang (2006) developed a geometric penalty for irregularly shaped clusters. Many algorithms frequently produce a solution that is nothing more than the collection of the highest incidence cells in the map, linked together forming a tree-shaped cluster spread through the map; the associated subgraph resembles a tree, except possibly for a few additional edges. This kind of cluster does not add new information with regard to its special geographical significance in the map. One easy way to avoid this problem is simply to set a smaller upper bound on the maximum number of cells within a zone. This approach is only effective when the cluster size is rather small (i.e., for detecting those clusters occupying roughly up to 10% of the cells of the map). For larger upper bounds in size, the increased geometric freedom favors the occurrence of very irregularly shaped tree-like clusters, thus impacting the power of detection. Another way to deal with this problem is to have some shape control for the zones that are being analyzed, penalizing the zones in the map that are highly irregularly shaped. For this purpose the geometric compactness of a zone is defined as the area of z divided by the circle with the perimeter of the convex hull of z. Compactness is dependent on the shape of the object, but not on its size. Compactness also penalizes a shape that has a small area compared to the area of its convex hull. A user-defined exponent α is attached to the penalty to control its strength; larger values of α increase the effect of

the penalty, allowing the presence of more compact clusters. Similarly, lower α values allow more freedom of shape. The idea of using a penalty function for spatial cluster detection, based on the irregularity of its shape, was first used for ellipses [Kulldorff *et al.* (2006)], although a different formula was employed.

The greedy algorithm idea was used by Yiannakoulias, Rosychuk and Hodgson (2007) to explore the space of all possible configurations. A new penalty function is now defined as the ratio of the number of edges $e(Z)$ to the total possible number of edges in the candidate cluster Z. The total possible number of edges is computed as $3(v(Z)-2)$ based solely on the number of vertices $v(Z)$ in the candidate cluster. The non-connectivity penalty is employed as an exponent to the LR, analogously to the geometric compactness penalty. In the same way, a user-defined exponent α is attached to the non-connectivity penalty to control its strength. Instead of stopping the candidate clusters' aggregation process before reaching a prespecified population proportion limit, another criterion is used, based on the failure to increase the LR to a higher value after a certain number u of steps. The parameter u is set by the user; larger values of u relax the search constraint, and making $u = 0$ halts the search when no vertices can be added that increase the LR. Although the non-connectivity penalty is in many ways similar to the geometric compactness penalty, it has an important difference: it does not rely on the geometric shape of the candidate cluster, which could be an interesting advantage when searching for real clusters that are highly irregularly shaped, but present good connectivity properties.

Conley, Gahegan and Macgill (2005) proposed a genetic algorithm to explore a configuration space of multiple agglomerations of ellipses for point data sets. The method employed a strategy to clean up the best configuration found in order to geometrically simplify the cluster.

Sahajpal, Ramaraju and Bhatt (2004) also used a genetic algorithm to find clusters shaped as intersections of circles of different sizes and centers in point data sets.

Duczmal *et al.* (2007) described a genetic algorithm scan for the detection and inference of irregularly shaped spatial clusters. Assuming a map divided into regions with given populations at risk and cases, the graph-related operations are minimized by means of a fast offspring generation and evaluation of Kulldorff's spatial scan statistic. The penalty function of Duczmal, Kulldorff and Huang (2006), based on the geometric non-compactness concept, is employed to avoid excessive irregularity of cluster geometric shape. This algorithm is an order of magnitude faster and exhibits less variance compared to the SA scan, and it is more flexible than the elliptic scan. It has about the same power of detection as the SA scan for mildly irregular clusters and is superior for the very irregular ones.

The oblique decision tree (ODT) of Gaudart *et al.* (2005) was a modification of the classification and regression tree (CART) strategy to obtain an optimal

partitioning procedure in order to detect spatial patterns and find the candidate clusters without prior specifications. Instead of using rectangular partitions of the covariate space as in CART, ODT provides oblique partitions maximizing the interclass variance of the independent variable, providing polygonal candidate clusters. Classical ODT algorithms in R^n rely on evolutionary algorithms or heuristics, but in this work an optimal ODT algorithm is developed in R^2, based on the directions defined by each couple of point locations. The procedure consists in finding several partitions of the plane. The first step finds the best oblique split of the plane between two adjacent classes, maximizing the interclass variance. Operating recursively, this algorithm will split the plane into several partitions, until it reaches a specific stopping criterion. Monte Carlo replications are used to test significance. The ODT is compared with Kulldorff's spatial scan.

Multi-resolution methods (MR) [Neill and Moore (2003) and Neill and Moore (2004)] maximized Kulldorff's scan statistic over the square regions S of a grid of $g \times g$ squares, each one with an assigned number of cases and controls. Instead of using a naïve approach, which would require $O(g^3)$ calculations (multiplied by R Monte Carlo replications), the MR algorithm partitions the grid into overlapping regions, bounds the maximum score of subregions contained in each region and prunes regions which cannot contain the maximum density region. The maximum density region is found using $O(g^2)$, for sufficiently dense regions. Neill *et al.* (2005) later introduced another algorithm, the fast spatial scan (FS), generalizing the original bidimensional MR to arbitrary dimensions and using rectangles instead of squares. Applications include multiple data streams in syndromic surveillance (emergency department visits and over-the-counter drug sales) and discovery of regions of increased brain activity corresponding to given cognitive tasks (from functional magnetic resonance imaging data).

Given n baseline and case points, Agarwal *et al.* (2006) presented an algorithm to compute exactly the maximum discrepancy rectangle in time $O(n^4)$. If the points lie in a $g \times g$ grid, the algorithm runs in time $O(g^4)$. This algorithm has the same asymptotic running time as the MR algorithm. A much better performance is achieved for the general family of *discrepancy functions* (including Kulldorff's scan), through the approx-linear Algorithm (AL) by representing the discrepancy function as the upper envelope of a collection of linear functions. It is shown that a thoroughly linear approximation of the discrepancy function, which would require many linear functions, is not strictly necessary, because the approximation needs only to preserve the ordering of points along the direction of the search. As a result, a much better algorithm can maximize the discrepancy function over axis parallel rectangles in time $O(n^2 \log n)$. The algorithm is also extended to aggregate data sets using a regular $g \times g$ grid. A further technique is presented, using sampling to compute an approximation to the maximum linear discrepancy.

Aldstadt and Getis (2006) proposed the AMOEBA (Multidirectional Optimum Ecotope-Based Algorithm). An *ecotope* or *habitat* is defined in the literature as a specialized region within a larger region. A local spatial autocorrelation statistic is employed to construct a spatial weights matrix, used to describe the association between contiguous spatial units. The weights matrix is used in the determination of the geometric form of spatial clusters. It searches for spatial association in all specified directions, starting from a selected collection of seed spatial units. The main objective is to identify the ecotopes, the spatially homogeneous subregions within the study area. AMOEBA is compared with SaTScan.

Duczmal, Cançado and Takahashi (2008) proposed an approach to the geographic delineation of irregularly shaped disease clusters, treating it as a multi-objective optimization problem. Irregularly shaped spatial disease clusters occur commonly in epidemiological studies, but their geographic delineation is poorly defined. Most current spatial scan software usually displays only one of the many possible cluster solutions with different shapes, from the most compact round cluster to the most irregularly shaped one, corresponding to varying degrees of penalization parameters imposed on the freedom of shape. Even when a fairly complete set of solutions is available, the choice of the most appropriate parameter setting is left to the practitioner, whose decision is often subjective. A quantitative criterion for choosing the best cluster solution is presented, simultaneously maximizing two competing objectives: regularity of shape ($K(z)$), and scan statistic value (LLR). The Pareto set is defined as the set of all cluster candidates z such that no other cluster has both higher LLR and higher regularity than z. For each value of $K(z)$, a separate empirical distribution of LLR under the null hypothesis is computed, constituting a two-dimensional p-value surface. The cluster with the lowest p-value is considered the most likely cluster. Instead of running a cluster-finding algorithm with varying degrees of penalization, the complete set of solutions is found in parallel, through a genetic algorithm. The p-value surface is computed using Gumbel approximations [Abrams, Kulldorff and Kleinman (2006)]. Although different shapes are dealt with simultaneously, multiple testing does not occur, since the null hypothesis maps also produce Pareto sets using exactly the same algorithm as the observed cases map. The introduction of the concept of Pareto set in this problem, followed by the choice of the most significant solution, is shown to allow a rigorous statement about what is such a 'best solution', without the need of arbitrary parameters.

Maps with irregularly shaped or multiple clustering, when there is not a clearly dominating primary cluster, occur frequently. Moura *et al.* (2007) developed a method to analyze more thoroughly the several levels of clustering that arise naturally in a disease map divided into m regions. Instead of using a genetic algorithm, this method incorporates the simplicity and speed of the

circular scan, being able to detect and evaluate irregularly shaped clusters. The circular occupation (CO) of a cluster candidate is defined roughly as its population divided by the population inside the smallest circle containing it. The CO concept, computationally faster and relying on familiar concepts, substitutes here the compactness definition as the measure of regularity of shape. A multi-objective modification of the circular scan algorithm is applied, using CO and LLR as the objectives. The comparison of Pareto sets for observed cases with those computed under the null hypothesis provides valuable hints for the spatial occurrence of diseases. The potential for monitoring incipient spatial-temporal clusters at several geographic scales simultaneously is a promising tool in syndromic surveillance, especially for contagious diseases when there is a mix of short- and long-range spatial interactions. The presence of knees in the Pareto sets indicates sudden transitions in the clusters structure, corresponding to rearrangements due to the coalescence of loosely knitted (usually disconnected) clusters.

Yiannakoulias *et al.* (2007) employed quad trees to generate non-uniform grid points in order to detect spatial clusters in study areas provided with a large number of points. This strategy is compared with another scheme, which uses uniform grid points. The quad tree approach is more sensitive to high-resolution spatial clusters and is also more flexible, compared with the uniform grid approach.

Boscoe (2003) proposed a tool to visualize relative risk and statistical significance simultaneously. Given a map of n regions, with their respective centroids, the procedure builds a grid of equidistant points between all combinations of two, three and four adjacent region centroids. For each grid point the distances to the region centroids are computed and sorted. These distances are used to define almost circular groupings of regions, with their respective cumulative numbers of observed and expected cases. The relative risk and the LLR are then calculated for each circular grouping. The LLR values are compared to the results of a Monte Carlo simulation under the null hypothesis. Groupings with LLR values exceeding 95% of those obtained from the simulation are stored and stratified into ten levels of relative risk. Within each risk level, the grouping with the largest LLR is then mapped. Circular groupings with lower LLR are also mapped if they did not overlap any grouping previously mapped. The final result is a ten-color-shaded map of regions with statistically significant relative risks, providing a very effective visualization tool to grasp these two concepts.

There exist many methods to detect boundaries and to detect clusters; Jacquez, Kaufmann and Goovaerts (2007) proposed the b-statistic as a tool for the simultaneous detection of boundaries and clusters. It evaluates boundaries between adjacent areas with different values, and also the existing links between adjacent areas with similar values. Clusters are constructed by joining similarly high valued areas, which are then connected through a link. Unlike the

local Moran and other statistics, which describe local spatial variation in the immediate local neighborhood about a central location, the b-statistic describes properties of the edge between two areas. The b-statistic was compared with polygon wombling [cf. Womble (1951)] for detecting boundaries and the local Moran test [Moran (1948) and Moran (1950)].

Haiman and Preda (2002) derived approximations for the estimation of the distribution of scan statistics for a two-dimensional Poisson process. Through extensive numerical tests, Abrams, Kulldorff and Kleinman (2006) showed that, under the null hypothesis, the empirical distribution of values of Kulldorff's scan statistic for circular clusters is approximated by the well-known Gumbel distribution. The authors calculated that, using this semi-parametric approach, 100 Monte Carlo replications suffice to provide the same accuracy in significance estimation as 10,000 replications using the usual empirical distribution.

Kulldorff, Tango and Park (2003) presented a large collection of simulated benchmark data sets generated under different cluster models and the null hypothesis, to be used for power evaluations. These data sets are used to compare the power of the spatial scan statistic, the maximized excess events test and the non-parametric M statistic.

Duczmal *et al.* (2007) described a graph-based model for cluster detection and inference on networks based on the scan statistic. Nodes, associated to cities, are linked by means of edges, which represent routes between cities. Instead of forming cluster candidates by grouping neighboring nodes of the original graph, the cluster candidates are chosen among the connected subgraphs of the dual graph. The objective is to find collections of plausible pathways by which the disease could be transmitted. The most likely cluster is naturally the most structurally stable connected subgraph, or arrangement of pathways, meaning that adding or subtracting pathways to it should decrease the observed signal-to-noise proportion. In this model, traffic between cities is analogous to population in the usual scan, and the number of syndromic individuals traveling between cities corresponds to the number of cases.

The prospective time periodic scan [Kulldorff (2001)] is a space-time scan statistic for regular time periodic disease surveillance to detect any active geographical clusters of disease. The statistical significance of such clusters is adjusted for multiple testing, taking account of all possible geographical locations and sizes, time intervals and time periodic analyses.

The pyramidal flexible shape space-time scan for point data sets proposed by Iyengar (2004), instead of building space-time cylinders, adopted the more flexible pyramid or cone shapes with its axis perpendicular to the space plane. It represents an advance over the usual cylindrical approach, because it is now possible to model emerging spatially growing or shrinking clusters over time.

Kulldorff *et al.* (2005) presented the space-time permutation scan statistics (STPSS) for outbreak detection in syndromic surveillance systems. Emerging

clusters are detected using cylinders of variable radius and height to scan the space-time region in order to select the candidate cluster with maximum likelihood. A data permutation procedure is executed through Monte Carlo simulations in order to estimate the p-value of the most likely cluster. This method does not require the previous knowledge of the population at risk. Costa, Kulldorff and Assunção (2007) extended the STPSS to detecting irregular space-time clusters.

7.3 Data-Driven Spatial Cluster Detection Models

In this section we review data-tailored algorithms for spatial cluster detection including censored survival data, spatial mobility, multiple data streams, parametric models different from the usual Poisson or Bernoulli distributions and non-parametric and learning models.

Cook, Gold and Li (2007) considered a spatial scan statistic for censored outcome data. In contrast to the traditional scan statistics, which usually require a complete specification of the model, this paper uses a statistic score of the model of proportional risks to allow more flexibility. Cluster significance is estimated through permutation tests.

Huang, Kulldorff and Gregorio (2007) proposed a spatial scan statistic based on an exponential model to include uncensored or censored continuous survival data. The method achieves good power and sensitivity for several survival distribution functions including the exponential, gamma and log-normal distributions. Huang *et al.* (2007) applied the previous methodology to investigate possible relationships between the cluster locations and social and health conditions using non-parametric methods, and to compare socioeconomic factors inside and outside of the detected clusters and evaluate the effect of related covariates on significant long- and short-survival detected clusters.

Kulldorff *et al.* (2007) proposed the multivariate scan statistic. Frequently, more than one data stream may be available in disease surveillance systems. When analyzed separately instead of combined, the power of detection of an outbreak signal that is present in all data streams may diminish due to low counts in each. Besides, the simple summation of all data stream counts may obliterate a signal that is primarily present in just one data stream, due to random noise present in the other data sets. These two problems are tackled by defining an extension of the space-time scan statistic as the sum of the individual log likelihoods for those data sets for which the observed case count is more than the expected.

The multivariate Bayesian scan statistic (MBSS) of Neill, Moore and Cooper (2007) proposed modeling different outbreak types employing multiple data

streams. However, this approach uses fixed methods and models for analysis, and cannot improve their performance over time. Neill and Makatchev (2007) incorporated machine learning algorithms in the MBSS system. Two methods were devised for overcoming this limitation, employing a priori information over outbreak regions and learning outbreak models from user feedback. The authors demonstrate through simulations that learning can enable systems to improve detection performance over time.

Motivated by the fact that the regions inside a cluster candidate are not homogeneous, Takahashi and Tango (2007) proposed an alternative scan statistic that can take the variability of the relative risks of regions included in Z into account, employing Anscombe's variance stabilization transformation.

Tango (2007) proposed a modified likelihood ratio test statistic which accounts for each individual region's risk. This modified scan includes an indicator variable based on the p-value for the zone consisting of the individual region i. Given a prespecified α_1 p-value, and if p_i is the p-value of the zone consisting of the individual region i, then the modified LR scan for a cluster including i is taken as zero when $p_i > \alpha_1$.

Neill and Moore (2006) presented the expectation-based scan statistic (EBSS) as an extension of the usual spatial and space-time scan statistics by inferring expected counts for each location from past data and detecting regions where recent counts are higher than expected. Neill and Lingwall (2007) presented the nonparametric scan statistic (NPSS), a general detector of space-time clusters in syndromic surveillance using multiple data streams. It does not assume a parametric model, but instead combines empirical p-values across multiple locations, days and data streams to detect anomalies.

A discrete event model was used by Beeker, Bauer and Mohtashemi (2007) to simulate the spread of infectious diseases through an agent-based, stochastic model of transmission dynamics. The objective is to generate a benchmark from a network of individual contacts in an urban environment using publicly available population data. This benchmark can be used to test the performance of various temporal and spatio-temporal detection algorithms when real data are not available or cannot be used due to confidentiality issues.

Duczmal and Buckeridge (2006) have derived an extension to the spatial scan statistic that accounts for the mobility of individuals between home address and workplace. An analyst can use the workflow scan statistic to search for disease clusters due to workplace exposure when health records contain only the residential address. The effect of the workflow scan statistic is to pull back the scattered workers that were contaminated in the workplace. Simulation studies demonstrate that in most scenarios, the workflow scan statistic has greater power than the usual scan statistic for detecting disease outbreaks due to workplace exposures. The workflow scan statistic is particularly useful when clusters are not circularly symmetrical, and thus more easily recognized by the workflow scan than by the usual spatial scan algorithm.

Cami, Wallstrom and Hogan (2007) presented a refinement of a Bayesian algorithm used for aerosol detection (BARD) incorporating a model that includes the mobility of the individuals. The population is subdivided into groups based on the residential and workplace information.

Local, global and focused tests were developed by Jacquez *et al.* (2005) to evaluate clustering in case-control data that take into account individual mobility. Matrices of nearest neighbor relationships are employed to represent the changing topology of cases and controls. The model includes the latency between exposure and disease manifestation. Jacquez *et al.* (2006) analyzed case-control clustering with individual mobility accounting for risk factors and covariates. Meliker and Jacquez (2007) extended those previous ideas to space-time clustering of case-control data with individual mobility. Using the Q-statistic, a statistic that includes time-dependent nearest-neighbors, the authors evaluate empirical induction periods, age-specific susceptibility and calendar year-specific effects.

Zhang and Lin (2007) presented a decomposition of Moran's I test into three components so that each component represents a global test statistic. The three components test for the existence of high-value clustering low-value clustering and negative autocorrelation. A set of simulations shows that the first test statistic is likely to be significant only for high-value clustering, the second test statistic is likely to be significant only for low-value clustering and the last test statistic is likely to be significant only for negatively correlated spatial structures. Two real data examples were studied, and in both cases low-value clustering and high-value clustering were shown to exist simultaneously.

Lin and Zhang (2007) combined the permutation test of Moran's I to the residuals of a log-linear model under the asymptotic normality assumption. It provides the versions of Moran's I based on Pearson residuals and deviance residuals so that they can be used to test for spatial clustering while at the same time account for potential covariates and heterogeneous population sizes.

Aggregation is commonly used as a mask to protect health data confidentiality of individuals. Ozonoff *et al.* (2007) studied the association between spatial resolution and power of detection through thousands of simulations with the spatial scan statistic. Power to detect clusters decreased from nearly 100% when using exact locations to roughly 40% at the coarsest level of spatial resolution. The authors conclude that aggregation has the potential to obliterate existing clusters.

The usefulness of individual-level health data point locations in providing high quality data for epidemiological research must be balanced with the easiness of breaking the confidentiality of the identities of the individuals. Geographic masking is being employed as a tool for achieving an appropriate balance between data utility and confidentiality. Usually the masks employ perturbation, aggregation of areas and a combination of both. Zimmerman and

Pavlik (2008) discussed whether certain characteristics of the mask (mask metadata) should be disclosed to data users and whether two or more distinct masked versions of the data can be released without breaching confidentiality.

Glaz and Zhang (2006) defined a maximum scan score-type statistic for testing the null hypotheses that the observed data are independent and identically distributed according to a specified distribution, against a class of window clustering-type alternatives. The maximum scan score-type statistic detects clustering effectively in the situation where the window size is unknown. The extension to multivariate data is discussed by the authors.

In disease surveillance, anomalies may be detected either by computing confidence intervals for region rates or by running a disease cluster detection algorithm. Rosychuk (2006) attempts to determine when those two approaches give the same answers. The study compared Besag and Newell's (1991) cluster detection method with confidence intervals for crude and directly standardized rates. Simulations suggest that the cluster detection method is preferred when the cluster size exceeds the number of cases in a region or when the expected number of cases exceeds a threshold.

In some situations of disease surveillance, it is preferable to use disease-related events instead of individuals as the units of analysis.

Rosychuk, Huston and Prasad (2006) proposed a compound Poisson method that detects event clusters by testing individual areas that may be combined with their nearest neighbors. This technique is useful where the population sizes are diverse and the population distribution by important strata may differ by area.

Song and Kulldorff (2003) compared the statistical power of several disease clustering tests: Besag–Newell's R, Cuzick–Edwards' k-nearest neighbors (k-NN), the spatial scan statistic, Tango's maximized excess events test (MEET), Swartz' entropy test, Whittemore's test, Moran's I and a modification of Moran's I. Except for Moran's I and Whittemore's test, all other tests have good power for detecting some kind of clustering. The spatial scan statistic is good at detecting localized clusters. Tango's MEET is good at detecting global clustering. With appropriate choice of parameter, Besag–Newell's R and Cuzick–Edwards' k-NN also perform well.

Aamodt, Samuelsen and Skrondal (2006) conducted a simulation study to compare three methods: SaTScan, generalized additive models and Bayesian disease mapping.

Ozdenerol *et al.* (2005) compared the results of Kulldorff's spatial scan statistic with the results of Rushton's spatial filtering technique through increasing sizes of spatial filters.

7.4 Applications

We finish this review by providing a short list of interesting applications. Of course, this list is not exhaustive. Its only purpose is to illustrate some of the ideas presented in the previous sections, in disease surveillance [Croner and De Cola (2001), Dunyak, Mohtashemi and Mandl (2006), Heffernan *et al.* (2004), Gardner, Strickland and Correa (2007), Grannis *et al.* (2007), Nordin *et al.* (2005), Gunn, Pendarvis and Barry (2006), Sabhnani, Neill and Moore (2005), Goranson *et al.* (2006) and Johnson *et al.* (2005)], terrorism surveillance [Porter and Brown (2007)] and epidemiology [Ali *et al.* (2006), Brooker *et al.* (2004), Dunchin (2003), Chaput, Meek and Heimer (2002), Durand and Wilson (2006), Viel, Floret and Mauny (2005), Kulldorff *et al.* (1998), Fukuda *et al.* (2005), Onozuka and Hagihara (2007), Goovaerts, Jacquez and Greiling (2005), Kulldorff *et al.* (1997), Mather *et al.* (2006), Moore (2005), Myers *et al.* (2006), Norström, Pfeiffer and Jarp (2000), Nunes (2007), Oliver *et al.* (2006), Hanson and Wieczorek (2002), Ozonoff *et al.* (2005), McNally and Colver (2008), Perez *et al.* (2005), Sabel *et al.* (2002), Sanchez *et al.* (2005), Tiwari *et al.* (2006), Ward and Carpenter (2000), Washington *et al.* (2004), Wheeler (2007)].

Acknowledgment

This work was supported by CNPq, Brazil.

References

1. Aamodt, G., Samuelsen, S.O. and Skrondal, A. (2006). A simulation study of three methods for detecting disease clusters, *International Journal of Health Geographics*, **5**, 15.

2. Abrams, A.M., Kulldorff, M. and Kleinman, K. (2006). Empirical/ asymptotic p-values for Monte Carlo-based hypothesis testing: an application to cluster detection using the scan statistic, *Advances in Disease Surveillance*, **1**, 1.

3. Agarwal, D., McGregor, A., Venkatasubramanian, S. and Zhu, Z (2006). Spatial Scan Statistics Approximations and Performance Study, *Conference on Knowledge Discovery in Data Mining 2006.*

4. Aldstadt, J. and Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters, *Geographical Analysis*, **38**, 327–343.

5. Ali, M., Goovaerts, P., Nazia, N., Haq, M.Z., Yunus, M. and Emch, M. (2006). Application of Poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh, *International Journal of Health Geographics*, **5**, 45.

6. Assunção, R.M., Costa, M.A., Tavares, A. and Neto, S.J.F. (2006). Fast detection of arbitrarily shaped disease clusters, *Statistics in Medicine*, **25**, 723–742.

7. Balakrishnan, N. and Koutras, M.V. (2002). *Runs and Scans with Applications*, John Wiley & Sons, New York.

8. Beeker, E., Bauer, D.W. and Mohtashemi, M. (2007). Benchmark data generation from discrete event contact network models, *Advances in Disease Surveillance*, **4**, 235.

9. Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society*, **A154**, 143–155.

10. Boscoe, F.P. (2003). Visualization of the spatial scan statistic using nested circles, *Health & Place*, **9**, 273–277.

11. Brooker, S., Clarke, S., Njagi, J.K., Polack, S., Mugo, B., Estambale, B., Muchiri, E., Magnussen, P. and Cox, J. (2004). Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya, *Tropical Medicine and International Health*, **9**, 757–766.

12. Buckeridge, D.L., Burkom, H., Campbell, M., Hogan, W.R. and Moore, A.W. (2005). Algorithms for rapid outbreak detection: a research synthesis, *Journal of Biomedical Informatics*, **38**, 99–113.

13. Cami, A., Wallstrom, G.L. and Hogan, W.R. (2007). Effect of work-related mobility in the simulation of aerosol anthrax releases with BARD, *Advances in Disease Surveillance*, **4**, 239.

14. Chaput, E.K., Meek, J.I. and Heimer, R. (2002). Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut, *Emerging Infectious Diseases*, **8**, 943–948.

15. Conley, J., Gahegan, M. and Macgill, J. (2005). A genetic approach to detecting clusters in point data sets, *Geographical Analysis*, **37**, 286–314.

16. Cook, A.J., Gold, D.R. and Li, Y. (2007). Spatial cluster detection for censored outcome data, *Biometrics*, **63**, 540–549.

17. Costa, M.A., Kulldorff, M. and Assunção, R.M. (2007). A space time permutation scan statistic with irregular shape for disease outbreak detection, *Advances in Disease Surveillance*, **4**, 243.

18. Croner, C.M. and De Cola, L. (2001). Visualization of Disease Surveillance Data with Geostatistics, *Work Session on Methodological Issues Involving Integration of Statistics and Geography.*

19. Dematteï, C., Molinari, N. and Daurès, J.P. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data, *Computational Statistics & Data Analysis*, **51**, 3931–3945.

20. Duczmal, L. and Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, *Computational Statistics & Data Analysis*, **45**, 269–286.

21. Duczmal, L. and Buckeridge, D.L. (2006). A workflow spatial scan statistic, *Statistics in Medicine*, **25**, 743–754.

22. Duczmal, L., Kulldorff, M. and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters, *Journal of Computational & Graphical Statistics*, **15**, 428–442.

23. Duczmal, L., Cançado, A.L.F. and Takahashi, R.H.C. (2008). Geographic delineation of disease clusters through multi-objective optimization, *Journal of Computational & Graphical Statistics*, **17**, 243–262.

24. Duczmal, L., Cançado, A.L.F., Takahashi, R.H.C. and Bessegato, L.F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics, *Computational Statistics & Data Analysis*, **52**, 43–52.

25. Duczmal, L., Moreira, G.J.P., Ferreira, S.J. and Takahashi, R.H.C. (2007). Dual graph spatial cluster detection for syndromic surveillance in networks, *Advances in Disease Surveillance*, **4**, 88.

26. Dunchin, J.S. (2003). Epidemiological response to syndromic surveillance signals, *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, **80**, 115–116.

27. Dunyak, J., Mohtashemi, M. and Mandl, K. (2006). Temporal-spatial surveillance techniques from non-homogenous random geometric graphs, *Advances in Disease Surveillance*, **1**, 23.

28. Durand, M. and Wilson, G. (2006). Spatial analysis of respiratory disease on an urbanized geothermal field, *Environmental Research*, **101**, 238–245.

29. Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses, *Annals of Mathematical Statistics*, **28**, 181–187.

30. Elliott, P., Martuzzi, M. and Shaddick, G. (1995). Spatial statistical methods in environmental epidemiology: a critique, *Statistical Methods in Medical Research*, **4**, 137–159.

31. Fukuda, Y., Umezaki, M., Nakamura, K. and Takano, T. (2005). Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan, *International Journal of Health Geographics*, **4**, 16.

32. Gardner, B.R., Strickland, M.J. and Correa, A. (2007). Application of the automated spatial surveillance program to birth defects surveillance data, *Birth Defects Research Part A: Clinical and Molecular Teratology*, **79**, 559–564.

33. Gaudart, J., Poudiougou, B., Ranque, S. and Doumbo, O. (2005). Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk, *BMC Medical Research Methodology*, **5**, 22.

34. Glaz, J. and Zhang, Z. (2006). Maximum scan score-type statistics, *Statistics & Probability Letters*, **76**, 1316–1322.

35. Glaz, J., Naus, J. and Wallenstein, S. (2001). Scan Statistics, in *Springer Series in Statistics*, Springer, Berlin-Heidelberg-New York.

36. Goovaerts, P., Jacquez, G.M. and Greiling, D. (2005). Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms, *Geographical Analysis*, **37**, 152–182.

37. Goranson, C., Konty, K., Lu, J. and Mostashari, F. (2006). Visualization of syndromic surveillance using GIS, *Advances in Disease Surveillance*, **1**, 26.

38. Grannis, S., Egg, J., Cassa, C.A., Olson, K., Mandl, K. and Overhage, J.M. (2007). Evaluating the performance of a spatial scan statistic using simulated outbreak characteristics, *Advances in Disease Surveillance*, **2**, 200.

39. Gunn, J., Pendarvis, J. and Barry, A. (2006). Syndromic surveillance and zip code data: the role of zip codes in understanding populations, *Advances in Disease Surveillance*, **1**, 29.

40. Haiman, G. and Preda, C. (2002). A new method for estimating the distribution of scan statistics for a two-dimensional poisson process, *Methodology and computing in Applied Probability*, **4(4)**, 393–407.

41. Hanson, C.E. and Wieczorek, W.F. (2002). Alcohol mortality: a comparison of spatial clustering methods, *Social Science & Medicine*, **55**, 791–802.

42. Heffernan, R., Mostashari, F., Das, D., Karpati, A., Kulldorff, M. and Weiss, D. (2004). Syndromic surveillance in public health practice, New York City, *Emerging Infectious Diseases*, **10**, 858–864.

43. Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data, *Biometrics*, **63**, 109–118.

44. Huang, L., Pickle, L.W., Stinchcomb, D. and Feuer, E.J. (2007). Detection of spatial clusters: application to cancer survival as a continuous outcome, *Epidemiology*, **18**, 73–87.

45. Iyengar, V.S. (2004). Space-Time Clusters with Flexible Shapes, *IBM Research Report RC23398 (W0408-068)*.

46. Jacquez, G.M., Kaufmann, A., Meliker, J., Goovaerts, P., AvRuskin, G. and Nriagu, J. (2005). Global, local and focused geographic clustering for case-control data with residential histories, *Environmental Health: A Global Access Science Source*, **4**, 4.

47. Jacquez, G.M., Meliker, J., AvRuskin, G., Goovaerts, P., Kaufmann, A., Wilson, M. and Nriagu, J. (2006). Case-control geographic clustering for residential histories accounting for risk factors and covariates, *International Journal of Health Geographics*, **5**, 32.

48. Jacquez, G.M., Kaufmann, A. and Goovaerts, P. (2007). Boundaries, links and clusters: a new paradigm in spatial analysis? *Environmental and Ecological Statistics* (Published online).

49. Johnson, G.D., Eidson, M., Schmit, K., Ellis, A. and Kulldorff, M. (2005). Geographic prediction of human onset of West Nile virus using dead crow clusters: an evaluation of year 2002 data in New York State, *American Journal of Epidemiology*, **163**, 171–180.

50. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26(6)**, 1481–1496.

51. Kulldorff, M. (1999). Spatial Scan Statistics: Models, Calculations, and Applications, in *Scan Statistics and Applications* (Ed., N. Balakrishnan and J. Glaz), pp. 303–322, Birkhäuser.

52. Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society*, **164(1)**, 61–72.

53. Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine*, **14**, 799–810.

54. Kulldorff, M., Feuer, E.J., Miller, B.A. and Freedman, L.S. (1997). Breast cancer clusters in the northeast United States: a geographic analysis, *American Journal of Epidemiology*, **146**, 161–170.

55. Kulldorff, M., Athas, W.F., Feuer, E.J., Miller, B.A. and Key, C.R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico, *American Journal of Public Health*, **88**, 1377–1380.

56. Kulldorff, M., Tango, T. and Park, P.J. (2003). Power comparisons for disease clustering tests, *Computational Statistics & Data Analysis*, **42**, 665–684.

57. Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R.M. and Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection, *PLoS Medicine*, **2(3)**, 216–224.

58. Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic, *Statistics in Medicine*, **25**, 3929–3943.

59. Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K. and Platt, R. (2007). Multivariate scan statistics for disease surveillance, *Statistics in Medicine*, **26**, 1824–1833.

60. Lawson, A. (2001). In Large Scale: Surveillance. *Statistical Methods in Spatial Epidemiology*, ISBN: 0471975729 (Ed., A. Lawson), 197–206, Wiley.

61. Lawson, A., Biggeri, A., Böhning, D., Lesare, E., Viel, J.F. and Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*, Wiley, London.

62. Lin, G. and Zhang, T. (2007). Loglinear residual tests of Moran's I autocorrelation and their applications to Kentucky breast cancer data, *Geographical Analysis*, **39**, 293–310.

63. McNally, R.J.Q. and Colver, A.F. (2008). Space-time clustering analyses of occurrence of cerebral palsy in northern England for births 1991 to 2003, *Annals of Epidemiology*, **18**, 108–112.

64. Mather, F.J., Chen, V.W., Morgan, L.H., Correa, C.N., Shaffer, J.G., Srivastav, S.K., Rice, J.C., Blount, G., Swalm, C.M., Wu, X. and Scribner, R.A. (2006). Hierarchical modeling and other spatial analyses in prostate cancer incidence data, *American Journal of Preventive Medicine*, **30**, 88–100.

65. Meliker, J.R. and Jacquez, G.M. (2007). Space-time clustering of case-control data with residential histories: insights into empirical induction periods, *Journal of Stochastic Environmental Research & Risk Assessment*, **21**, 625–634.

66. Modarres, R. and Patil, G.P. (2007). Hotspot detection with bivariate data, *Journal of Statistical Planning and Inference*, **137**, 3643–3654.

67. Moore, D.A. and Carpenter, T.E. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology, *Epidemiologic Reviews*, **21**, 143–161.

68. Moore, G.E. (2005). A space-time cluster of adverse events associated with canine rabies vaccine, *Vaccine*, **23**, 5557–5562.

69. Moran, P.A.P. (1948). The interpretation of statistical maps, *Journal of the Royal Statistical Society*, **10**, 243–251.

70. Moran, P.A.P. (1950). A test for the serial independence of residuals, *Biometrika*, **37**, 178–181.

71. Moura, F.R., Duczmal, L., Tavares, R. and Takahashi, R.H.C. (2007). Exploring multi-cluster structures with the multi-objective circular scan, *Advances in Disease Surveillance*, **2**, 48.

72. Myers, W.L., Kurihara, K., Patil, G.P. and Vraney, R. (2006). Finding upper-level sets in cellular surface data using echelons and SaTScan, *Environmental and Ecological Statistics*, **13**, 379–390.

73. Naus, J.I. (1965). Clustering of random points in two dimensions, *Biometrika*, **52**, 263–267.

74. Neill, D.B. and Moore, A.W. (2003). A Fast Multi-Resolution Method for Detection of Significant Spatial Overdensities, *Carnegie Mellon CSD Technical Report*.

75. Neill, D.B. and Moore, A.W. (2004). A fast multi-resolution method for Detection of Significant Spatial Disease Clusters, *Advances in Neural Information Processing Systems*, **16**, 651–658.

76. Neill, D.B., Moore, A.W., Pereira, F. and Mitchell, T. (2005). Detecting significant multidimensional spatial clusters, *Advances in Neural Information Processing Systems*, **17**, 969–976.

77. Neill, D.B. and Moore, A.W. (2006). Methods for detecting spatial and spatio-temporal clusters, *Handbook of Biosurveillance*, Elsevier, Amsterdam, 243–254.

78. Neill, D.B., Moore, A.W. and Cooper, G.E. (2007). A multivariate Bayesian scan statistic, *Advances in Disease Surveillance*, **2**, 60.

79. Neill, D.B. and Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance, *Advances in Disease Surveillance*, **4**, 106.

80. Neill, D.B. and Makatchev, M. (2007). Incorporating learning into disease surveillance systems, *Advances in Disease Surveillance*, **4**, 107.

81. Nordin, J.D., Goodman, M.J., Kulldorff, M., Ritzwoller, D.P., Abrams, A.M., Kleinman, K., Levitt, M.J., Donahue, J. and Platt, R. (2005). Simulated anthrax attacks and syndromic surveillance, *Emerging Infectious Diseases*, **11** 1394–1398.

82. Norström, M., Pfeiffer, D.U. and Jarp, J. (2000). A spacetime cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds, *Preventive Veterinary Medicine*, **47**, 107–119.

83. Nunes, C. (2007). Tuberculosis incidence in Portugal: spatiotemporal clustering, *International Journal of Health Geographics*, **5(1)**, 51.

84. Oliver, M.N., Smith, E., Siadaty, M., Hauck, F.R. and Pickle, L.W. (2006). Spatial analysis of prostate cancer incidence and race in Virginia, 1990-1999, *American Journal of Preventive Medicine*, **30**, 67–76.

85. Onozuka, D. and Hagihara, A. (2007). Geographic prediction of tuberculosis clusters in Fukuoka, Japan, using the space-time scan statistic, *BMC Infectious Diseases*, **7**, 26.

86. Ozdenerol, E., Williams, B.L., Kang, S.Y. and Magsumbol, M.S. (2005). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters, *International Journal of Health Geographics*, **4**, 19.

87. Ozonoff, A., Webster, T., Vieira, V., Weinberg, J., Ozonoff, D. and Aschengrau, A. (2005). Cluster detection methods applied to the Upper Cape Cod cancer data, *Environmental Health: A Global Access Science Source*, **4**, 19.

88. Ozonoff, A., Jeffery, C., Manjourides, J., White, L.F. and Pagano, M. (2007). Effect of spatial resolution on cluster detection: a simulation study, *International Journal of Health Geographics*, **6**, 52.

89. Patil, G.P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

90. Patil, G.P., Modarres, R., Myers, W.L. and Patankar, P. (2006). Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics, *Environmental and Ecological Statistics*, **13**, 365–377.

91. Perez, A.M., Thurmond, M.C., Grant, P.W. and Carpenter, T.E. (2005). Use of the scan statistic on disaggregated province-based data: foot-and-mouth disease in Iran, *Preventive Veterinary Medicine*, **71**, 197–207.

92. Porter, M.D. and Brown, D.E. (2007). Detecting local regions of change in high-dimensional criminal or terrorist point processes, *Computational Statistics & Data Analysis*, **51**, 2753–2768.

93. Rosychuk, R.J. (2006). Identifying geographic areas with high disease rates: when do confidence intervals for rates and a disease cluster detection method agree? *International Journal of Health Geographics*, **5**, 46.

94. Rosychuk, R.J., Huston, C. and Prasad, N.G.N. (2006). Spatial event cluster detection using a compound Poisson distribution, *Biometrics*, **62**, 465–470.

95. Sabel, C.E., Boyle, P.J., Lytnen, M., Gatrell, A.C., Jokelainen, M., Flowerdew, R. and Maasilta, P. (2002). Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death, *American Journal of Epidemiology*, **157**, 898–905.

96. Sabhnani, M.R., Neill, D.B. and Moore, A.W. (2005). Detecting anomalous patterns in pharmacy retail data, *Conference on Knowledge Discovery in Data Mining 2005*.

97. Sahajpal, R., Ramaraju, G.V. and Bhatt, V. (2004). Applying Niching Genetic Algorithms for Multiple Cluster Discovery in Spatial Analysis, *International Conference on Intelligent Sensing and Information Processing*.

98. Sanchez, J., Stryhn, H., Flensburg, M., Ersboll, A.K. and Dohoo, I. (2005). Temporal and spatial analysis of the 1999 outbreak of acute clinical infectious bursal disease in broiler flocks in Denmark, *Preventive Veterinary Medicine*, **71**, 209–223.

99. Song, C. and Kulldorff, M. (2003). Power evaluation of disease clustering tests, *International Journal of Health Geographics*, **2**, 9.

100. Takahashi, K. and Tango, T. (2007). A scan statistic based on Anscombe's variance stabilization transformation, *Advances in Disease Surveillance*, **4**, 116.

101. Takahashi, K., Kulldorff, M., Tango, T. and Yie, K. (2007). A flexible space-time scan statistic for disease outbreak detection and monitoring, *Advances in Disease Surveillance*, **2**, 70.

102. Tango, T. (2007). A spatial scan statistic scanning only the regions with elevated risk, *Advances in Disease Surveillance*, **4**, 117.

103. Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11.

104. Tiwari, N., Adhikari, C.M.S., Tewari, A. and Kandpal, V. (2006). Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic, *International Journal of Health Geographics*, **5**, 33.

105. Viel, J.F., Floret, N. and Mauny, F. (2005). Spatial and space-time scan statistics to detect low rate clusters of sex ratio, *Environmental and Ecological Statistics*, **12**, 289–299.

106. Yiannakoulias, N., Rosychuk, R.J. and Hodgson, J. (2007). Adaptations for finding irregularly shaped disease clusters, *International Journal of Health Geographics*, **6**, 28.

107. Yiannakoulias, N., Karosas, A., Schopflocher, D.P., Svenson, L.W. and Hodgson, M.J. (2007). Using quad trees to generate grid points for application in geographic disease surveillance, *Advances in Disease Surveillance*, **3**.

108. Waller, L.A. and Jacquez, G.M. (2000). Disease models implicit in statistical tests of disease clustering, *Epidemiology*, **6**, 584–590.

109. Ward, M.P. and Carpenter, T.E. (2000). Techniques for analysis of disease clustering in space and in time in veterinary epidemiology, *Preventive Veterinary Medicine*, **45**, 257–284.

110. Washington, C.H., Radday, J., Streit, T.G., Boyd, H.A., Beach, M.J., Addiss, D.G., Lovince, R., Lovegrove, M.C., Lafontant, J.G., Lammie, P.J. and Hightower, A.W. (2004). Spatial clustering of filarial transmission before and after a Mass Drug Administration in a setting of low infection prevalence, *Filaria Journal*, **3**, 3.

111. Wheeler, D.C. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003, *International Journal of Health Geographics*, **6**, 13.

112. Wieland, S.C., Brownstein, J.S., Berger, B. and Mandl, K.D. (2007). Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes, *PNAS*, **104(22)**, 904–909.

113. Womble, W.H. (1951). Differential systematics, *Science*, **114**, 315–322.

114. Zhang, T. and Lin, G. (2007). A decomposition of Moran's I for clustering detection, *Computational Statistics & Data Analysis*, **51**, 6123–6137.

115. Zimmerman, D.L. and Pavlik, C. (2008). Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data, *Geographical Analysis*, **40**, 52–76.

8

1-Dependent Stationary Sequences and Applications to Scan Statistics

George Haiman[1] **and Cristian Preda**[2]

[1] *UFR de Mathématiques, Université de Lille 1, Lille, France*
[2] *Faculté de Médecine, Université de Lille 2, Lille, France*

Abstract: A new method of estimating the distribution function of scan statistics was presented and studied by the authors in a series of papers. This method is based on the application of some results concerning the distribution function of the partial maximum sequence generated by a 1-dependent stationary sequence. We present a review of our results and compare the method with other existing methods.

Keywords and phrases: Scan statistic, 1-dependence, Poisson process

8.1 Introduction

Let N be a Poisson process of intensity λ on the real line and let $u > 0$ and $T > u$ be fixed constants. Let $\nu_t = N(t+u) - N(t)$ be the number of points in the interval $[t, t+u[$, $t \in [0, T-u]$.

The one-dimensional continuous scan statistic is defined [see Glaz *et al.* (2001)] as

$$S = S(u, \lambda, T) = \max_{0 \le t \le T-u} \nu_t. \tag{8.1}$$

Let $T = \tau u$, $\tau \in \mathbf{N}$ and let

$$X_n = \max_{(n-1)u \le t < nu} \nu_t, \;\; n = 1, \ldots, \tau - 1. \tag{8.2}$$

It can be easily seen that $\{X_n\}$ forms a 1-dependent stationary sequence and

$$S = S_\tau = \max_{1 \le n \le \tau - 1} X_n. \tag{8.3}$$

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_8,

Then, in order to approximate the distribution function (d.f.) of S, we can apply either one of the following equivalent versions of Haiman (1999), Theorems 3 and 4.

Let $\{X_n\}$ be a general 1-dependent stationary sequence of random variables (r.v.'s) and let

$$q_n = q_n(x) = \mathbf{P}\left\{\max(X_1, \ldots, X_n) \leq x\right\}, \ \ n \geq 1.$$

Theorem 8.1.1 *For any x such that $1 - q_1(x) \leq 0.025$ and any integer $n > 3$ such that $88n(1-q_1)^3 \leq 1$, we have*

$$\left| q_n - \frac{4q_3 - 3q_4 + 6(q_1 - q_2)^2}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^n} \right| \Big/ q_n \leq (1-q_1)^3[88n(1 + 124n(1-q_1)^3) + 561]. \tag{8.4}$$

Theorem 8.1.2 *For any x such that $1 - q_1(x) \leq 0.025$ and any integer $n > 3$ such that $3.3n(1-q_1)^2 \leq 1$, we have*

$$\left| q_n - \frac{2q_1 - q_2}{(1 + q_1 - q_2 + 2(q_1 - q_2)^2)^n} \right| \Big/ q_n \leq (1-q_1)^2[3.3n(1 + 4.7n(1-q_1)^2) + 9 + 561(1-q_1)]. \tag{8.5}$$

From Theorem 8.1.1 and Theorem 8.1.2 we deduce, respectively, the approximations

$$\mathbf{P}(S_\tau \leq x) \approx \frac{4q_3 - 3q_4 + 6(q_1 - q_2)^2}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^{\tau-1}} \tag{8.6}$$

with a relative error bound of about $88\tau(1-q_1)^3$ and

$$\mathbf{P}(S_\tau \leq x) \approx \frac{2q_1 - q_2}{(1 + q_1 - q_2 + 2(q_1 - q_2)^2)^{\tau-1}} \tag{8.7}$$

with a relative error bound of about $3.3\tau(1-q_1)^2$.

The approximations (8.6) and (8.7) for the d.f. of continuous scan statistics have been introduced and studied in Haiman (2000). A characteristic of these approximations is that they depend on a prior knowledge of $q_i = q_i(x) = \mathbf{P}(S_{i+1} \leq x)$, $i = 1, \ldots, 4$, respectively, $i = 1, 2$.

Let $Z_1, \ldots, Z_N$ be a sequence of integer-valued r.v.'s that are independent and identically distributed (i.i.d.), typically Bernoulli $\mathcal{B}(1, p)$. Let $1 \leq m \leq N$ be a fixed positive integer, let

$$\mu_t = \sum_{i=t}^{t+m-1} Z_i, i \leq t \leq N - m + 1, \tag{8.8}$$

and define the one-dimensional discrete scan statistic [see Glaz *et al.* (2001)] by

$$S = S(m, p, N) = \max_{1 \le t \le N-m+1} \mu_t. \tag{8.9}$$

Let $N = \tau m$, $\tau \in \mathbf{N}$, $\tau \ge 1$, and let

$$Y_n = \max_{(n-1)m+1 \le t \le nm+1} \mu_t, n \in \mathbf{N}, n \ge 1. \tag{8.10}$$

Then $\{Y_n\}$ similarly forms a stationary 1-dependent sequence, $S = S_\tau = \max_{1 \le n \le \tau-1} Y_n$ and the d.f. of S can again be approximated by either one of the corresponding versions of approximations (8.6) and (8.7).

This type of approximation for the d.f. of discrete scan statistics was introduced and studied in Haiman (2007).

In Section 8.2 we present and discuss the main aspects related to the application of approximations (8.6) and (8.7) to continuous and discrete one-dimensional scan statistics.

Let N be a two-dimensional Poisson process of intensity λ. For fixed positive u and v, let $\nu_{t,s}(u, v)$ be the number of points in the rectangle $[t, t+u) \times [s, s+u)$, i.e.,

$$\nu_{t,s} = \nu_{t,s}(u, v) = N([t, t + u) \times [s, s + v)). \tag{8.11}$$

For $0 < u < L$ and $0 < v < K$, the two-dimensional continuous scan statistic

$$S = S((u, v), \lambda, L, K) = \max_{\substack{0 \le t \le L-u \\ 0 \le s \le K-v}} \nu_{t,s} \tag{8.12}$$

represents the largest number of points in any rectangle of dimension $u \times v$ within the rectangular region $[0, L] \times [0, K]$. Observing that for any $0 < u < L$ and $0 < v < K$ we have

$$\mathbf{P}\left(S((u, v), \lambda, L, K) \le k\right) = \mathbf{P}\left(S((1, 1), \lambda uv, \frac{L}{u}, \frac{K}{v}) \le k\right),$$

we now suppose that $u = v = 1$.

Let K and L be positive integers and let

$$X_k = \max_{\substack{0 \le t \le L-1 \\ k-1 \le s \le k}} \nu_{t,s}, k = 1, \ldots, K-1. \tag{8.13}$$

We first observe that $\{X_k\}$ is a stationary 1-dependent sequence and

$$S = S_{L,K} = \max_{1 \le k \le K-1} X_k.$$

Then, a first application of Theorem 8.1.2 (under the required conditions) leads to the approximation

$$\mathbf{P}(S \leq n) \approx (2q_1 - q_2)(1 + q_1 + q_2 + 2(q_1 - q_2)^2)^{-(K-1)}, n \in \mathbf{N}. \quad (8.14)$$

with an error bound of about $3.3(K-1)(1-q_1)^2$. Here $q_1 = \mathbf{P}(X_1 \leq n)$ and $q_2 = \mathbf{P}(X_1 \leq n, X_2 \leq n)$. In order to obtain the final approximation of $\mathbf{P}(S \leq n)$, q_1 and q_2 are replaced in (8.14) by their approximations obtained using again Theorem 8.1.2. Indeed,

$$Y_l = \max_{\substack{l-1 \leq t \leq l \\ 0 \leq s \leq 1}} \nu_{t,s}, l = 1, \ldots, L-1 \quad (8.15)$$

is a 1-dependent stationary sequence and

$$q_1 = \mathbf{P}\left(\max_{0 \leq l \leq L-1} Y_l \leq n\right). \quad (8.16)$$

Analogously,

$$Z_l = \max_{\substack{l-1 \leq t \leq l \\ 0 \leq s \leq 2}} \nu_{t,s}, l = 1, \ldots, L-1$$

is also a 1-dependent stationary sequence and

$$q_2 = \mathbf{P}\left(\max_{0 \leq l \leq L-1} Z_l \leq n\right). \quad (8.17)$$

Then, Theorem 8.1.2 provides the approximations

$$q_1 \approx (2q_{2,2} - q_{2,3})(1 + q_{2,2} + q_{2,3} + 2(q_{2,2} - q_{2,3})^2)^{-(L-1)}, \quad (8.18)$$

and

$$q_2 \approx (2q_{3,2} - q_{3,3})(1 + q_{3,2} + q_{3,3} + 2(q_{3,2} - q_{3,3})^2)^{-(L-1)}, \quad (8.19)$$

where $q_{2,2} = \mathbf{P}(S_{2,2} \leq n)$, $q_{2,3} = q_{3,2} = \mathbf{P}(S_{2,3} \leq n)$ and $q_{3,3} = \mathbf{P}(S_{3,3} \leq n)$.

Thus, in the two-dimensional case, the final approximation of $\mathbf{P}(S_{L,K} \leq n)$ depends on a prior knowledge of $q_{2,2}$, $q_{2,3}$ and $q_{3,3}$. If $q_{2,2}$, $q_{2,3}$ and $q_{3,3}$ are known and $L \leq K$, it can be shown that the resulting error on the approximation of $\mathbf{P}(S \leq n)$ is bounded by about

$$e = 3.3(L-1)(K-1)\left((1-q_{2,2})^2 + (1-q_{3,2})^2 + (L-1)(q_{2,2} - q_{3,2})^2\right). \quad (8.20)$$

The main difficulty in the two-dimensional case arises from the fact that currently there are no, exact formulas for $q_{2,2}$, $q_{3,2}$ and $q_{3,3}$. This type of approximation for the d.f. of two-dimensional scan statistics generated by a Poisson process was introduced and studied in Haiman and Preda (2002).

As in the one-dimensional case, the method can be adapted to the two-dimensional discrete scan statistics defined as follows.

Let N_1 and N_2 be positive integers and $\{X_{i,j}; 0 \le i \le N_1 - 1, 0 \le j \le N_2 - 1\}$ be a family of i.i.d. nonnegative integer valued r.v.'s from some specified distribution (typically $\mathcal{B}(n,p)$ or Poisson(λ)). For $\le i \le N_1 - 1$ and $0 \le j \le N_2 - 1$, $X_{i,j}$ represents the number of some events observed in the elementary square subregion $[i, i+1] \times [j, j+1]$. Let m_1, m_2 be positive integers, $1 \le m_1 \le N_1$, $1 \le m_2 \le N_2$. For $0 \le t \le N_1 - m_1$, $0 \le s \le N_2 - m_2$, let

$$\nu_{t,s} = \nu_{t,s}(m_1, m_2) = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} X_{i,j}. \tag{8.21}$$

The two-dimensional discrete scan statistic is defined as the largest number of events in any $m_1 \times m_2$ rectangular scanning window within the rectangular region $[0, N_1] \times [0, N_2]$, i.e.,

$$S = S_{N_1,N_2}(m_1, m_2) = \max_{\substack{0 \le t \le N_1 - m_1 \\ 0 \le s \le N_2 - m_2}} \nu_{t,s}. \tag{8.22}$$

Let $N_1 = Lm_1$, $N_2 = Km_2$, with L and K integers, $L > 3$, $K > 3$. In this case, the same arguments and formulas as those leading to the approximation for the continuous scan statistics can be used with the following changes:

- X_k in formula (8.13) is now

$$X_k = \max_{\substack{0 \le t \le (L-1)m_1 \\ (k-1)m_2 \le s \le km_2}} \nu_{t,s}, \quad k = 1, \ldots, K-1,$$

- Y_l in formula (8.15) becomes

$$Y_l = \max_{\substack{(l-1)m_1 \le t \le lm_1 \\ 0 \le s \le m_2}} \nu_{t,s}, \quad l = 1, \ldots, L-1,$$

- and Z_l is

$$Z_l = \max_{\substack{(l-1)m_1 \le t \le lm_1 \\ 0 \le s \le 2m_2}} \nu_{t,s}, \quad l = 1, \ldots, L-1.$$

This type of approximation for the d.f. of discrete two-dimensional scan statistics was studied in Haiman and Preda (2006).

The main aspects related to the application of this method to two-dimensional scan statistics are presented and discussed in Section 8.3.

8.2 Application of the Approximations (8.6) and (8.7) to One-Dimensional Scan Statistics

The approximations (8.6) and (8.7) require a prior knowledge of $q_i = q_i(x) = \mathbf{P}(S_{i+1} \leq x)$, $i = 1, 2, 3, 4$, respectively, $i = 1, 2$. In the one-dimensional case, for both continuous and discrete scan statistics, there are exact formulas for q_n (see the references in Subsections 8.2.1 and 8.2.2). However, these formulas become rapidly intractable as n becomes large. There are also bounds and approximations as those mentioned below. The approximation formulas are based on heuristics, and their accuracy is evaluated using simulation results only in some particular configurations.

Our formulas include error bounds from which one can characterize completely their domain of applicability. A typical application of our method is the following. Suppose we want to establish, for a large $\tau \in \mathbf{N}$, the value of $x_{0.95}$ such that $q_{\tau-1}(x_{0.95}) \approx 0.95$. In the case of the continuous scan statistic, $x_{0.95}$ represents the critical value for testing the intensity λ of the underlying Poisson distribution at the 5% level of significance, i.e., reject the null hypothesis $(\lambda = \lambda_0)$ if $S_{\tau-1} > x_{0.95}(\lambda_0)$.

Under the condition "large τ", $1 - q_1$ is then necessarily small with respect to 0.05 and condition $1 - q_1 \leq 0.025$ is satisfied. Indeed, we then have

$$\begin{aligned} q_{\tau-1}(x_{0.95}) &\approx (2q_1 - q_2)\left(1 + (q_1 - q_2) + 2(q_1 - q_2)^2\right)^{-(\tau-1)} \\ &\approx 1 - (\tau - 1)(q_1 - q_2) \approx 0.95 \text{ as } q_1 \to 1. \end{aligned} \tag{8.23}$$

By Haiman *et al.* (1998), Proposition 2.1, page 490, if $1-q_1$ is sufficiently small, we have $1 - q_1 \leq 2(q_1 - q_2)$.

Thus,

$$1 - q_{\tau-1}(x_{0.95}) \approx 0.05 \approx (\tau - 1)(q_1 - q_2) \geq 2(\tau - 1)(1 - q_1), \tag{8.24}$$

whereas the error bound is about $3.3\tau(1 - q_1)^2$, thus very small with respect to the approximated value of 0.05. When q_3 and q_4 are available, the approximation (8.6) is more accurate (error of order $(1-q_1)^3$ instead of $(1-q_1)^2$), but generally, the approximation (8.7) appears to be sufficiently precise.

We now examine separately the application of the method to continuous and discrete scan statistics.

8.2.1 Application to one-dimensional continuous scan statistics

Let $S = S(u, \lambda, T)$ be the scan statistic generated by a Poisson process as defined in (8.1). Huntington and Naus (1975) give an exact formula for $\mathbf{P}(S \leq n)$ for $n \geq 0$ and $T \geq u$ that sums many products of determinants and for large

T requires excessive computation time. This formula is used in Neff and Nauss (1980) to establish tables for the d.f. of $S(1, \lambda, \tau)$ (notice that $S(u, \lambda, T) = S(1, \lambda u, \frac{T}{u})$) for several discrete values of λ and $\tau \leq 100$.

In Haiman (2000) we have applied the approximation (8.7) with q_1 and q_2 from Neff and Naus tables and $\tau = 1000$. Notice that when we mention a numerical application of the approximations (8.6) or (8.7), it means that we also provide the corresponding error bound.

Naus (1982), making a reasoning based on the hypothesis of a Markov-like behavior of the sequence $\{X_n\}$ defined in (8.2), proposes the approximation

$$q_\tau = q_{\tau(x)} = \mathbf{P}(S_\tau \leq x) \approx q_1 \left(\frac{q_2}{q_1}\right)^{\tau-2}, \tau > 2. \tag{8.25}$$

He shows, using the exact formula, that for λ and τ ranging in a certain domain, and also compared to other existing approximations, approximation (8.25) is remarkably accurate. This fact is not surprising: if we denote, respectively, by q_τ^H and q_τ^N the approximations in (8.7) and (8.25), it can be shown that for τ sufficiently large we have $|q_\tau^H - q_\tau^N| \leq 5(1 - q_1)^2$. Table 8.1 presents some numerical examples of these approximations and illustrates this fact.

Another scan statistic of interest generated by a Poisson process is defined as

$$S^* = S^*(u, \lambda, T) = \min_{0 \leq t \leq T-u} \nu_t, \tag{8.26}$$

where, as in (8.1), $\nu_t = N(t + u) - N_t$.

Let $T = \tau u$, $\tau \in \mathbf{N}$, $\tau > 0$ and let

$$X_k^* = - \min_{(k-1)u \leq t \leq ku} \nu_t. \tag{8.27}$$

Then $\{X_k^*\}$ forms a 1-dependent stationary sequence and

$$\overline{q}_\tau^*(n) = \mathbf{P}(S_\tau^* > n) = \mathbf{P}\left(\max_{1 \leq k \leq \tau-1} X_k^* < -n\right), \ n \geq 0. \tag{8.28}$$

Theorems 8.1.1 and 8.1.2 can also be applied here and corresponding versions of the approximations (8.6) and (8.7) can be used to estimate $\overline{q}_\tau^*(n)$. The values

Table 8.1. Approximations for $\mathbf{P}(S \leq x)$ by approximations (8.25) and (8.7). $T = 1001$.

x	λ	Naus (1982)	Haiman (2000)	Error
4	0.1	0.985399334	0.9854	2×10^{-6}
6	0.5	0.930142831	0.9302	2.5×10^{-5}
9	1.3	0.940503808	0.9405	1.7×10^{-5}

of $\overline{q}_i^*(n)$, $i = 1, 2, 3, 4$, or $i = 1, 2$ used in these approximations can be obtained from the exact formulas established in Huntington (1978) (these exact formulas also become intractable as τ becomes large).

Janson (1984) gives upper and lower bounds for q_τ and $\overline{q}_\tau^*$. In Haiman (2000) we have shown that the approximation (8.7) and Janson's bounds have similar precision.

The waiting time until the first occurrence of n points within an interval of length u, W_n, is an r.v. whose distribution is important in several applications (see Naus (1982)). For $n \geq 1$ and $t \geq 2$ we have

$$\mathbf{P}(W_n > t) = \mathbf{P}\left(\max_{0 \leq s \leq t-u} \nu_s < n\right). \tag{8.29}$$

Let W_n^* be the corresponding discretized waiting time defined as

$$W_n^* = \left[\frac{\min\{s \geq 0 : \nu_s = n\}}{u}\right] u, \tag{8.30}$$

where $[\,\cdot\,]$ stands for integer part.

For $n \geq 1$ and $\tau = 2, 3, \ldots$, we have

$$\mathbf{P}(W_n^* > \tau) = \mathbf{P}(S_\tau \leq n-1) = q_{\tau-1}(n-1). \tag{8.31}$$

We then can apply the approximations (8.6) or (8.7) to estimate the expected waiting time, $\mathbf{E}(W_n^*)$. Details about this application and a numerical example are given in Haiman (2000).

Let $M(T)$ be the number of subintervals, each of length u, dropped so that their midpoints are the occurrence points of a homogenous Poisson process N in the interval $[0, T]$. We say that a point x is covered by a subinterval with midpoint y if $y - \frac{u}{2} \leq x \leq y + \frac{u}{2}$. The calculation of the probability of the event $E_n =$ *"all points of the interval [0,T] are covered by at least n subintervals"* is of interest in several applications [see Glaz and Naus (1978)]. Let $T = \tau u$, $\tau = 2, 3, \ldots$. In Haiman (2000) we use the fact that the calculation of $\mathbf{P}(E_n)$ is related to the calculation of $\overline{q}_\tau^*$. Thus, via the approximation (8.7) we obtain an approximation formula for $\mathbf{P}(E_n)$.

Let F_n be the event *"there does not exist a subarc of length $u = 1$ of a circle with circumference τ, $\tau = 2, 3, \ldots$, that contains n points."* Using similar arguments, in Haiman (2000) we obtain an approximation formula for $\mathbf{P}(F_n)$.

8.2.2 Application to one-dimensional discrete scan statistics

Let $Z_1, \ldots, Z_N$ be a sequence of integer-valued r.v.'s that are i.i.d. and consider the discrete scan statistic S defined in (8.9). Exact formulas for $\mathbf{P}(S \geq k)$ exist, and some of them are tractable only in a limited number of situations. The Bernoulli case ($Z_i \sim \mathcal{B}(1, p)$) plays an important role in the applications. In

this case, exact formulas have been obtained by Naus (1982) for $N = 2m$ and $N = 3m$, i.e., for q_1 and q_2. As for continuous scan statistics, Naus uses q_1 and q_2 to estimate $q_\tau = q_\tau(k) = \mathbf{P}(S_\tau \leq k)$ ($N = \tau m$, $\tau \geq 3$) by formula (8.25).

Fu (2001) employed a finite Markov chain embedding method to derive exact formulas for $\mathbf{P}(S_\tau \leq k)$. However, this method involves quite complicated computations, and it may become difficult to use for large or very large values of m and $\tau = \frac{N}{m}$.

Thereby, various approximation methods and bounds for $\mathbf{P}(S \leq k)$ have been proposed by several authors. However, the quality of these approximations and bounds can be evaluated for a limited number of particular configurations. An overview of these results as well as a complete bibliography on the subject are given in Glaz *et al.* (2001). In Haiman (2007) we have illustrated by several numerical examples the application of our approximation (8.7) in parallel with formula (8.25) of Naus. In these examples, $m = 30$, $p = 0.1$ and N ranges from 256×30 to 1024×30. As for continuous scan statistics and for a similar reason (see Section 8.2.1) the approximations (8.7) and (8.25) give very close results.

Let $V(N)$ denote the *length of the longest success run* in N Bernoulli $\mathcal{B}(1, p)$ trials (1 = success, 0 = failure). We then have

$$\mathbf{P}(V(N) \geq m) = \mathbf{P}(S \geq m) = \mathbf{P}(S = m). \tag{8.32}$$

Thus, if $N = \tau m$, $\tau \geq 2$,

$$\mathbf{P}(V(N) \geq m) = \mathbf{P}(S_\tau \geq m) = 1 - q_{\tau-1}(m). \tag{8.33}$$

An exact formula for $\mathbf{P}(V(N) \geq m)$ of Bateman (1948) allows in this case an easy calculation of $q_i(m)$, $i = 1, 2, 3, 4$, from which $\mathbf{P}(V(N) \geq m)$ can be approximated by either one of the approximations (8.6) or (8.7). In Haiman (2007) we have used numerical examples to illustrate and compare these two approximations. It appears that the approximations (8.6) and (8.7) provide very close results. Table 8.2 presents some numerical examples of these approximation and illustrates this fact.

Fu *et al.* (2003) have used the finite Markov chain embedding to obtain the exact distribution of $V(N)$. They also obtained a large deviation approximation of the above distribution [in relationship to this problem, see also Lou (1996), Vaggelatou (2003) and the references quoted in these papers].

In Haiman (2007), we also compare the approximation (8.6) and exact values of $V(N)$ calculated in Fu *et al.* (2003).

Let k and m, $1 \leq k \leq m$, be positive integers and define the waiting time, until "$k - in - m$ quota" by

$$T = T_{k,m} = \inf\{t \geq 1 : \mu_t \geq k\}, \tag{8.34}$$

where μ_t is defined in (8.8).

Table 8.2. Approximations for $\mathbf{P}(S \leq x)$ by Haiman (2007) and Naus (1982), $X_i \sim \mathcal{B}(1,p)$, $p = 0.1$, $m = 30$.

x	9	10	11
$\mathbf{P}(S(30, 256 \times 30) \leq x)$:			
App. (8.7)	0.5161	0.85979	0.970613
Error	0.008	0.0023	10^{-6}
App. (8.25)	0.5172	0.86028	0.970726
$\mathbf{P}(S(30, 512 \times 30) \leq x)$:			
App. (8.7)	0.2658	0.73888	0.941997
Error	0.017	0.00046	0.000017
App. (8.25)	0.2663	0.739295	0.9421067

Huntington (1974) derives an exact and quite complicated formula for $\mathbf{E}(T)$, in terms of ratios of determinants of some matrices. Naus (1982), using the fact that

$$\mathbf{E}(T_{k,m}) = \sum_{N=0}^{\infty} (1 - \mathbf{P}(S_N < k)),$$

uses the approximation (8.25) to obtain the approximation

$$\mathbf{E}(T_{k,m}) \approx 2m + \frac{q_2}{(1 - \frac{q_2}{q_1})^{\frac{1}{m}}}. \tag{8.35}$$

In Haiman (2007), we similarly use the approximation (8.7) to establish upper and lower bounds for $\mathbf{E}(T_{k,m})$ and give some numerical examples.

Let now r.v. Z_i, $i = 1, \ldots, N$ take values -1, 0 and 1. The corresponding discrete scan statistic S is associated to the "charge problem." Exact results for $\mathbf{P}(S \leq k)$ have been obtained in this case by Saperstein (1976) for $N \leq 2m$ and by Karwe (1993) for $N \in \{2m-1, 2m \text{ (thus } q_1), 3m-1 \text{ and } 3m \text{ (thus } q_2)\}$. In Haiman (2007) we give numerical examples and compare the approximations (8.7) and (8.25) using values of q_1 and q_2 provided in Karwe (1993).

8.3 Application of the Method to Two-Dimensional Scan Statistics

As mentioned in Section 8.1, the main difficulty in applying the method to both, continuous and discrete two-dimensional scan statistics arises from the fact that at present there are no exact formulas allowing us to calculate $q_{i,j}$, $i, j = 2, 3$.

There are some approximation formulas (see references below) based on heuristics; their accuracy is evaluated using simulation results only in some particular configurations. As in the one-dimensional case, the characteristic of our approximation formulas is that they include error bounds.

8.3.1 Application to continuous scan statistics

Let S be defined in (8.12) and for $u = v = 1$ and K, L integers, $K, L > 3$, put

$$S_{L,K} = S = S((1,1), \lambda, L, K). \tag{8.36}$$

Previously, Aldous (1989) and Alm (1997) have established approximation formulas for the d.f. of $S_{L,K}$.

Let

$$q^n_{L,K}(k) = \mathbf{P}\left(S_{L,K} \le k \middle| N([0,L] \times [0,K]) = n\right), 1 \le k \le n \tag{8.37}$$

denote the d.f. of the conditional scan statistic, i.e., the scan statistic given that a fixed number n of points fall in $[0,L] \times [0,K]$. Notice that $q^n_{L,K}$ is the d.f. of the r.v. $S^n_{L,K}$ = *maximum number of points obtained by scanning with the* $[0,1] \times [0,1]$ *window a rectangle* $[0,L] \times [0,K]$ *in which n independent points are drawn uniformly.*

We then have

$$\begin{aligned} q_{L,K}(k) &= \mathbf{P}\left(S_{L,K} \le k\right) \\ &= e^{-\lambda LK}\left(\sum_{j=0}^{k} \frac{(\lambda LK)^j}{j!} + \sum_{j=k+1}^{kLK} q^j_{L,K}(k)\frac{(\lambda LK)^j}{j!}\right). \end{aligned} \tag{8.38}$$

In Haiman and Preda (2002) we have developed a method of "perfect" simulation of independent replications of r.v.'s $S^n_{i,j}$, $i, j = 2, 3$. We construct (Theorem 2) a stopping time T with respect to the filtration generated by a sequence $\{Z_n\}_{n\ge1}$ of Bernoulli $\mathcal{B}(1, \frac{1}{2})$ i.i.d. r.v's together with functions $f_t(z_1, \ldots, z_t)$ such that the r.v. $\mathcal{S}^n_{i,j} = f_T(Z_1, \ldots, Z_T)$ has the same distribution as $S^n_{i,j}$. We use this method to obtain via formula (8.38) empirical estimations of $q^n_{i,j}(k)$, $i, j = 2, 3$ and then we calculate (see Section 8.1) the final approximation of $q_{L,K}(k)$.

The empirical estimation of $q^n_{i,j}$ generates additional errors. These errors are bounded at the 95% confidence level by $\varepsilon_{i,j}$, where $\varepsilon_{i,j} \approx 1.96\sqrt{\frac{q^n_{i,j}(1-q^n_{i,j})}{M}}$. M is the number of replications of r.v.'s $S^n_{i,j}$, $i, j = 2, 3$. The total error on $\mathbf{P}(S_{L,K} \le k)$ is then bounded by about

$$E = e + LK(\varepsilon_{2,2} + \varepsilon_{2,3} + \varepsilon_{3,3}), \tag{8.39}$$

with e, the error bound when $q_{i,j}$ are known, given in (8.20). Naus (1965) and Neff (1978) give exact formulas for $q^m_{L,K}(m-1)$ and $q^m_{L,K}(m-2)$. In Haiman

Table 8.3. Approximation for $\mathbf{P}(S \leq n)$. $L = 500$, $K = 500$, $\lambda = 0.01$.

n	App. (8.14)	Error	Alm (1997)	Aldous (1989)
2	0.69318103	0.008570775	0.7839302629	0.8484459199
3	0.998401542	6.37679E-05	0.9987785770	0.9990759644

and Preda (2002) we use these formulas for $L, K = 2, 3$ to evaluate our simulation results. We then give numerical examples for several values of L, K and λ ($L, K = 10, 50, 100, 1000$, $\lambda = 0.01, 0.05, 0.1, 1$) and compare our results with corresponding results obtained by other approximation formulas in Aldous (1989) and Alm (1997).

In order to obtain error bounds $\varepsilon_{i,j}$ such that their contribution to the total error E has the same order of magnitude as e, we use in our examples up to 10^7 replications of r.v.'s $S^n_{i,j}$, $i, j = 2, 3$.

Table 8.3 presents some numerical examples of application of our method and the corresponding results obtained using the methods of Aldous and Alm.

8.3.2 Application to discrete scan statistics

Let $S = S_{N_1,N_2}$ be defined in (8.22) where the underlying $X_{i,j}$ are binomial $\mathcal{B}(n, p)$ or Poisson $\mathcal{P}(\lambda)$. Since there are no exact formulas for $\mathbf{P}(S \leq k)$, various methods of approximation and bounds have been proposed by several authors. An overview of these methods as well as a complete bibliography on the subject are given in Glaz *et al.* (2001). In particular, the case where $X_{i,j}$ are binary variables, with application to reliability (two-dimensional $r - within\ m_1 \times m_2 - out - of\ N_1 \times N_2$) has received considerable research interest during the last years. In this framework, several approximations and bounds have been proposed and studied in the literature [see, e.g., Chen and Glaz (1996), Boutsikas and Koutras (2003) and references therein].

Let $N_1 = Lm_1$ and $N_2 = Km_2$ with L and K integers, $L, K > 3$. In Haiman and Preda (2006) we have applied our approximation method of $\mathbf{P}(S_{N_1,N_2} \leq k)$ using, similarly to the previous continuous case, empirical estimations of

$$q_{i,j}(k) = \mathbf{P}(S_{im_1,jm_2} \leq k)$$

obtained by simulating i.i.d. replications of r.v.'s S_{im_1,jm_2}, $i, j = 2, 3$. The error bound due to simulation, e_{sim}, is then also proportional to $\frac{LK}{\sqrt{M}}$, where M is the number of replications and the total error bound, as in (8.39), is

$$E = e + e_{sim}.$$

For $X_{i,j}$ binomial and Poisson we give numerical examples and compare our results with those obtained using the product-type approximation, the Poisson approximation and Bonferroni inequality techniques, as presented in Glaz *et al.* (2001).

Table 8.4. Approximation for $\mathbf{P}(S \leq x) : X_{i,j} \sim \text{Poisson}(0.25)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $M = 10^9$.

x	$\hat{\mathbf{P}}(S \leq x)$	P-T	Bonferroni	Poisson	H-P	Error
15	0.8596	0.8374	0.7700	0.8292	0.860427482	0.067409646
16	0.9402	0.9351	0.9130	0.9314	0.940749305	0.010867255
17	0.9783	0.9764	0.9691	0.9750	0.977260378	0.001546897
18	0.9930	0.9920	0.9896	0.9916	0.991966851	0.000217233

Table 8.5. Approximation for $\mathbf{P}(S \leq x) : X_{i,j} \sim \mathcal{B}(5, 0.05)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $M = 10^9$.

x	$\hat{\mathbf{P}}(S \leq x)$	P-T	Bonferroni	Poisson	H-P	Error
15	0.8932	0.8830	0.8387	0.8768	0.896135764	0.035108915
16	0.9617	0.9577	0.9441	0.9554	0.960112719	0.004770939
17	0.9868	0.9862	0.9819	0.9854	0.986256278	0.000584065
18	0.9948	0.9958	0.9946	0.9956	0.995633424	8.08015E-05

For binary $X_{i,j}$ we compare our approximations with bounds obtained in Boutsikas and Koutras (2003).

In all these examples we use up to $M = 10^9$ replications of r.v. S_{im_1,jm_2}, $i, j = 2, 3$. Tables 8.4 and 8.5 present some numerical examples of the application of our method and the corresponding results obtained using the product-type (P-T), the Poisson and the Bonferroni approximation methods. $\hat{\mathbf{P}}(S \leq x)$ denotes the empirical estimation of $\mathbf{P}(S \leq x)$ using 10,000 trials [see Glaz *et al.* (2001)].

For binomial $X_{i,j}$, and in particular Bernoulli, the current work of the authors consists in constructing computer algorithms allowing one to obtain, without using simulations, exact values or sufficiently accurate (with respect to the method) approximations of $q_{i,j}(m)$, $i, j = 2, 3$.

References

1. Aldous, D. (1989). *Probability Approximation via the Poisson Clumping Heuristic*, Springer-Verlag, New York.

2. Alm, S.E. (1997). On the distribution of scan statistics of two-dimensional Poisson processes, *Advances Applied Probability*, **29**, 1–18.

3. Bateman, G.I. (1948). On the power function of the longest run as a test for randomness in a sequence of alternatives, *Biometrika*, **35**, 97–112.

4. Boutsikas, M. and Koutras, M. (2003). Bounds for the distribution of two dimensional binary scan statistics, *Probability in the Engineering and Information Sciences*, **17**, 509–525.

5. Chen, J. and Glaz, J. (1996). Two-dimensional discrete scan statistics, *Statistics and Probability letters*, **31**, 59–68.

6. Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bistate trials, *Journal of Applied Probability*, **38**, 4, 908–916.

7. Fu, J.C., Wang, L. and Lou, W. (2003). On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials, *Journal of Applied Probability*, **40**, 2, 346–360.

8. Glaz, J. and Naus, J.I. (1978). Multiple coverage on the line, *Annals of Probability* **7**, 900–906.

9. Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer Series in Statistics, Springer-Verlag, New York.

10. Haiman, G. (1999). First passage time for some stationary processes, *Stochastic Processes and Their Applications*, **80**, 231–248.

11. Haiman, G. (2000). Estimating the distribution of scan statistics with high precision, *Extremes*, **3:4**, 349–361.

12. Haiman, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences, *Journal of Statistical Planning and Inference*, **137:3**, 821–828.

13. Haiman, G., Mayeur, N., Nevzorov, V. and Puri, M.L. (1998) Records and 2-block records of 1-dependent stationary sequences under local dependence, *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **34:4**, 481–503.

14. Haiman, G. and Preda, C. (2002). A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process, *Methodology and Computing in Applied Probability*, **4**, 393–407.

15. Haiman, G. and Preda, C. (2006). Estimation for the distribution of two-dimensional discrete scan statistics, *Methodology and Computing in Applied Probability*, **8**, 373–382.

16. Huntington, R.J. (1974). Distributions and expectations for clusters in continuous and discrete cases, with applications, *Ph.D. Thesis*, Rutgers University.

17. Huntington, R.J. (1978). Distribution of the minimum number of points in a scanning interval on the line. *Stochastic Processes and Their Applications*, **7**, 73–78.

18. Huntington, R.J. and Naus, J.I. (1975). A simpler expression for kth nearest-neighbor coincidence probabilities, *Annals of Probability*, **3**, 894–896.

19. Janson, S. (1984). Bounds on the distribution of extremal values of a scanning process, *Stochastic Processes and Their Applications*, **18**, 313–328.

20. Karwe, V.V. (1993). The distribution of the supremum of integer moving average processes with applications to the maximum net charge in DNA sequences, *Ph.D. Thesis*, Rutgers University.

21. Lou, W. (1996). On runs and longest run tests: a method of finite Markov chain imbedding, *J. Amer. Statist. Assoc.* **91**, 1595–1601.

22. Naus, J.I. (1965). A power comparison of two tests of non-random clustering, *Technometrics*, **8**, 493–517.

23. Naus, J.I. (1982). Approximations for distributions of scan statistics, *Journal of the American Statistical Association*, **77**, 177–183.

24. Neff, N. (1978) Piecewise polynomials for the probability of clustering on the unit interval, *Unpublished PhD. dissertation*, Rutgers University.

25. Neff N.D. and Naus, J.I. (1980). The distribution of the size of the maximum cluster points on a line, In *IMS Series Selected Tables in Mathematical Statistics*, Vol. **IV**, American Mathematical Society, Providence, RI.

26. Saperstein, B. (1976). The analysis of attribute moving averages: MIL-STD-105D reduced inspection plan, *Sixth Conference Stochastic Processes and Applications*, Tel Aviv.

27. Vaggelatou, E. (2003). On the length of the longest run in a multi-state Markov chain, *Statistics & Probability Letters*, **62:3**, 211–221.

9

Scan Statistics in Genome-Wide Scan for Complex Trait Loci

Josephine Hoh[1] **and Jurg Ott**[2]

[1] *Department of Epidemiology and Public Health, Yale University, New Haven, CT, USA*
[2] *Beijing Institute of Genomics, Beijing, China*

Abstract: In genome-wide genetic scans for disease susceptibility loci, true peaks have been shown to be wider than false peaks. We describe scan statistics to make use of this extra information, which is not generally taken into account otherwise. Our methods are applied to four disease datasets.

Keywords and phrases: Genome-wide scans, linkage analysis, genetic association analysis, autism, schizophrenia, Parkinson's disease, age-related macular degeneration

9.1 Introduction

We have used the scan-statistics approach on linkage mapping by jointly analyzing information at a number of microsatellite marker loci covering a contiguous region of the genome [Hoh and Ott (2000), Hoh and Ott (2003)]. This paper is based on our previous work, which will be extended to mapping with single nucleotide polymorphisms (SNPs) in genome-wide association studies.

In linkage analysis, logarithms of likelihood ratios (lod scores) are computed for many DNA markers on the genome. The likelihood in the numerator refers to the presence of a susceptibility locus at a given position, and the likelihood in the denominator assumes absence of that locus. Methods implemented in programs such as Aspex [Schwab *et al.* (1995)], Genehunter [Kruglyak *et al.* (1996), Friddle *et al.* (2000)], and Allegro [Gudbjartsson *et al.* (2000)] can make use of microsatellite marker loci on a chromosome and render any point along the chromosome as informative as possible. True peaks of such lod score curves are known to be wider than false peaks [Terwilliger *et al.* (1997)]. Consequently, higher positive lod scores and a larger number of them are expected around true rather than around false peaks. This property of lod scores is not generally

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_9,

taken into account in the search for susceptibility loci, but ad hoc approaches have suggested increased power when information from a small number of neighboring markers is combined [Goldin and Chase (1997), Goldin *et al.* (1999)]. Theoretical investigations show that the power gain from this information for a single susceptibility locus might be modest [Siegmund (2001)] but can be more substantial for two loci in close proximity [Hernandez *et al.* (2005)].

Further extension of the traditional linkage methods using scan statistics is a novel way of testing for disease association/linkage. This new method combines information from marker loci clustering around a local peak and assesses its genome-wide significance by permutation tests. The information to be combined is based on single-marker statistics, which might be lod scores in general small families, allele-sharing proportions for sib pairs, or Pearson χ^2 association statistics in case-control association studies. For a given "length" (the number of combined single-marker statistics) of the scan statistic, we assess its significance by permutation tests. Comparing p-values for varying lengths of scan statistics, we treat the smallest observed p-value as our statistic of interest and determine its overall significance level. This method has been illustrated in an autism pedigree dataset. A susceptibility region was found (genome-wide significance level = 0.038), which was missed in conventional analyses (see below).

9.2 Methods

Consider a sequence of random variables, $X_1, ..., X_N$. For $1 \leq L \leq N$, let $Y_L(t) = \sum_{i+t}^{t+L-1} X_i$ be a moving sum of L consecutive observations. The linear (unconditional) scan statistic is then defined as

$$S_L = \max[Y_L(1), Y_L(2), ..., Y_L(N - L + 1)], \tag{9.1}$$

that is, as the largest moving sum of length L [Glaz and Balakrishnan (1999)]. Scan statistics have been used in epidemiology, molecular biology, and many other areas of science and engineering to detect clustering, for example, in DNA sequence analysis [Karlin and Brendel (1992)].

Here, X_i is an observation or a statistic based on the genotypes at the i-th marker, and the sum $Y_L(t)$ refers to the combined information of ordered markers, moving along the chromosomes. Alternatively, there is precedent for using log likelihood ratios as "observations"—correlations between lod scores [MacLean *et al.* (1993)] or allele sharing proportions [Cox *et al.* (1999)] at specific loci have been interpreted as evidence for genetic interaction between these loci. It is clear that a scan statistic based on lod scores captures the particular feature of true peaks being wider than false peaks (see Introduction).

Therefore, scan statistics are expected to be more powerful for detection of susceptibility loci than a statistic focused only on a single marker locus.

The corresponding statistical test is mathematically intractable but can be achieved by computer-based methods, bootstrap or permutation, which have been proven effective [Efron and Tibshirani (1991)]. Below, we employ permutation tests to search for clusters of consecutive markers that point to a gene underlying the trait studied.

Under the null hypothesis of no disease association or linkage, any set of marker genotypes in an individual is equally likely to occur with a binary outcome. This implies that data matrices with any permutation of the n binary outcomes have equal probabilities of occurrence. For each permutation sample, the proportion, p_L, of permutation samples with a scan statistic at least as large as an observed scan statistic, S_L, represents the significance level associated with S_L.

The p_L-value so computed represents the global significance level, as opposed to a locus-specific significance level [Lander and Kruglyak (1995)], for a given value of L. However, there may be no *a priori* reason for choosing any particular value for L. Rather, one would like to try any one of the values from 1 through, say, $L_{\max} = 10$ and focus on the smallest p_L-value obtained. This minimum p_L-value, $p_{\min}$, then represents the statistic whose significance level is to be determined. It is obtained from the permutation samples as follows. We view the statistics, $S_1, S_2, ..., S_{L_{\max}}$, as multiple (correlated) measurements. In each permutation sample, a minimum significance level, $p^*_{\min}$, is obtained in analogy to the one observed in the real data. Then, the overall significance level, p_{global}, associated with $p_{\min}$ is given by the proportion of permutation samples with $p^*_{\min} \leq p_{\min}$ [Manly (2006)].

We recommend the following strategy to make this approach as efficient as possible: testing by adding lengths sequentially until the scan statistic starts decreasing, which is equivalent to the sequential test until the p-value starts increasing.

9.3 Applications

9.3.1 Autism data

In a genome screen for autism, independent sib pairs were genotyped for a total of 324 microsatellites [Liu *et al.* (2001)]. With a broad disease definition, 86 affected-affected (AA) and 91 affected-unaffected (AU) sib pairs were available. At each marker locus, the Allegro program [Gudbjartsson *et al.* (2000)] determined the lod score associated with the allele-sharing proportion in each sib pair. The statistic used for the i-th marker was the difference, $X_i = u_{AA} - u_{AU}$,

where u_{AA} and u_{AU} are total lod scores in AA and AU sib pairs, respectively. Scan statistics of lengths 1 through 10 were tried.

For marker number 159, the total lod score observed in AA pairs was 1.21 and that for AU pairs was −2.29. Neither lod score is remarkable. The difference in lod scores, 3.50, is the largest such difference observed in the data and, in a permutation test, is associated with a genome-wide significance level of 0.131. With an associated significance level of $p_{\min} = 0.015$, the most significant scan statistic is that of length 6.

The genome-wide significance level associated with $p_{\min}, p_{\text{global}} = 0.038$, was estimated with 100,000 permutations. For a map of 400 microsatellite markers, a genome-wide significance level of $p_{\text{global}} = 0.05$ corresponds to a locus-specific significance level of $p = 0.000022$ or a lod score of 3.6 [Lander and Kruglyak (1995)]. Similarly, $p_{\text{global}} = 0.038$ translates into $p = 0.0000163$ or a lod score of 3.8.

9.3.2 Schizophrenia data

Over the years, various genome screens for schizophrenia susceptibility loci had generally furnished rather modest results. On chromosome 10, two peaks relatively close to each other had been observed in European-American families but none of them was significant [Faraone *et al.* (1998)]. The experiment-wise significance level for the peak seen at marker D10S1423 was only $p \approx 0.20$. In our analysis with permutation testing in multipoint analysis, the peak nonparametric linkage score at D10S1423 in those data reached an experiment-wise significance level of $p = 0.016$ while application of scan statistics resulted in $p = 0.008$ [Dewan and Ott (2004)].

9.3.3 Parkinson's disease data

We carried out association tests for Parkinson's disease (PD) data [Fung *et al.* (2006), Simon-Sanchez *et al.* (2007)]. To reduce the computational effort, we focused on chromosome 11 because some SNPs on that chromosome were previously reported as being associated with PD [Fung *et al.* (2006)]. A total of 19,494 SNPs had been genotyped on 270 case and 271 control individuals and passed our quality control measures. Genotypes AA, AB, and BB were coded 1, 2, and 3, respectively. As a test statistic, we applied a t-test to each SNP, that is, we tested for a difference in the mean number of A alleles between cases and controls. All p-values were estimated on the basis of 20,000 randomization samples. The single best SNP (scan statistic of length 1) had an associated p-value of 0.2404, so was far from significant. The best scan statistic had length 13 and comprised SNPs between $rs7951781$ and $rs647248$ that are 61,810 bp apart. Its p-value was $p_{\min} = 0.0142$ with a corresponding significance level

of $p_{\text{global}} = 0.0620$. Thus, these results provide a hint of a region of disease association, but genome-wide statistical significance is lacking.

As is often the case, different test statistics provide different answers. For example, when we apply a chi-square test to 2×3 tables of genotypes (case/control versus 3 genotypes), the single best SNP has p-value 0.0010 while the best scan statistic with length > 1 has p-value 0.0486. Evidently, this statistic gives no evidence for a cumulative effect of neighboring SNPs over that of single SNPs.

9.3.4 Age-related macular degeneration (AMD) data

We applied the scan statistics method to our two AMD genome-wide association (GWA) datasets [Klein *et al.* (2005), Dewan *et al.* (2006)]. Single SNPs were found highly significant in these data. Neither scan statistics nor haplotypes gained further information surrounding these SNPs.

9.4 Discussion

In the autism data of the Applications section, the significance level is improved from 0.131 by traditional analysis to 0.038 by our scan statistics method. It demonstrates the usefulness of the scan statistic approach. This method can also be useful in association studies when thousands of dense SNP markers are tested.

Different single-marker statistics may have unequal properties. For example, if markers have different numbers of alleles, and allele frequencies are compared between cases and controls, a suitable statistic is chi-square for a $2 \times n$ table, with n being the number of alleles. Markers with different numbers of alleles will yield statistics with different numbers of degrees of freedom. One may convert these statistics to empirical significance levels, p, and use $\log[\log(p)]$ as the statistics of interest, which now are all on an equal scale.

The increased power provided by scan statistics may yield a stronger result than conventional statistics. That is, with a given number of observations, conventional methods will detect susceptibility loci of some minimum effect, but scan statistics will detect loci of smaller effects. However, the proposed scan statistics method does not aim at narrowing a candidate region to a point mutation, which is always the ultimate goal in gene mapping.

In the GWA studies, using dense SNP information from scan statistics may be comparable with using that obtained from the haplotype effects.

Acknowledgments

Support of this work for JO includes China NSFC grant, project number 30730057 and MH44292 from the U.S. National Institute of Mental Health (NIH); and for JH includes NIH R01 EY 015771, R21 EY018127, K25 HG000060; the Macular Vision Research Foundation; the Ellison Foundation for Medical Research; and the American Health Assistance Foundation. This study used data from the SNP Database at the NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds), as well as clinical data. The original genotyping was performed in the laboratories of Drs. Singleton and Hardy, (NIA, LNG), Bethesda, MD, USA.

References

1. Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I., and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nature Genetics*, **21**, 213–215.

2. Dewan, A., Liu, M., Hartman, S., Zhang, S. S., Liu, D. T., Zhao, C., Tam, P. O., Chan, W. M., Lam, D. S., Snyder, M., Barnstable, C., Pang, C. P., and Hoh, J. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration, *Science*, **314**, 989–992.

3. Dewan, A., and Ott, J. (2004). Reanalysis of a genome scan for schizophrenia loci using multigenic methods, *Human Heredity*, **57**, 191–194.

4. Efron, B., and Tibshirani, R. (1991). Statistical data analysis in the computer age, *Science* **253**, 390–395.

5. Faraone, S. V., Matise, T., Svrakic, D., Pepple, J., Malaspina, D., Suarez, B., Hampe, C., Zambuto, C. T., Schmitt, K., Meyer, J., Markel, P., Lee, H., Harkavy Friedman, J., Kaufmann, C., Cloninger, C. R., and Tsuang, M. T. (1998). Genome scan of European-American schizophrenia pedigrees: results of the NIMH Genetics Initiative and Millennium Consortium, *American Journal of Medical Genetics*, **81**, 290–295.

6. Friddle, C., Koskela, R., Ranade, K., Hebert, J., Cargill, M., Clark, C. D., McInnis, M., Simpson, S., McMahon, F., Stine, O. C., Meyers, D., Xu, J., MacKinnon, D., Swift-Scanlan, T., Jamison, K., Folstein, S., Daly, M., Kruglyak, L., Marr, T., DePaulo, J. R., and Botstein, D. (2000). Full-genome scan for linkage in 50 families segregating the bipolar affective disease phenotype, *American Journal of Human Genetics*, **66**, 205–215.

7. Fung, H. C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J. R., Langefeld, C., Stiegert, M. L., Schymick, J., Okun, M. S., Mandel, R. J., Fernandez, H. H., Foote, K. D., Rodriguez, R. L., Peckham, E., De Vrieze, F. W., Gwinn-Hardy, K., Hardy, J. A., and Singleton, A. (2006). Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data, *Lancet Neurology*, **5**, 911–916.

8. Glaz, J., and Balakrishnan, N. (1999). Introduction to scan statistics, in *Scan Statistics and Applications*, Editors: Glaz, J., Balakrishnan, N., 3–26, Birkhäuser, Boston, MA.

9. Goldin, L. R., and Chase, G. A. (1997). Improvement of the power to detect complex disease genes by regional inference procedures, *Genetic Epidemiology*, **14**, 785–789.

10. Goldin, L. R., Chase, G. A., and Wilson, A. F. (1999). Regional inference with averaged P values increases the power to detect linkage, *Genetic Epidemiology*, **17**, 157–164.

11. Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis, *Nature Genetics*, **25**, 12–13.

12. Hernandez, S., Siegmund, D. O., and de Gunst, M. (2005) On the power for linkage detection using a test based on scan statistics, *Biostatistics*, **6**, 259–269.

13. Hoh, J., and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes, *Proceedings of the National Academy of Sciences USA*, **97**, 9615–9617.

14. Hoh, J., and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes, *Nature Reviews Genetics* **4**, 701–709.

15. Karlin, S., and Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis, *Science*, **257**, 39–49.

16. Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration, *Science*, **308**, 385–389.

17. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics*, **58**, 1347–1363.

18. Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, *Nature Genetics*, **11**, 241–247.

19. Liu, J., Nyholt, D. R., Magnussen, P., Parano, E., Pavone, P., Geschwind, D., Lord, C., Iversen, P., Hoh, J., Ott, J., and Gilliam, T. C. (2001). A genome-wide screen for autism susceptibility loci, *American Journal of Human Genetics*, **69**, 327–340.

20. MacLean, C. J., Sham, P. C., and Kendler, K. S. (1993). Joint linkage of multiple loci for a complex disorder, *American Journal of Human Genetics* **53**, 353–366.

21. Manly, Bryan F. J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall/CRC, New York.

22. Schwab, S. G., Albus, M., Hallmayer, J., Honig, S., Borrmann, M., Lichtermann, D., Ebstein, R. P., Ackenheil, M., Lerer, B., Risch, N., Maier, W., and Wildenauer, W.B. (1995). Evaluation of a susceptibility gene for schizophrenia on chromosome 6p by multipoint affected sib-pair linkage analysis, *Nature Genetics*, **11**, 325–327.

23. Siegmund, D. (2001) Is peak height sufficient? *Genetic Epidemiology*, **20**, 403–408.

24. Simon-Sanchez, J., Scholz, S., Del Mar Matarin, M., Fung, H. C., Hernandez, D., Gibbs, J. R., Britton, A., Hardy, J., and Singleton, A. (2007). Genome-wide SNP assay reveals mutations underlying Parkinson disease, *Human Mutation*, **9**, 9.

25. Terwilliger, J. D., Shannon, W. D., Lathrop, G. M., Nolan, J. P., Goldin, L. R., Chase, G. A., and Weeks, D. E. (1997). True and false positive peaks in genome-wide scans: applications of length-biased sampling to linkage mapping, *American Journal of Human Genetics*, **61**, 430–438.

10

On Probabilities for Complex Switching Rules in Sampling Inspection

W.Y. Wendy Lou[1] **and James C. Fu**[2]

[1] *Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada*
[2] *Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada*

Abstract: Switching rules between different levels of sampling are widely used in quality control, as in the well-known Military Standard 105E (MIL STD 105E) and similar acceptance-sampling schemes. These switching rules are typically defined by specific patterns of inspection outcomes within a sequence of previous inspections. The probability distributions of the rules are usually hard to find, and many of them remain unknown. In this chapter, we will provide a general and simple method, the finite Markov chain imbedding technique, to obtain the distributions of switching rules. We demonstrate the utility of this method primarily by (i) deriving the generating function of a basic switching rule (k consecutive acceptances) for the practically important case of a two-state, first-order autoregressive AR(1) sequence, (ii) treating jointly the normal and tightened inspection regimes of MIL STD 105E including the overall probability of discontinuing inspection, and (iii) considering a stratified sampling scheme with three possible inspection outcomes.

Keywords and phrases: Runs and patterns, finite Markov chain imbedding, acceptance sampling, switching rules

10.1 Introduction

Acceptance sampling is an important tool of statistical quality control. Rather than 100% inspection of lots, only random samples are tested and used as the basis for lot sentencing (acceptance or rejection). In the simplest and most common case, only one random sample is tested per lot, and sample items are classified by attributes (*e.g.* pass/fail, or poor/acceptable/good) as opposed to

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_10,

variables on a quantitative scale; such procedures are called "single-sampling plans by attributes." Montgomery (2001) gives a detailed review of acceptance sampling.

Military Standard 105E (MIL STD 105E) of the United States Department of Defense, and its nearly equivalent civilian counterpart ANSI/ASQ Z1.4-2003, have generally been the most prevalent set of acceptance-sampling plans for attributes worldwide. Contained within these (and other) sampling systems are switching rules between different sampling plans to allow for normal, tightened, and reduced levels of inspections, depending on the vendor's recent quality history. Switching rules are usually defined by the occurrences of specific patterns in the outcomes of recent inspections, and are often designed based more on empirical experience rather than exact probabilistic considerations.

There are a few theoretical and numerical analyses of switching rules in the literature, including Dodge (1963, 1965), Brown and Rutmiller (1975), and Schilling and Sheesley (1978a and 1978b). Hald (1981) derived the probability generating functions for some switching rules using the theory of recurrent events [Feller (1968)], but only means and variances of the distributions were computed due to the complexity of the generating functions; Shmueli and Cohen (2000) used the method of partial fraction expansion to obtain the exact probability distributions of several switching rules from their known generating functions. However, there are many switching rules with generating functions and probability distributions that remain unknown, especially when considering several switching rules jointly (*e.g.* normal, tightened, and discontinuation rules in MIL STD 105E) to examine the overall distributions of waiting times. This has led to the introduction of alternatives to MIL STD 105E in the Japanese counterpart [see Koyama *et al.* (1970), Koyama (1978), and Hald (1981)]. Furthermore, with regard to MIL STD 105E, Montgomery (2001) states the following:

> *In particular, some engineers dislike the switching rules because there is often a considerable amount of misswitching from normal to tightened or normal to reduced inspection when the process is actually producing lots of AQL* (acceptable quality level) *quality. Also, there is a significant probability that production would even be discontinued, even though there has been no actual quality deterioration* (p. 714).

There is an urgent need to study the probability structures of complex switching rules and their interactions within sampling systems.

Let's consider a sequence of lots under the level of tightened inspection, and a switching rule for shifting to normal inspection that is defined by the occurrence of the pattern $\Lambda = AA \ldots A$ of k consecutive lots accepted under tightened sampling inspection. Let $W(\Lambda)$ be the waiting time for switching

from tightened to normal inspection. Under the assumption that the sequence of inspections is one of Bernoulli trials, Feller (1968), using the theory of recurrent events, provided the generating function for the waiting time $W(\Lambda)$:

$$\varphi_w(s) = \frac{p^k s^k (1 - ps)}{1 - s + qp^k s^{k+1}}, \tag{10.1}$$

where $0 < p < 1$ is the probability of acceptance. Hirano (1986) and Philippou and Makri (1986), independently using combinatory analysis, obtained the exact distribution of $W(\Lambda)$:

$$P(W(\Lambda) > n) = \sum_{m=0}^{k-1} \sum_{x_1+2x_2+\cdots+kx_k=n-m} \binom{x_1 + \cdots + x_k}{x_1, x_2, \cdots, x_k} p^n \left(\frac{q}{p}\right)^{x_1+\cdots+x_k}. \tag{10.2}$$

Recently, a sequence of papers by Fu and Koutras (1994), Koutras and Alexandrou (1995), Fu (1996), and Lou (1996) treated the distributions of runs and patterns using the finite Markov chain imbedding technique. In another sequence of papers, Hirano and Aki (1993), Han and Aki (1998), and Aki and Hirano (1999) studied the distributions of runs and patterns via the method of conditional probability generating functions. Since most current switching rules are based either on simple or compound patterns, in this chapter, we will present a general and numerically efficient method based on the finite Markov chain imbedding technique for studying the distributions of switching rules not only under the assumption that the sampling inspections $\{X_i\}$ are Bernoulli trials, but also allowing for two-state AR(1) trials and multi-state Markov-dependent trials. This paper is organized as follows. Section 10.2 introduces the notation and the finite Markov chain imbedding. The main results are presented in Section 10.3. Section 10.4 provides four numerical examples to illustrate the implementation of the results. Section 10.5 discusses more general cases and extensions.

10.2 Notation and Finite Markov Chain Imbedding

Suppose the quality of each lot is classified into m ($m \geq 2$) levels denoted by the m symbols $a_1, a_2, \ldots, a_m$. Let $\mathcal{S} = \{a_1, \ldots, a_m\}$, and $X_1, \ldots, X_n$ be a sequence of sampling inspections with outcomes in $\mathcal{S}$.

Definition 10.2.1 A pattern Λ is called a simple pattern if Λ is composed of a specified sequence of k symbols: *i.e.* $\Lambda = a_{i_1} a_{i_2} \cdots a_{i_k}$, where the length of the pattern k is finite, and the symbols are allowed to be repeated.

Let Λ_1 and Λ_2 be two distinct simple patterns (neither Λ_1 nor Λ_2 is a subsequence of the other). We define the union of Λ_1 and Λ_2, $\Lambda = \Lambda_1 \bigcup \Lambda_2$, as the occurrence of either Λ_1 or Λ_2.

Definition 10.2.2 Λ is called a compound pattern if it is a union of l $(2 \leq l < \infty)$ distinct simple patterns $\Lambda_1, \Lambda_2, \ldots, \Lambda_l$, *i.e.* $\Lambda = \bigcup_{i=1}^{l} \Lambda_i$.

Throughout this chapter, we shall consider only switching rules that are based purely on simple or compound patterns. In MIL STD 105E, for example, the switching rules to and from the reduced level of inspection involve several qualitative conditions unrelated to the inspection outcomes, such as whether the production is deemed steady or irregular; most switching rules, however, are entirely pattern based.

Given a pattern Λ, let $X_n(\Lambda)$ be the number of patterns Λ that have occurred in the sequence of sampling inspections $\{X_i\}_{i=1}^{n}$. Define the indicator random variable

$$I_n(\Lambda) = \begin{cases} 0 \text{ if } X_n(\Lambda) = 0 \\ 1 \text{ if } X_n(\Lambda) > 0, \end{cases} \tag{10.3}$$

where $I_n(\Lambda) = 0$ means that no pattern Λ has occurred in the sequence $\{X_i\}_{i=1}^{n}$, and $I_n(\Lambda) = 1$ means that there has been at least one occurrence of the pattern Λ. It has been shown that the random variables $X_n(\Lambda)$ and $I_n(\Lambda)$ are finite Markov chain imbeddable in the following sense.

Definition 10.2.3 A non-negative integer random variable $X_n(\Lambda)$ is finite Markov chain imbeddable if there exists

(a) a finite Markov chain $\{Y_t : t = 1, 2, \ldots, n\}$ defined on a finite state space $\Omega = \{b_1, \ldots, b_r\}$ with initial probability ξ_0,

(b) a finite partition $\{C_x : x = 0, 1, \ldots, l\}$ on the state space Ω, and

(c) for every $x = 0, 1, \ldots, l$, a probability such that

$$P(X_n(\Lambda) = x|\xi_0) = P(Y_n \in C_x|\xi_0). \tag{10.4}$$

To simplify the notation, unless specified otherwise, the probability $P(X_n(\Lambda) = x|\xi_0)$ will be written as $P(X_n(\Lambda) = x)$ throughout this manuscript.

10.3 Main Results

Let Λ be the pattern corresponding to a specific switching rule. Then the following result holds.

Theorem 10.3.1 *If Λ is a simple (or compound) pattern of length k and $\{X_i\}_{i=1}^n$ is a sequence of homogeneous Markov dependent (or i.i.d.) multi-state trials, then $I_n(\Lambda)$ is finite Markov chain imbeddable, and the distribution of the waiting time $W(\Lambda)$ for the switching rule is given by, for $n = k, k+1, \ldots,$*

$$P(W(\Lambda) = n) = (1, 0, \ldots, 0)M^{n-1}(I - M)(1, 1, \ldots, 1, 0)', \tag{10.5}$$

where M is a $(k+1) \times (k+1)$ transition probability matrix associated with an imbedded homogeneous Markov chain $\{Y_t : t = 0, 1, \ldots, n\}$.

The detailed proof of the above general theorem can be seen from Fu (1996); here we omit the proof. The above result can also be extended as follows to the case where $\{X_i\}_{i=1}^n$ corresponds to a sequence of independent but not identically distributed multi-state sampling inspections:

$$P(W(\Lambda) = n) = (1, 0, \ldots, 0)\left(\prod_{t=1}^{n-1} M_t\right)(I - M_n)(1, 1, \ldots, 1, 0)', \tag{10.6}$$

where the imbedded Markov chain $\{Y_t\}$ has transition probabilities M_t, $t = 0, 1, \ldots, n$.

Next we show how the finite Markov chain imbedding technique can be applied in cases where the sequence of inspections is taken as two-state AR(1)-dependent trials (autocorrelation is often found in manufacturing processes for sufficiently small lot sizes—see Montgomery, 2001). In particular, for the common switching rule defined by the occurrence of the pattern $\Lambda = AA \cdots A$ of k ($k = 1, 2, \ldots$) consecutive acceptances, we show how to use this technique to readily derive the probability generating function for its waiting time $W(\Lambda)$.

Let $\{X_i\}$ be a sequence of two-state AR(1) trials with autocorrelation coefficient ρ: *i.e.*

$$\rho = \frac{Cov(X_i, X_{i+1})}{Var(X_i)}, \ \forall i = 1, \ldots, n-1,$$

where the two states, acceptance (A) and rejection (R), carry the marginal probabilities $P(X_i = A) = p$ and $P(X_i = R) = q$, respectively, for $i = 1, 2, \ldots, n$. For clarity, to distinguish the transition probabilities p_{ij} of this dependent sequence, we use 1 and 0 as the subscripts instead of A and R, respectively.

Theorem 10.3.2 *If $\{X_i\}$ is a sequence of two-state AR(1) sampling inspection outcomes, then*

(i) for $\rho \neq 0$, the probability generating function for $W(\Lambda)$ is

$$\varphi_{W(\Lambda)}(t) = \frac{(p - pp_{00}t + qp_{01}t)p_{11}^{k-1}t^k}{1 - p_{00}t - \sum_{i=2}^{k} p_{11}^{i-2}p_{10}p_{01}t^i}, \tag{10.7}$$

where $p_{11} = 1 - p_{10} = p + \rho(1-p)$ and $p_{00} = 1 - p_{01} = (1-p) + \rho p$,

(ii) for $\rho = 0$, the recursive equation for $P(W(\Lambda) = n)$ is given by

$$P(W(\Lambda) = n) = \sum_{i=1}^{k} qp^{i-1} P(W(\Lambda) = n - i), \tag{10.8}$$

(iii) for $\rho = 0$,

$$\varphi_{W(\Lambda)}(t) = \frac{p^k t^k}{1 - \sum_{i=1}^{k} qp^{i-1} t^i}. \tag{10.9}$$

Proof. By definition, $\rho = Cov(X_i, X_{i-1})/Var(X_i)$, so that $p_{11} = p + \rho(1-p)$. Further, using the ergodic property of the sequence $\{X_i\}$,

$$(p, 1-p) \begin{pmatrix} p + \rho(1-p) & 1 - p - \rho(1-p) \\ 1 - p_{00} & p_{00} \end{pmatrix} = (p, 1-p),$$

which yields $p_{00} = (1-p) + \rho p$. This implies that $\{X_i\}_{i=1}^n$ is also a two-state Markov chain with transition matrix

$$\mathcal{A} = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} \equiv \begin{pmatrix} p + \rho(1-p) & 1 - p - \rho(1-p) \\ p - \rho p & (1-p) + \rho p \end{pmatrix}. \tag{10.10}$$

Since $W(\Lambda)$ is finite Markov chain imbeddable, we can construct a corresponding finite Markov chain $\{Y_t\}$ defined on the state space

$$\begin{aligned} \Omega &= \{\emptyset\} \cup \{A, R\} \cup \{A, AA, \ldots, \underbrace{A \ldots A}_{k-1}, \underbrace{A \ldots A}_{k} = \alpha\} \\ &= \{\emptyset, R, A, AA, \ldots, \underbrace{A \ldots A}_{k-1}, \underbrace{A \ldots A}_{k} = \alpha\}, \end{aligned}$$

where $\emptyset$ stands for the dummy state and α for the absorbing state, with the transition probability matrix

$$M = \begin{array}{c} \emptyset \\ R \\ A \\ AA \\ \vdots \\ A \cdots A \\ \alpha \end{array} \begin{pmatrix} 0 & q & p & 0 & \cdot & \cdot & 0 & 0 \\ 0 & p_{00} & p_{01} & 0 & \cdot & \cdot & 0 & 0 \\ 0 & p_{10} & 0 & p_{11} & \ddots & \cdot & \cdot & \cdot \\ 0 & p_{10} & 0 & 0 & p_{11} & \ddots & \cdot & \cdot \\ \vdots & \vdots & \cdot & \cdot & 0 & \cdot & \vdots & \vdots \\ 0 & p_{10} & 0 & \cdot & \cdots & \cdot & 0 & p_{11} \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} = \left(\begin{array}{c|c} N & C \\ \hline 0 & 1 \end{array}\right), \tag{10.11}$$

where p_{ij}, $i, j = 0, 1$, are defined by Equation (10.10). The dummy state $\emptyset$ is used for handling the initial distribution ξ_0, especially when $\{X_i\}$ is Markov dependent. It follows from the general result in Theorem 10.3.1 that

$$P(W(\Lambda) > n) = P(Y_n \in \Omega - \alpha | \xi_0) = \xi_0 N^n 1', \tag{10.12}$$

where $\xi_0 = (1, 0, \ldots, 0.\ )_{1\times(k+1)}$, $1 = (1, 1, \ldots, 1)_{1\times(k+1)}$, and N is given by Equation (10.11). Applying Theorem 10.3.1, a result by Fu and Chang (2002, p. 74), and the matrix N defined by Equation (10.11), we have

$$\varphi_{W(\Lambda)}(t) = 1 + (1 - \frac{1}{t})\Phi_W(t), \tag{10.13}$$

$$\Phi_W(t) = \phi_1(t) + \cdots + \phi_{k+1}(t), \tag{10.14}$$

where $(\phi_1(t), \ldots, \phi_{k+1}(t))$ is the solution of the simultaneous equations

$$\phi_i(t) = t\xi_0 e_i^{'} + t(\phi_1(t), \ldots, \phi_{k+1}(t))N(i) \tag{10.15}$$

for $i = 1, \ldots, k+1$, with $N(i)$ denoting the i-th column of the matrix N. After some simple algebra, this yields

$$\varphi_{W(\Lambda)}(t) = \frac{(p - pp_{00}t + qp_{01}t)p_{11}^{k-1}t^k}{1 - p_{00}t - \sum_{i=2}^{k} p_{11}^{i-2}p_{10}p_{01}t^i}.$$

This completes the proof of (i).

For $\rho = 0$, $\{X_i\}_{i=1}^n$ is a sequence of Bernoulli trials. Inserting $p_{11} = p_{01} = p$ and $p_{00} = p_{10} = q$ into the matrix M in Equation (10.11), the result (ii) follows from Equation (10.12) and

$$P(W(\Lambda) = n) = P(Y_{n-1} \in \Omega - \alpha) - P(Y_n \in \Omega - \alpha) = \xi_0 N^{n-1} 1^{'} - \xi_0 N^n 1^{'},$$

with $e_1 N = qe_2 + pe_3$ and $e_i N = qe_2 + pe_{i+1}$, for $i = 2, \ldots, k+1$. The result (iii) follows directly from the result (i) by taking $p_{11} = p_{01} = p$ and $p_{00} = p_{10} = q$. This completes the proof. ■

The result (ii) yields the equation

$$\varphi_{W(\Lambda)}(t) = t^k p^k + \sum_{i=1}^{k} qp^{i-1}t^i \varphi_{W(\Lambda)}(t). \tag{10.16}$$

Hence the result (iii) is also an immediate consequence of result (ii). With some modifications, it can be extended to switching rules based on compound patterns. The probability of $W(\Lambda) = n$ can also be obtained by using the probability generating function (10.7), since

$$P(W(\Lambda) = n) = \frac{1}{n!}\, \varphi_{W(\Lambda)}^{(n)}(t)|_{t=0},$$

where $\varphi_{W(\Lambda)}^{(n)}(t)$ stands for the n-th derivative of $\varphi_{W(\Lambda)}(t)$. Several applications of our methodology are presented in the following section.

10.4 Numerical Examples of Switching Rules

10.4.1 Example 1: Tightened to normal inspection

Let us consider the common switching rule from tightened to normal inspection defined as the occurrence of k consecutive acceptances ($\Lambda = A \ldots A$) of lots under tightened inspection. We assume the sequence $\{X_i\}$ follows a two-state, AR(1) model with autocorrelation coefficient ρ and ergodic probabilities $P(X = A) = p$ and $P(X = R) = q$. Based on the methodology presented in the previous section, Table 10.1 provides the waiting time distributions $W(\Lambda)$ for some selected ρ, p, and k.

10.4.2 Example 2: Normal to tightened inspection

Let us consider a switching rule from normal to tightened sampling inspection defined as the occurrence of 2 rejected lots out of a maximum of k consecutive lots under normal inspection. For $k = 5$, this is the switching rule from normal to tightened inspection in MIL STD 105E. Let $\Lambda_1 = RR$, $\Lambda_2 = RAR, \ldots,$ $\Lambda_{k-1} = RA \ldots AR$ be $(k-1)$ distinct simple patterns, and $\Lambda = \bigcup_{i=1}^{k-1} \Lambda_i$ be a compound pattern generated by the Λ_i. Hence, the above switching rule from normal to tightened inspection is *equivalent* to the compound pattern Λ having occurred in a maximum of k consecutive lots under normal sampling inspection.

Table 10.1. Distribution of $W(\Lambda)$ for some selected ρ, p and k in Example 1.

$k = 2$, $p = 0.95$			$k = 5$, $p = 0.8594$			$k = 3$, $p = 0.5$		
n	$\rho = 0$	$\rho = 0.5$	n	$\rho = 0$	$\rho = 0.3$	n	$\rho = 0$	$\rho = 0.3$
2	0.9025	0.9263	5	0.4688	0.5678	3	0.1250	0.2113
3	0.0451	0.0233	6	0.0659	0.0559	4	0.0625	0.0739
4	0.0451	0.0233	7	0.0659	0.0559	5	0.0625	0.0739
5	0.0044	0.0124	8	0.0659	0.0559	6	0.0625	0.0739
6	0.0024	0.0068	9	0.0659	0.0559	7	0.0547	0.0630
7	0.0003	0.0037	10	0.0659	0.0559	8	0.0508	0.0559
			11	0.0350	0.0337	9	0.0469	0.0499
			12	0.0307	0.0248	10	0.0430	0.0443
			13	0.0263	0.0200	11	0.0396	0.0394
			14	0.0220	0.0163	12	0.0364	0.0350
			15	0.0176	0.0131	13	0.0334	0.0311
			16	0.0133	0.0099	14	0.0308	0.0276
			17	0.0110	0.0076	15	0.0283	0.0246

The waiting time $W(\Lambda)$ is the sooner waiting time of the individual waiting times $W(\Lambda_i)$, $i = 1, \ldots, k$, *i.e.*

$$W(\Lambda) = \inf\{W(\Lambda_i), i = 1, \ldots, k\}. \tag{10.17}$$

It is easy to see that there are $k-1$ distinct intermediate states R, RA, RAA, ..., $R\underbrace{A \ldots A}_{k-2}$ generated by the $(k-1)$ simple patterns Λ_i, $i = 1, \ldots, k-1$. Define a state space $\Omega = \{0, 1, \ldots, k-1, \alpha\}$, where the states $0 \equiv A$, $1 \equiv R$, $2 \equiv RA$, ..., $k-1 \equiv RA \ldots A$, are referred to as ending blocks. We also define a finite Markov chain $\{Y_t\}$ on Ω as

$$Y_t = \begin{cases} 0 \text{ if no pattern } \Lambda \text{ has occurred in the first } t \text{ trials and} \\ \quad \text{no } R \text{ has occurred among the } X_{t-k+1}, \ldots, X_t \text{ trials,} \\ \text{E if no pattern } \Lambda \text{ has occurred in the first } t \text{ trials} \\ \quad \text{and } X_{t-E+1} = R,\ X_{t-E+2} = \cdots = X_t = A, \\ \alpha \text{ if a pattern } \Lambda \text{ has occurred in the first } t \text{ trials,} \end{cases} \tag{10.18}$$

where $E = 1, \ldots, k-1$, and α is an absorbing state. To make our definition of Y_t in Equation (10.18) more transparent, we present the following example. Consider $k = 5$ and the 15 outcomes of normal inspections $ARAAAARAAAARARR$. Then the realization of the Markov chain $\{Y_t : t = 1, 2, \ldots, 15\}$ is $\{Y_1 = 0, Y_2 = 1, Y_3 = 2, Y_4 = 3, Y_5 = 4, Y_6 = 0, Y_7 = 1, Y_8 = 2, Y_9 = 3, Y_{10} = 4, Y_{11} = 0, Y_{12} = 1, Y_{13} = 2, Y_{14} = \alpha, Y_{15} = \alpha\}$. If $\{X_i\}$ consists of i.i.d. two-state trials with probability q of rejecting a lot, then the Markov chain Y_t has transition probability matrix

$$M = \begin{array}{c} 0 \\ 1 \\ 2 \\ \cdot \\ \cdot \\ k-1 \\ \alpha \end{array} \begin{bmatrix} p & q & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & p & 0 & \cdots & 0 & 0 & q \\ 0 & 0 & 0 & p & \cdots & 0 & 0 & q \\ & & & \ldots\ldots\ldots\ldots\ldots & & & & \\ 0 & 0 & 0 & \cdots & \cdots & 0 & p & q \\ p & 0 & \cdots & \cdots & \cdots & 0 & 0 & q \\ 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}. \tag{10.19}$$

The distribution of the waiting time $W(\Lambda)$ for this switching rule can then be computed by using Equation (10.5) with $M_t = M$ for all $t = 1, \ldots, n$, where the transition probability matrix M is given by Equation (10.19), or by using the recursive equation in Theorem 10.3.2(ii). Samples of this distribution for $k = 5$, in accordance with MIL STD 105E, are given in Table 10.2.

10.4.3 Example 3: Discontinuation of inspection

In this example, we wish to demonstrate the utility of the finite Markov chain imbedding technique by considering the waiting-time distribution for discontinuation of inspections in MIL STD 105E, a distribution for which the generating

Table 10.2. Samples of the waiting time distribution of $W(\Lambda)$ in Example 2 with $\Lambda = \bigcup_{i=1}^{4} \Lambda_i$ and $k = 5$.

n	$p = 0.789$	$p = 0.9$	$p = 0.95$
2	0.0445	0.0100	0.0025
3	0.0703	0.0180	0.0048
4	0.0831	0.0243	0.0068
5	0.0875	0.0292	0.0086
6	0.0690	0.0262	0.0081
7	0.0581	0.0243	0.0078
8	0.0516	0.0230	0.0076
9	0.0475	0.0223	0.0075
10	0.0446	0.0220	0.0075

function is not known. Since switching to/from a reduced level of inspection involves various inspection-independent conditions, such as on the regularity of production, only the normal and tightened inspection regimes are considered here. Switching from normal to tightened inspection occurs when two out of five consecutive lots are rejected, as treated in Example 2. From tightened inspection, a switch back to the normal regime requires the acceptance of five consecutive lots; on the other hand, inspection is discontinued (and not restarted until corrective action has been taken) if the cumulative number of rejections under the tightened regime reaches five.

As described in the book by Montgomery (2001), in using MIL STD 105E, one first selects the AQL, determines the lot size to be inspected, and chooses the General Inspection Level (I, II, or III). Let's suppose we desire an AQL of 1 percent, have a total lot size to be inspected at each inspection of 10,000, and choose Inspection Level I, the least stringent level. Then, based on the tables of the standard, a random sample of the lot of size $N = 80$ should be examined, and the acceptance number, the allowable number of rejections within the sample, is either $c_n = 2$ under normal inspection, or $c_t = 1$ under tightened inspection. The actual distribution of the number of defectives in a random sample of N items is approximately binomial with parameters N and f, where f is the fraction of defective items in the lot. Here we focus on the case where the manufacturer is producing so that f is equal to the AQL, *i.e.* $f = 0.01$. The probabilities p_n and p_t that the observed number of defectives, d, is less than c_n and c_t under the normal and tightened regimes, respectively, may then be determined using the binomial distribution as $p_n = P\{d \leq 2\} = 0.953$ and $p_t = P\{d \leq 1\} = 0.809$. The probabilities p_n and p_t represent the acceptance/success probabilities of the sequence assuming the manufacturer is producing at exactly the desired AQL. In general,

$p_t < p_n$, so that rejection is more likely under tightened as opposed to normal inspection. For Inspection Level II, a similar calculation yields $p_n = 0.984$, $p_t = 0.858$, and for Inspection Level III, the most stringent level, $p_n = 0.985$, $p_t = 0.901$.

Hence, we are considering here a two-state sequence of inspections that start off with success probability p_n (normal inspection), and then if two out of five trials are failures, the success probability decreases to $p_t < p_n$ (tightened inspection). Then, if there are five consecutive successes under the tightened inspection, the success probability returns to p_n (normal inspection); on the other hand, if there are a total of five failures under tightened inspection, then inspection is discontinued. To determine the distribution of the waiting time for discontinuation of inspection, we define the following state space:

$$\Omega = \{N, A, R, RA, RAA, RAAA, T, (S_1, S_2), \alpha\},$$

where $S_1 \in \{A, AA, AAA, AAAA\}$ and $S_2 \in \{0, 1, 2, 3, 4\}$. This state space differs from that of Example 2 by the addition of states N and T to account for the switching between normal and tightened regimes, and by the two-dimensional array of states (S_1, S_2) to capture the switching from tightened inspection to normal inspection or to the absorbing state α. The two coordinates of the array, S_1 and S_2, represent counters for the number of consecutive acceptances and for the total number of rejections, respectively, within the tightened regime. The total number of states is 32. A Markov chain $\{Y_t\}$ to obtain the distribution of the waiting time for discontinuation of inspection, $P[W(D) = n]$, may be constructed in a manner similar to that in Example 2, and we omit the details here.

Figure 10.1 shows the distribution of $P[W(D) = n]$ versus the number of sampling inspections n for the three inspection levels discussed above, where $n = 1$ is the first inspection under the normal level of sampling. After about 25 inspections, the distribution for each inspection level follows a very slow exponential decay governed by a single time-constant, as implied by the straight line of slightly negative slope on the semi-log plot shown. This is to be expected, since in repeated matrix multiplications the largest relevant eigenvalue will eventually dominate the multiplications (see, for example, Fu, Wang, and Lou 2003). The expected values of the waiting time for discontinuation of inspection, $W(D)$, are $EW(D) = 1271$, 24421, and 109381, for Inspection Levels I, II, and III, respectively (rounded to the nearest integer). Based on the results from this example, it appears that, somewhat contrary to the engineers' suspicions quoted in Section 10.1, the probabilities of discontinuing inspections within MIL STD 105E are sufficiently small for practical purposes, even under the least stringent Inspection Level I, when the lots are produced so that the fraction of defectives matches the AQL.

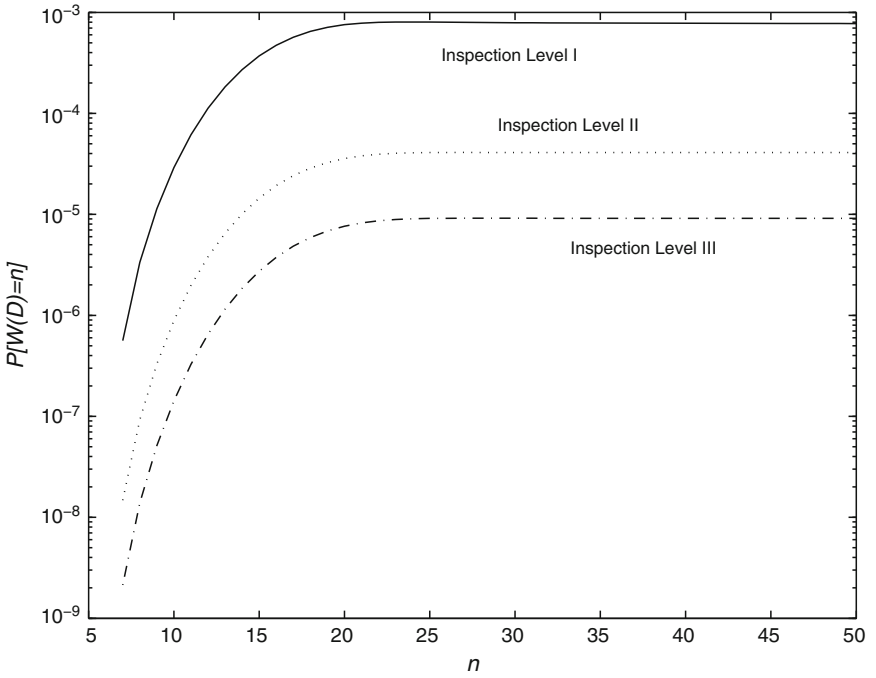


Figure 10.1. The distributions of the waiting time for discontinuation of inspection, $P[W(D) = n]$ versus n for the three inspection levels of MIL STD 105E in Example 3. For Inspection Level I, $p_n = 0.953$, $p_t = 0.809$, and $EW(D) = 1271$; for Inspection Level II, $p_n = 0.984$, $p_t = 0.858$, and $EW(D) = 24421$; and for Inspection Level III, $p_n = 0.985$, $p_t = 0.901$, and $EW(D) = 109381$.

10.4.4 Example 4: Three-level modeling

Traditionally, each lot under sampling inspection is sentenced as either a rejection or an acceptance. For more general practical applications, we may assume that each lot is classified into one of the three Levels 1, 2, and 3, where Level 1 indicates highest quality, Level 2 indicates intermediate quality, and Level 3 indicates lowest quality; *i.e.* $X_i = 1$, 2, or 3, for $i = 1, \ldots, n$. Consider a more complex switching rule from normal to tightened inspection defined as the occurrence of l, where $l = 2$ or 3, consecutive lots whose qualities are at Level 2 or Level 3, and whose sum is greater than or equal to 6. Mathematically, this switching rule can be stated as the existence of an i and an l such that

$$\{X_i + X_{i+1} + \cdots + X_{i+l-1} \geq 6, X_j \geq 2, i \leq j \leq i + l - 1\}. \tag{10.20}$$

This switching rule may be desirable from a practical point of view, but such complex rules are often not used in practice due to the difficulty in determining

their probabilistic structures. Here we show how the finite Markov chain imbedding technique can be readily applied to complex switching rules.

Let $\Lambda_1 = 33$, $\Lambda_2 = 323$, $\Lambda_3 = 322$, $\Lambda_4 = 232$, $\Lambda_5 = 233$, $\Lambda_6 = 223$, and $\Lambda_7 = 222$ be seven simple patterns, and $\Lambda = \bigcup_{i=1}^{7} \Lambda_i$ be the compound pattern generated by the seven simple patterns. The switching rule from normal to tightened inspection defined by Equation (10.20) can be viewed as the occurrence of the pattern Λ under normal inspection.

The intermediate states generated by the seven simple patterns are 2, 3, 22, 23, and 32. Let $\Omega = \{1, 2, 3, \beta, \alpha\}$ be the state space with five states, where β represents the intermediate states 22, 23, and 32, and the absorbing state α represents the simple patterns $\Lambda_1, \Lambda_2, \ldots, \Lambda_7$. With respect to a sequential counting procedure, we define the imbedded Markov chain $\{Y_t : t = 1, 2, \ldots, n\}$ as

$$Y_t = \begin{cases} E \text{ if no pattern } \Lambda \text{ has occurred in the first } t \text{ inspections} \\ \quad \text{with ending block } E,\ E = 1, 2, 3, \beta, \\ \alpha \text{ if the pattern } \Lambda \text{ has occurred in the first } t \text{ normal inspections,} \end{cases} \tag{10.21}$$

for $t = 1, \ldots, n$.

Let's consider the following sequence of outcomes of ten normal inspections: 1221321332. It follows from the definition of Y_t that we have $Y_1 = 1$, $Y_2 = 2$, $Y_3 = \beta$, $Y_4 = 1$, $Y_5 = 3$, $Y_6 = \beta$, $Y_7 = 1$, $Y_8 = 3$, $Y_9 = \alpha$, and $Y_{10} = \alpha$. Furthermore, $I_t(\Lambda) = 0$ for $t = 1, \ldots, 8$, $I_t(\Lambda) = 1$ for $t = 9, 10$, and the switching time $W(\Lambda) = 9$. Assuming the sequence $\{X_i\}_{i=1}^{n}$ consists of i.i.d. trials having probabilities p_1, p_2, and p_3 for the three quality levels 1, 2, and 3, respectively, then the imbedded Markov chain $\{Y_t : t = 0, 1, 2, \ldots, n\}$ has the following transition probability matrix:

$$M = \begin{array}{c} 1 \\ 2 \\ 3 \\ \beta \\ \alpha \end{array} \begin{bmatrix} p_1 & p_2 & p_3 & 0 & 0 \\ p_1 & 0 & 0 & p_2 + p_3 & 0 \\ p_1 & 0 & 0 & p_2 & p_3 \\ p_1 & 0 & 0 & 0 & p_2 + p_3 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{10.22}$$

The distribution of the waiting time for this switching rule can be computed by

$$P(W(\Lambda) = n) = (1, 0, 0, 0, 0) M^{n-1} (I - M)(1, 1, 1, 1, 0)', \tag{10.23}$$

for $n = 2, 3, \ldots$, where M is given by Equation (10.22). For example, the distribution of $W(\Lambda)$ for $p_1 = 0.7$, $p_2 = 0.2$, and $p_3 = 0.1$ is shown in Figure 10.2.

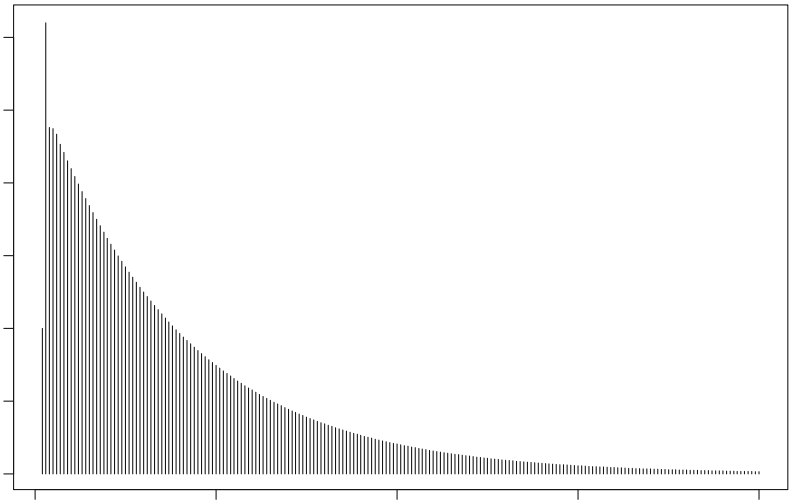


Figure 10.2. Distribution of $W(\Lambda)$ for Example 4 with $\Lambda = \bigcup_{i=1}^{7} \Lambda_i$ and $p_1 = 0.7, p_2 = 0.2, p_3 = 0.1$.

10.5 Summary and Discussion

In view of the examples in Section 10.4, the imbedded Markov chain $\{Y_t\}$ is defined by whether the sequence contains the pattern of interest Λ and by suitable ending blocks E that keep track of intermediate states. The construction of the imbedded Markov chain $\{Y_t\}$ is simple and direct. Regardless of what the switching rule is, as long as it is defined by a simple or compound pattern, the imbedded Markov chain $\{Y_t\}$ always exists and the formula for computing the distribution remains the same. In Example 1, if the sequence $\{X_i\}$ of inspections has a Markov chain structure, with only simple modifications to the transition probability matrix M, the exact distribution of the waiting time $W(\Lambda)$ could also be computed with the same formula. This is a great advantage of the finite Markov chain imbedding technique, one which would be difficult to achieve by any other method.

Since Equation (10.6) involves only the multiplication of the transition probability matrix n times, the size of the state space Ω (or M) is a good index for the speed of computation. For most switching rules, the size of the transition probability matrix M is rather small. All of our numerical results given in Section 10.4 were computed on an average PC with the S-plus program, and the CPU times were only a fraction of a second. With the speed of today's computers, the accuracy and computability of exact distributions, means, and variances for switching rules should not be an issue in most applications.

In general, the imbedded Markov chain $\{Y_t\}$ associated with the specified switching rules is not unique. For the switching rule of Example 4, we could define a Markov chain $\{Y_t^\star\}$ on the state space $\Omega^\star = \{1, 2, 3, 22, 23, 32, \text{and } \alpha\}$ with transition probability matrix

$$M^\star = \begin{array}{c} 1 \\ 2 \\ 3 \\ 22 \\ 23 \\ 32 \\ \alpha \end{array} \begin{bmatrix} p_1 & p_2 & p_3 & 0 & 0 & 0 & 0 \\ p_1 & 0 & 0 & p_2 & p_3 & 0 & 0 \\ p_1 & 0 & 0 & 0 & 0 & p_2 & p_3 \\ p_1 & 0 & 0 & 0 & 0 & 0 & p_2 + p_3 \\ p_1 & 0 & 0 & 0 & 0 & 0 & p_2 + p_3 \\ p_1 & 0 & 0 & 0 & 0 & 0 & p_2 + p_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The distributions of the switching rule obtained from two different imbedded Markov chains $\{Y_t\}$ and $\{Y_t^\star\}$ are the same. However, the size of the state space Ω is smaller than the size of $\Omega^\star$, and from a computational point of view, the imbedded Markov chain $\{Y_t\}$ is more efficient.

Our results could also be extended to switching rules defined by the multiple occurrence of a pattern, say m times. For example, consider a switching rule from tightened to normal sampling inspection defined as $m = 2$ occurrences of k consecutive acceptances (using non-overlap counting) under tightened inspection. An imbedded Markov chain $\{Y_t\}$ associated with this switching rule could be defined on the state space

$$\Omega = \{(0,0), (0,1), \ldots, (0,k-1), (1,0), \ldots, (1,k-1), \alpha\}.$$

We leave the details of constructing such a Markov chain and its transition probability matrix to the reader. Further, and more generally, this method could also be extended to the case of switching rules defined by the occurrences of a sequence of m specified simple or compound patterns. For example, let $\Lambda_1, \Lambda_2, \ldots, \Lambda_m$ be compound patterns. The waiting time for the sequence of patterns $\Lambda_1, \Lambda_2, \ldots, \Lambda_m$ is very different from the waiting time for a compound pattern $\Lambda = \cup_{i=1}^m \Lambda_i$.

The finite Markov chain imbedding technique is a useful tool for determining the probabilistic structure of complex switching rules in sampling inspection, and even exact results for dependent sequences of inspection outcomes can be readily computed. Other possible applications of this technique in the area of quality control lie in the study of complex quality control charts, such as the Shewhart chart with runs rules (see Koutras, Bersimis, and Maravelakis, 2007).

Acknowledgments

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program.

References

1. Aki, S. and Hirano, K. (1999). Sooner and later waiting time problems for runs in Markov dependent bivariate trials, *Annals of the Institute of Statistical Mathematics*, **51**, 17–29.
2. Brown, G. G. and Rutmiller, H. C. (1975). An analysis of the long range operating characteristics of the MIL-STD-105D sampling scheme and some suggested modifications, *Naval Research Logistics Quarterly*, **22**, 667–679.
3. Dodge, H. F. (1963). A general procedure for sampling inspection by attributes-based on the AQL concept, *ASQC Annual Convention Transactions 1963*, 7–19.
4. Dodge, H. F. (1965). Evaluation of a sampling inspection system having rules for switching between normal and tightened inspection, *Technical Report*, **14**, Statistics Center, Rutgers University, Piscataway, NJ.
5. Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, (Vol. 1, 3rd ed.), Wiley, New York.
6. Fu, J. C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials, *Statistica Sinica*, **6**, 957–974.
7. Fu, J. C. and Chang, Y. M. (2002). On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials, *Journal of Applied Probability*, **39**, 70–80.
8. Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach, *Journal of the American Statistical Association*, **89**, 1050–1058.
9. Fu, J. C., Wang, L., and Lou, W. Y. W. (2003). On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials, *Journal of Applied Probability*, **40**, 346–360.
10. Hald, A. (1981). *Statistical Theory of Sampling Inspection by Attributes*, Academic Press, London.

11. Han, Q. and Aki, S. (1998). Formulae and recursions for the joint distributions of success runs of several lengths in a two-state Markov chain, *Statistics and Probability Letters*, **40**, 203–214.

12. Hirano, K. (1986). Some properties of the distributions of order k. *Fibonacci Numbers and Their Applications* (eds. A. N. Philippou, G. E. Bergum, and A. F. Horadam), Reidel, Dordrecht, 43–53.

13. Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, **3**, 313–320.

14. Koutras, M. V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach, *Annals of the Institute of Statistical Mathematics*, **47**, 743–766.

15. Koutras, M. V., Bersimis, S., and Maravelakis, P. E. (2007). Statistical process control using Shewhart control charts with supplementary runs rules, *Methodology and Computing in Applied Probability*, **9**, 207–224.

16. Koyama, T. (1978). Modified switching rules for sampling schemes such as MIL-STD-105D, *Technometrics*, **20**, 95–102.

17. Koyama, T., Ohmae, Y., Suga, R., Yamamoto, T., Yokoh, T. and Pabst, W. R. (1970). MIL-STD-105D and the Japanese modified standard, *Journal of Quality Technology*, **2**, 99–108.

18. Lou, W. Y. W. (1996). On runs and longest run tests: A method of finite Markov chain imbedding, *Journal of the American Statistical Association*, **91**, 1595–1601.

19. Montgomery, D. C. (2001). *Introduction to Statistical Quality Control* (4th ed.), John Wiley, New York.

20. Philippou, A. N. and Makri, F. S. (1986). Success runs and longest runs. *Statistics and Probability Letters*, **4**, 211–215.

21. Schilling, E. G. and Sheesley, J. H. (1978a). The performance of MIL-STD-105D under the switching rules, Part 1: Evaluation, *Journal of Quality Technology*, **10**, 76–83.

22. Schilling, E. G. and Sheesley, J. H. (1978b). The performance of MIL-STD-105D under the switching rules, Part 2: Tables, *Journal of Quality Technology*, **10**, 104–124.

23. Shmueli, G. and Cohen, A. (2000). Run-related probability functions applied to sampling inspection, *Technometrics*, **42**, 188–202.

11

Bayesian Network Scan Statistics for Multivariate Pattern Detection

Daniel B. Neill,[1,2] **Gregory F. Cooper,**[3] **Kaustav Das,**[2] **Xia Jiang,**[3] **and Jeff Schneider**[2]

[1] *H.J. Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA*
[2] *School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*
[3] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA*

Abstract: We review three recently proposed scan statistic methods for multivariate pattern detection. Each method models the relationship between multiple observed and hidden variables using a Bayesian network structure, drawing inferences about the underlying pattern type and the affected subset of the data. We first discuss the multivariate Bayesian scan statistic (MBSS) proposed by Neill and Cooper (2008). MBSS is a stream-based event surveillance framework that detects and characterizes events given the aggregate counts for multiple data streams. Next, we describe the agent-based Bayesian scan statistic (ABSS) proposed by Jiang *et al.* (2008). ABSS performs event detection and characterization given individual-level data for each agent in a population. Finally, we review the anomalous group detection (AGD) method proposed by Das, Schneider, and Neill (2008). AGD is a general pattern detection approach which learns a Bayesian network structure from data and detects anomalous groups of records.

Keywords and phrases: Pattern detection, event detection, scan statistic, Bayesian networks, biosurveillance

11.1 Introduction

In this chapter, we focus on the problem of *multivariate event surveillance*, in which we monitor multiple data sources with the goal of identifying patterns that correspond to emerging events. More generally, our goal is *pattern detection*: we wish to find subsets of a large, complex dataset that are relevant, either because the group of data records corresponds to some known statistical pattern which we are interested in detecting, or because it is highly anomalous

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_11,

given our current understanding of the data. Here we review three recently proposed Bayesian variants of the spatial scan statistic [Kulldorff (1997)], which extend the scan statistic methodology to enable rapid detection and accurate characterization of events in multivariate datasets. The three methods include the multivariate Bayesian scan statistic (MBSS) method proposed by Neill and Cooper (2008), the agent-based Bayesian scan statistic (ABSS) method proposed by Jiang *et al.* (2008), and the anomalous group detection (AGD) method proposed by Das, Schneider, and Neill (2008). MBSS is a stream-based event surveillance framework that detects and characterizes events given the aggregate counts for multiple data streams, while ABSS performs event detection and characterization given individual-level data for each agent in a population. Finally, AGD is a general pattern detection approach which detects anomalous groups of records in categorical datasets. These methods use Bayesian networks to model the relationship between multiple observed variables, extending the univariate Bayesian spatial scan statistic methodology of Neill *et al.* (2006) to integrate multiple data streams and differentiate between multiple types of events. MBSS and ABSS assume fixed Bayesian network structures, focusing on stream-based and agent-based event surveillance scenarios, respectively, while AGD learns the Bayesian network structure from data and can be applied to pattern detection in general multivariate datasets.

11.1.1 Event surveillance

Event surveillance systems monitor massive quantities of multivariate data in order to detect and identify emerging patterns. For example, government agencies responsible for public safety must respond rapidly to potential threats including wars, disease outbreaks, crime waves, natural disasters, and terrorist attacks. Timely and informed responses to such events may substantially reduce the resulting costs to society, while delayed or incorrect responses can have catastrophic results. As a concrete example, we consider the task of disease surveillance, in which we monitor electronically available public health data such as hospital visits and medication sales in order to detect emerging outbreaks of disease. Major health threats such as emerging infectious diseases or bioterrorist attacks require rapid and appropriate responses in order to control the spread of disease and treat infected individuals. However, taking appropriate actions often requires knowledge of the characteristics of the disease (e.g. source, method of transmission, and available treatments) and which areas have been affected. Similarly, serious outbreaks requiring urgent responses must be distinguished from less serious outbreaks (e.g. seasonal influenza) and from irrelevant patterns in the data (e.g. increases in medication sales due to store promotions).

The main goals of event surveillance are to achieve *early detection* and *accurate characterization* of events, identifying which events have occurred and

which subsets of the data have been affected by each event. However, the massive size, high dimensionality, and complex spatial and temporal structure of the multivariate data make these goals difficult to achieve. As discussed by Neill and Cooper (2008), an event surveillance system must meet three general criteria to achieve timely and accurate detection.

1. To achieve high detection power, the system must integrate spatial and temporal information from multiple data streams (or from multiple individuals in a population) in a coherent probabilistic framework, incorporating both prior knowledge and historical data into its models.

2. To achieve accurate characterization of events, the system must be able to model and differentiate between multiple types of events.

3. To achieve a rapid response to emerging events, the system must be computationally efficient, detecting patterns in large real-world datasets in near real time.

We now discuss a variety of commonly used methods for event detection, and consider how well the methods fit these criteria.

11.1.2 The spatial scan statistic

The spatial scan statistic [Kulldorff and Nagarwalla (1995), Kulldorff (1997)] is a well-known method for spatial cluster detection. It is in wide use for monitoring health data, detecting clusters of disease cases due to chronic environmental exposures [Kulldorff *et al.* (1997), Hjalmars *et al.* (1996)], infectious disease outbreaks [Mostashari *et al.* (2003)], or bioterrorist attacks [Neill (2006)]. Given a set of spatial locations s_i, each with a count (e.g. number of disease cases) c_i and an underlying population p_i, the spatial scan finds the most significant clusters by searching over a given set of spatial regions, finding those regions which maximize a likelihood ratio statistic, and computing the statistical significance of the detected regions by randomization testing (Figure 11.1). Assuming that the counts in region S are distributed with some unknown rate of incidence q, the goal of the scan statistic is to find regions where the incidence rate is higher than expected. We can either compare the counts inside and outside region S [Kulldorff (1997)], or alternatively, compare the counts inside region S to their expected values obtained from historical data [Neill *et al.* (2005b)]. In either case, we define the null hypothesis H_0, which assumes no clusters, and the alternative hypothesis $H_1(S)$, which assumes a cluster in region S. We then find the region that maximizes the *likelihood ratio statistic*:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}. \tag{11.1}$$

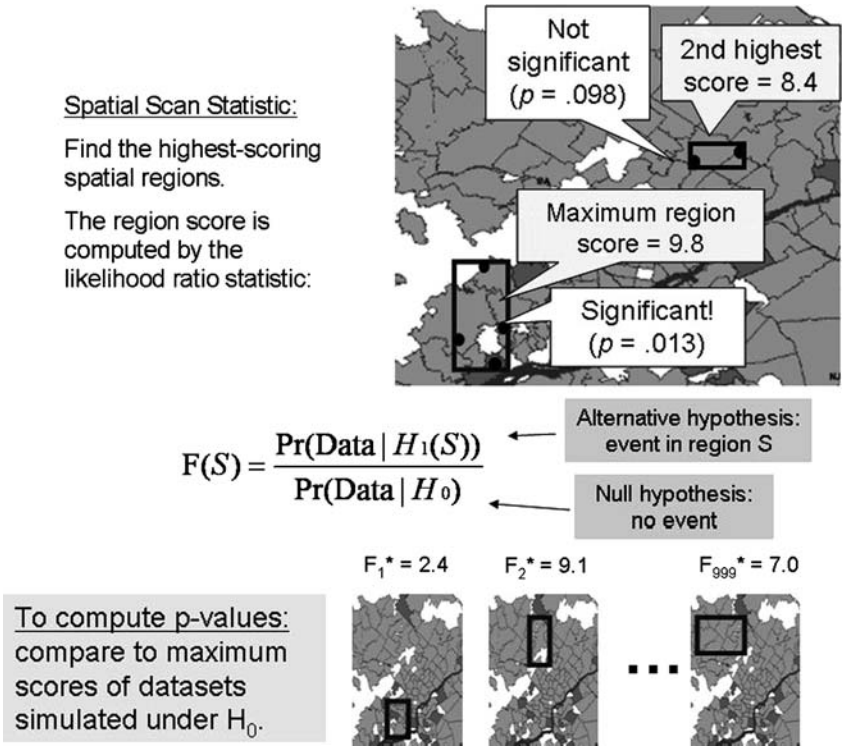


Figure 11.1. Demonstration of the spatial scan statistic.

The original presentation of the spatial scan statistic [Kulldorff (1997)] considers two different models, the Bernoulli model and the Poisson model. In the Bernoulli model, each individual is characterized by some binary variable (e.g. whether the individual goes to the emergency department with a fever). Under the null hypothesis of no clusters, H_0, every individual has a constant probability q_{all} of having this property, while under the alternative hypothesis of a cluster in region S, $H_1(S)$, the incidence rate is higher inside region S than outside (i.e. $q_{in} > q_{out}$). In the Poisson model, we measure the total count of some event type (for example, the number of over-the-counter cough/cold drugs sold) in each spatial region. Assuming that counts are Poisson distributed with mean proportional to the product of the population p_i and the incidence rate q, we can again compare the rates inside and outside region S. Likelihood ratio statistics for each model are derived by Kulldorff (1997).

While Kulldorff's original spatial scan statistic did not take the time dimension into account, later work generalized this method to the "space-time scan statistic" by considering a time series of counts c_i^t for each spatial location s_i and scanning over variable size temporal windows [Kulldorff *et al.* (1998), Kulldorff (2001)]. Recent extensions such as the expectation-based scan statistic [Neill *et al.* (2005b)] and model-based scan statistic [Kleinman *et al.* (2005)]

also take the time dimension into account by using historical data to model the expected distribution of counts in each spatial location.

Many variants of the spatial and space-time scan statistics have been proposed, differing in both the set of regions to be searched and the underlying statistical models. While Kulldorff's original method [Kulldorff (1997)] assumed circular search regions, other methods have searched over rectangles [Neill *et al.* (2005a)], ellipses [Kulldorff *et al.* (2006)], and various sets of irregularly shaped regions [Duczmal and Assuncao (2004), Patil and Taillie (2004), Tango and Takahashi (2005)]. Similarly, many different statistical models have been considered, ranging from simple Poisson and Gaussian statistics [Neill *et al.* (2005b), Neill (2006)] to robust and non-parametric models [Neill and Sabhnani (2007), Neill and Lingwall (2007)].

Kulldorff *et al.* (2007) recently proposed a multivariate variant of the Poisson spatial scan statistic. This work directly extends the original spatial scan to multiple data streams by assuming that all data streams are independent, thus calculating the likelihood ratio score for a given region as the product of the likelihood ratios for each individual data stream. However, we expect streams to be correlated by spatial and temporal trends and other covariates under the null hypothesis, and by the parameters of an event (e.g. outbreak severity) under the alternative hypothesis. Additionally, Kulldorff's method does not characterize events, differentiate between multiple event types, or incorporate prior information. Nevertheless, it can integrate information from multiple data streams for faster and more accurate detection, and it performs well as a "general detector" of anomalous patterns when no prior knowledge of events is assumed. Neill and Cooper (2008) use this method as a baseline for comparison when evaluating the detection power of their MBSS method.

11.1.3 The univariate Bayesian spatial scan statistic

The spatial scan approaches described in Section 11.1.2 fulfill some, but not all, of the criteria for event surveillance discussed above. Spatial scan methods integrate information from multiple spatial locations and multiple time steps, but with the exception of the multivariate Poisson spatial scan [Kulldorff *et al.* (2007)], they can monitor only a single data stream. These methods are also computationally expensive because randomization testing is used to determine the statistical significance of detected clusters, requiring a search over all spatial regions S for many randomly generated datasets. Most importantly, none of these methods can model and differentiate between multiple event types, limiting their usefulness for event characterization.

The Bayesian spatial scan statistic (BSS) method, developed by Neill *et al.* (2006), enables the incorporation of prior information into the event detection process. In the BSS framework, we are given a dataset D, consisting of a time series of counts c_i^t for each spatial location s_i, and we consider a given set of

space-time regions S with prior probabilities $\Pr(H_1(S))$. For some recent past period of time (e.g. the current day), BSS computes the posterior probability that an event has occurred in each spatial region using Bayes's theorem:

$$\Pr(H_1(S) \mid D) = \frac{\Pr(D \mid H_1(S))\Pr(H_1(S))}{\Pr(D)} \tag{11.2}$$

$$\Pr(H_0 \mid D) = \frac{\Pr(D \mid H_0)\Pr(H_0)}{\Pr(D)}. \tag{11.3}$$

The likelihood of the data under each hypothesis is computed using a gamma-Poisson model, and we can specify a probability distribution for the effects of an event on the affected region S. Neill *et al.* (2006) demonstrated that the Bayesian approach has several advantages over frequentist methods. Computation is much faster in the Bayesian framework since randomization testing is unnecessary, and the results of the BSS method (the posterior probability that each region has been affected) are easy to interpret and visualize. Most importantly, the BSS framework allows us to model the spatial and temporal distribution of events by specifying the region priors $\Pr(H_1(S))$, as well as modeling the effects of an event $H_1(S)$ on the monitored data stream in the affected region S. While the original BSS method only considers a single data stream and a single event type, the recently proposed MBSS [Neill and Cooper (2008)] extends this framework to multiple streams and multiple types of events. We discuss the MBSS method in more detail in Section 11.2. More generally, the Bayesian framework can be extended to multivariate data by specification of a Bayesian network relating the observed variables and the underlying event. Each of the three methods discussed in this chapter considers a different set of observations and thus assumes a different Bayesian network structure. In the following section, we briefly review Bayesian networks and their application to pattern detection.

11.1.4 Bayesian networks

A Bayesian network [Pearl (1988), Heckerman *et al.* (1995)], or Bayes net, is a commonly used graphical representation of the joint probability distribution of a set of variables. Bayes Nets are a valuable statistical tool for efficient inference and learning of multivariate probability distributions, and they provide a concise and interpretable visualization of the conditional dependencies between variables. They have been used in many anomaly detection applications, including network intrusion detection [Bronstein *et al.* (2001), Ye and Xu (2000)], detecting malicious e-mails [Dong-Her *et al.* (2004)] and outbreak detection [Wong *et al.* (2003a, 2003b)]. Formally, a Bayesian network can be represented as a directed acyclic graph, where each vertex X_i represents a variable, and each edge from a "parent" vertex X_p to a "child" vertex X_c represents

the dependence of X_c on X_p. The joint probability distribution can be concisely expressed as the product of each variable's conditional distribution given the values of that variable's parents: $\Pr(X_1 \ldots X_M) = \prod_{i=1\ldots M} \Pr(X_i|\text{Parents}(X_i))$. Conditional independencies between variables can also be easily inferred from the network structure: for example, any variable is conditionally independent of its non-descendents given its parents. Inference and learning in Bayesian networks are described in detail by Pearl (1988), Heckerman *et al.* (1995), and many others.

One general approach to anomaly detection using Bayesian networks is to report any individual records with unusually low likelihoods as potential anomalies. In this case, a Bayesian network is learned automatically from a large "training dataset." Established machine learning methods such as "optimal reinsertion" [Moore and Wong (2003)] can be used to efficiently learn the network structure, and the parameters can be optimized by maximum likelihood. We then compute the likelihood of each record in a separate "test dataset" given the Bayes net model, and report the least likely records. Unlike the scan statistic methods considered here, this method treats each individual data record separately, and does not incorporate any spatial or temporal data or other information about group structure. Das *et al.* (2008) use this method as a baseline for comparison in their evaluation of AGD, as discussed below.

Also relevant to our discussion is the PANDA system for disease surveillance proposed by Cooper *et al.* (2004, 2007), which uses Bayesian network models to differentiate between multiple outbreak types (e.g. the CDC Category A diseases), assuming an underlying agent-based model of emergency department visits. Unlike the event detection methods considered here, the baseline version of PANDA-CDCA [Cooper *et al.* (2007)] does not incorporate spatial information, and thus cannot determine which subset of the data has been affected by an event. However, Section 11.3 describes the agent-based Bayesian scan statistic [Jiang *et al.* (2008)], which extends the PANDA-CDCA model to spatial data.

In the remainder of this chapter, we discuss three recently proposed multivariate event detection methods: the MBSS [Neill *et al.* (2007), Neill and Cooper (2008)], the ABSS [Jiang *et al.* (2008)], and the AGD method [Das *et al.* (2008)]. All of these methods incorporate a Bayesian network structure to efficiently model the relationships between variables in the multivariate dataset, using the observed variables to draw inferences about which type of event has occurred and which subset of the data has been affected. The MBSS and ABSS methods each assume a fixed Bayesian network structure relating the underlying event to the observed variables and unobserved state variables, while the AGD method *learns* the Bayesian network structure from data. All three methods can be considered generalizations of the simple Bayesian network anomaly detection method discussed above, detecting self-similar groups of anomalous

records and characterizing the discovered patterns. They also generalize the use of scan statistics to detect clusters of counts, extending spatial scan methods from simple univariate models to multivariate datasets, thus providing a general and powerful framework for event detection. AGD can also be applied to more general pattern detection problems which may not have a spatial or temporal structure, such as knowledge discovery from scientific databases.

11.2 The Multivariate Bayesian Scan Statistic

The multivariate Bayesian scan statistic (MBSS) is a general framework for event detection and characterization using multivariate stream-based data. The MBSS method was first presented by Neill *et al.* (2007) and further developed by Neill and Cooper (2008). This approach extends the original, univariate BSS [Neill *et al.* (2006)] in two ways. First, rather than detecting patterns in a single stream of data, it integrates information from multiple data streams, improving the timeliness and accuracy of event detection. Second, MBSS extends the Bayesian framework to model and distinguish between multiple different types of events, thus enabling both detection and characterization of events.

In the stream-based event detection problem, we are given a dataset D consisting of multiple data streams D_m. Each data stream contains spatial time series data collected at a set of spatial locations s_i. For each stream D_m and location s_i, we have a time series of counts $c_{i,m}^t$, where $t = 0$ represents the current time step and $t = 1, \ldots, T$ represent the counts from 1 to T time steps ago, respectively. In disease surveillance, the data streams may include emergency department (ED) visits, with each stream representing the number of visits with a different chief complaint type, and over-the-counter (OTC) medication sales, with each stream representing the number of sales of a different product group. Thus a given count $c_{i,m}^t$ might represent the number of respiratory ED visits, or the number of cough/cold drugs sold, for zip code s_i on day t.

The goals of the MBSS method are event detection and characterization: to detect any relevant events occurring in the data, identify the type of event, and determine the event duration and affected locations. Thus, MBSS compares the set of alternative hypotheses $H_1(S, E_k)$, each representing the occurrence of some event of type E_k in some space-time region S, against the null hypothesis H_0 that no events have occurred. In disease surveillance, the event types may be either specific illnesses (e.g. influenza, anthrax), non-specific syndromes (e.g. influenza-like illness), or other non-outbreak events that may result in patterns of increased counts, such as promotional sales of OTC medications, inclement weather, or tourism. More generally, an event can be thought of as a process that affects some subset of the count data $c_{i,m}^t$ in some probabilistic

manner. In addition to the set of event types E_k, MBSS is also given the set of space-time regions S to search, where each region S contains some subset of the counts $c_{i,m}^t$. Typically, each search region represents some set of spatial locations s_i for some time duration w, and regions of varying size, shape, and duration are considered.

11.2.1 Methods

Given the set of event types E_k, the set of space-time regions S, and the multivariate dataset D, MBSS computes the posterior probability $\Pr(H_1(S, E_k) \mid D)$ that each event type E_k has affected each space-time region S, as well as the posterior probability $\Pr(H_0 \mid D)$ that no event has occurred. The prior probability of each event type occurring in each space-time region, $\Pr(H_1(S, E_k))$, and the prior probability of no events, $\Pr(H_0)$, are given. MBSS computes the likelihood of the multivariate data given each hypothesis, and then calculates the posterior probability of each hypothesis using Bayes's theorem:

$$\Pr(H_1(S, E_k) \mid D) = \frac{\Pr(D \mid H_1(S, E_k))\Pr(H_1(S, E_k))}{\Pr(D)} \tag{11.4}$$

$$\Pr(H_0 \mid D) = \frac{\Pr(D \mid H_0)\Pr(H_0)}{\Pr(D)}. \tag{11.5}$$

Here the total probability of the data, $\Pr(D)$, is equal to $\Pr(D \mid H_0)\Pr(H_0) + \sum_{S,E_k} \Pr(D \mid H_1(S, E_k))\Pr(H_1(S, E_k))$.

In the MBSS framework, counts are assumed to have been generated from the Bayesian network represented in Figure 11.2. The event type E_k is drawn

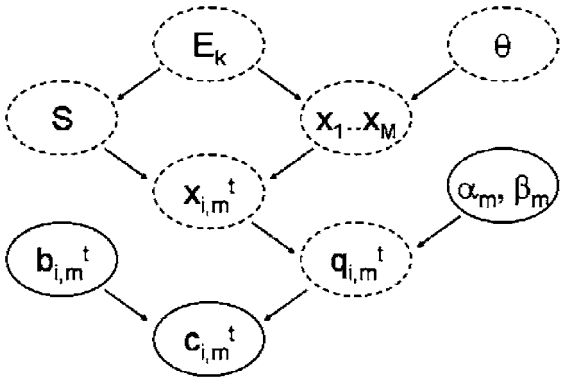


Figure 11.2. Bayesian network representation of the MBSS method. Solid ovals represent observed quantities, and dashed ovals represent hidden quantities that are modeled. The counts $c_{i,m}^t$ are directly observed, while the baselines $b_{i,m}^t$ and the parameter priors for each stream (α_m, β_m) are estimated from historical data.

from a multinomial distribution: here $k = 0$ represents the null hypothesis H_0 of no events, with probability $\Pr(H_0)$, and $k = 1 \ldots K$ represent the occurrence of event type E_k, with corresponding probabilities $\Pr(E_k)$. The region of effect S is conditional on the event type, with probabilities $\Pr(H_1(S, E_k) \mid E_k)$. The distribution of event types and regions can be learned from training data or obtained from expert knowledge.

The effects of an event $H_1(S, E_k)$ on the data are determined by a value $x^t_{i,m}$ for each location s_i, data stream D_m, and time step t. These effects are assumed to be multiplicative, increasing the expected value of each count $c^t_{i,m}$ by a factor of $x^t_{i,m}$. For the null hypothesis H_0, no events have occurred, and $x^t_{i,m} = 1$ everywhere. For an event $H_1(S, E_k)$, only locations and time steps inside the space-time region S have been affected, and thus $x^t_{i,m} = 1$ for all $i, m, t \notin S$. Each event type can have a different joint probability distribution over the effects $x^t_{i,m}$.

The current implementation of MBSS [Neill and Cooper (2008)], as applied to the disease surveillance domain, makes several additional assumptions. To determine the search regions S, spatial locations are mapped to a uniform grid, and all gridded rectangular regions are considered. This method yields computational efficiency and the ability to detect both compact and elongated clusters [Neill *et al.* (2005a)]. MBSS assumes a hierarchical gamma-Poisson model [Clayton and Kaldor (1987), Mollié (1999)]: each count $c^t_{i,m}$ is drawn from a Poisson distribution with mean proportional to the product of the expected count $b^t_{i,m}$ and the relative risk $q^t_{i,m}$. The expected counts (assuming no events taking place) are inferred from historical data, accounting for day-of-week and seasonal trends. Under the null hypothesis, all relative risks $q^t_{i,m}$ for a given data stream D_m are drawn independently from a gamma distribution with parameters (α_m, β_m). These parameters are estimated for each data stream by matching the mean and variance of the gamma-Poisson model to their observed values in historical data. Under the alternative hypothesis $H_1(S, E_k)$, the relative risks $q^t_{i,m}$ inside region S are drawn from a gamma distribution with parameters $(x^t_{i,m}\alpha_m, \beta_m)$. Neill and Cooper (2008) assume a simplified event model, in which an event's effect on each data stream D_m is some constant x_m. These constants are a function of the average effects $x_{km,avg}$ of event type E_k on data stream D_m, as well as the event severity θ: $x_m = 1 + \theta(x_{km,avg} - 1)$. For example, consider an event type E_k with average effects $x_{km,avg} = 1.5$, 1.2, and 1.0 on three data streams $D_1 \ldots D_3$. For an event of "average" severity ($\theta = 1$), the expected counts of streams D_1 and D_2 would be increased by 50% and 20%, respectively, with no effect on stream D_3. For a more severe event with severity $\theta = 2$, the expected counts of streams D_1 and D_2 would be increased by 100% and 40%, respectively. Neill and Cooper (2008) assume a fixed, discrete distribution for θ, and present a simple, smoothed maximum likelihood method for learning the average effects $x_{km,avg}$ from labeled training examples.

The marginal likelihood of each observed count $c^t_{i,m}$ can be computed given the effect $x^t_{i,m}$, baseline $b^t_{i,m}$, and parameter priors α_m and β_m. MBSS integrates over all possible values of the relative risk $q^t_{i,m}$, weighted by their respective probabilities. Neill and Cooper (2008) derive a closed form (negative binomial) solution for the marginal likelihood. Since the null hypothesis assumes $x^t_{i,m} = 1$ everywhere, and since the counts are conditionally independent given the baselines, the α and β parameters, and the effects $x^t_{i,m}$, the marginal likelihood of the data under the null hypothesis can be easily computed. To calculate the likelihood of the data given an alternative hypothesis $H_1(S, E_k)$, MBSS marginalizes over the distribution of effects $x^t_{i,m}$, computing a weighted average of the data likelihoods given each effects vector $(x_1 \ldots x_M)$, weighted by the conditional probability of those effects given $H_1(S, E_k)$. The simplified event model makes these marginals efficiently computable: for each possible event type and severity, MBSS computes log-likelihood ratios for each location, and then computes the log-likelihood ratios for all regions under consideration by summing the location log likelihoods. Alternatively, we can efficiently find those regions with highest log-likelihood ratios, using a variant of the fast spatial scan [Neill and Moore (2004)].

11.2.2 Evaluation

Neill and Cooper (2008) evaluated the event detection and characterization performance of the MBSS method, with and without incorporating prior information, on simulated outbreaks of influenza-like illness (ILI) injected into three streams of OTC medication sales data (cough/cold, anti-fever, and thermometers) from Allegheny County, Pennsylvania. A "general" MBSS detector was used to handle the case when no prior knowledge of events was available. This detector assumed $2^M - 1$ event models (one for each non-empty subset of the M data streams). Each event model assumed equal average effects on the affected subset of streams, and assumed a uniform prior over the event types and affected regions. A "specific" MBSS detector was used to handle the case when prior knowledge of one or more event types was available. This detector assumed a pre-specified event model for each event type, giving the average effects of this event type on each data stream. The main results of their evaluation are as follows.

1. The "general" MBSS detector achieved 1.5 days faster detection than univariate BSS detectors monitoring each data stream separately, demonstrating that MBSS increases detection power by integrating information from the multiple data streams.

2. The "general" MBSS detector and Kulldorff's multivariate spatial scan statistic [Kulldorff *et al.* (2007)] achieve very similar detection performance,

suggesting that either method can be used to detect a broad range of event types when no prior information is available.

3. The "specific" MBSS detector was able to detect outbreaks an average of 1.3 days faster than either the "general" MBSS detector or Kulldorff's multivariate scan. This demonstrates that MBSS can achieve higher detection power by incorporating information about an event's effects on the different data streams. Further performance gains result from using informative region priors that incorporate knowledge of the distribution of each event type in space and time [Neill (2007)].

4. Given an event model for each of two different outbreak types (one primarily causing respiratory symptoms, and one primarily causing fever), MBSS was able to accurately differentiate between the outbreaks by the second outbreak day. The posterior probability of the correct outbreak type increased rapidly over the course of the outbreak, while the probability of the incorrect outbreak type remained constant and small.

11.2.3 Discussion

Neill and Cooper (2008) demonstrate that the MBSS method has several advantages compared to prior event detection approaches. As in the univariate Bayesian spatial scan method [Neill *et al.* (2006)], MBSS can incorporate prior information of an event's effects and its distribution in space and time, increasing detection power. Similarly, the Bayesian scan statistics do not require randomization testing, resulting in 2–3 orders of magnitude faster computation compared to the standard frequentist spatial scan.

Extension of the Bayesian framework to the multivariate case has further, substantial benefits. Integration of information from multiple data streams enables MBSS to detect emerging patterns (e.g. the early stages of an emerging outbreak of disease) that would not be visible from monitoring only a single stream. Incorporating multiple event models not only increases detection power, but also allows MBSS to *characterize* events by specifying models for multiple event types and computing the probability that each type of event has occurred. This enables the user to distinguish relevant events requiring urgent responses from irrelevant events which can safely be ignored, as well as informing the user's response to these events. For example, patterns of ILI would be a high priority for public health officials if these cases were due to pandemic avian influenza or a bioterrorist anthrax attack, and different interventions would be necessary in each case.

Finally, the outputs of MBSS (posterior probabilities of each event type in each space-time region) are easy to interpret, visualize, and use for decision making. For example, considering the posterior probabilities of a given event

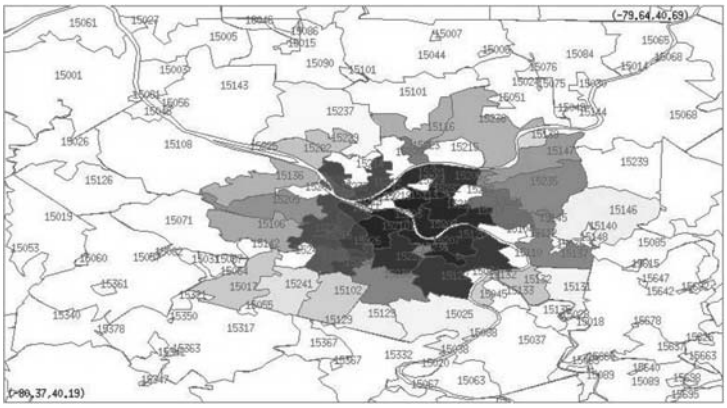


Figure 11.3. Example of a probability map computed by MBSS. Darker shading indicates a higher probability that the given zip code has been affected.

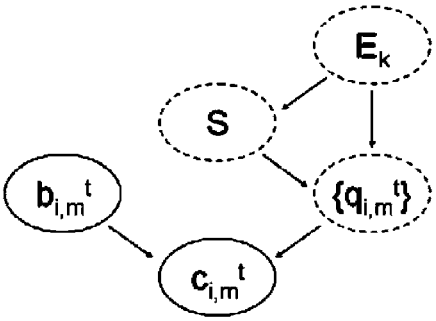


Figure 11.4. General Bayesian network representation of stream-based scan approaches. Relative risks $q^t_{i,m}$ are conditioned on the event type E_k and region S, and may be correlated. Counts $c^t_{i,m}$ are conditionally independent given the relative risks $q^t_{i,m}$ and baselines $b^t_{i,m}$.

type E_k on a given day t, we can compute the probability that each spatial location has been affected by summing the probabilities of all regions containing that location, and display the resulting "probability map" (Figure 11.3).

Comparison to prior methods

The Bayesian network shown in Figure 11.2 is a special case of the general stream-based scan statistic in Figure 11.4. In the general case, the counts $c^t_{i,m}$ are conditionally independent given the baselines $b^t_{i,m}$ and relative risks $q^t_{i,m}$. The joint distribution of the $q^t_{i,m}$ is conditional on the event type E_k and region S. However, the values of $q^t_{i,m}$ (for each location s_i, stream D_m, and time step t)

may be correlated by dependence on other hidden nodes. For example, in Figure 11.2, observing a stream with a high count makes it more likely that the event severity θ is large, and thus increases the probability that another stream has a high count.

Both the univariate BSS [Neill *et al.* (2006)] and Kulldorff's Poisson spatial scan statistic [Kulldorff (1997)] can be considered special cases of the Bayesian network in Figure 11.4, assuming a single data stream D_m and a single event type ($E_k = H_1$ or H_0). In either case, we assume three additional nodes in the Bayesian network (q_{in}, q_{out}, q_{all}). Under the null hypothesis H_0, $q^t_{i,m} = q_{all}$ everywhere, and under the alternative hypothesis $H_1(S)$, $q^t_{i,m} = q_{in}$ inside region S and $q^t_{i,m} = q_{out}$ outside region S. Kulldorff's Poisson scan statistic assumes the maximum likelihood values for q_{in}, q_{out}, and q_{all}. The BSS instead marginalizes over each value, assuming that $q_{in} \sim \text{gamma}(x_{in}\alpha_{in}, \beta_{in})$, $q_{out} \sim \text{gamma}(\alpha_{out}, \beta_{out})$, and $q_{all} \sim \text{gamma}(\alpha_{all}, \beta_{all})$. The values of the α and β parameters are learned from data, and a discrete uniform distribution of x_{in} is assumed.

We note that the MBSS model differs from the original univariate BSS model [Neill *et al.* (2006)] even for the case of a single data stream and single event type. Like Kulldorff's spatial scan statistic [Kulldorff (1997)], the original BSS assumes constant relative risks q_{in}, q_{out}, and q_{all}. The MBSS model allows these risks to vary over space, time, and for different data streams, assuming that each risk is drawn independently from the gamma distribution for that stream. Allowing risks to vary under the null hypothesis reduces the number of false positives due to overdispersion of counts, and the MBSS framework defines a simple and efficiently computable model for the impact of each event type on each data stream.

Incorporating learning into pattern detection

One important aspect of MBSS is the ability to learn new event models (and incrementally update existing models) from user feedback or from labeled training data. Neill and Cooper (2008) demonstrate that the average effects of each event type can be learned from a small number of labeled examples, and that the fitted models gained a large improvement (average of 1.3 days faster detection) compared to the general multivariate detectors. We note that learning from data may only be feasible for very common outbreaks (e.g. influenza), while models of rare events would still rely heavily on expert knowledge. Another possibility would be to learn models of common "confounding" events which are not relevant for detection, and use these models to reduce the false positive rates. For example, patterns of OTC sales of cough/cold medications may occur due to cold weather, poor air quality, short-term population fluctuations due to tourism, or even promotional sales of these medications.

Future work

The incorporation of incremental model learning into the multivariate Bayesian pattern detection framework will be an important aspect of future work. In addition to the effects of each event type on the multiple data streams, many other aspects of the event models can be learned from labeled data, including the prevalence, size, shape, and spread of each type of event. The preliminary results of Neill (2007) suggest that learning these aspects of the event model can also lead to significant improvements in detection performance. Additionally, "active learning" methods can be incorporated in order to choose potential events that are both most relevant to the user and most informative to the system, present these events to the user, and update models based on the user feedback. Finally, the current MBSS implementation assumes the occurrence of a single event, with constant effects over time. Future work will include extending MBSS to "dynamic models" (where events can move and grow over time, and can have spatially and temporally varying effects), as well as "synergistic models" (where multiple events with interacting effects can occur).

11.3 The Agent-Based Bayesian Scan Statistic

Most existing approaches to event detection are "stream-based" methods which monitor the aggregate counts of a set of data streams and report patterns of anomalously high counts. For example, a stream-based disease surveillance system such as MBSS may look at the daily sales of anti-diarrheal medication and numbers of gastrointestinal ED visits, with the goal of detecting an outbreak caused by *Cryptosporidium*. An alternative event detection approach is to model each individual (agent) in a population, observe one or more variables for each individual, and draw inferences about the underlying event. These "agent-based" approaches often rely on an explicit Bayesian network representation to model the causal relationships between the underlying event, the state of each individual (which usually cannot be directly observed), and the observable variables. The PANDA system developed by Cooper *et al.* (2004, 2007) is an agent-based Bayesian network approach for disease surveillance using ED data. Here we consider the agent-based Bayesian scan statistic (ABSS) method proposed by Jiang *et al.* (2008), which extends PANDA by incorporating spatial information.

The ABSS approach assumes a population of R agents, $r = 1, \ldots, R$. Each agent might represent an individual in the population, a measurement device (e.g. a sensor that monitors for the presence of microbes), or some other entity. Each agent r has a set of observable values C_r, which is conditioned on that

agent's underlying state D_r. As in Jiang *et al.* (2008), we assume here that agents are individuals in the population, and that each individual has a single observable value C_r drawn from some multivalued discrete distribution. For example, in the disease surveillance domain, D_r may represent an individual's underlying disease state, which is not directly observed, and C_r may represent that individual's ED visit or purchase of OTC medication. The underlying states, and therefore the observable values, are conditioned on the event type E_k and the affected region S, enabling us to draw inferences about the event and affected region given the set of observed values $\{C_r\}$.

11.3.1 Methods

As in the MBSS approach, the ABSS assumes a fixed set of event types E_k and a fixed set of spatial regions S. Given the multivariate dataset D, the goal of this method is to compute the posterior probability $\Pr(H_1(S, E_k) \mid D)$ that each event type has occurred in each spatial region, as well as the posterior probability $\Pr(H_0 \mid D)$ that no events have occurred. These probabilities can be computed by Bayes's theorem, Equation (11.4), combining the prior probability of each hypothesis with the data likelihood given that hypothesis.

However, the agent-based approach, rather than being given spatial time series data, is given a value C_r for each individual in the population, $r = 1, \ldots, R$. These values are assumed to be drawn from some multivalued discrete distribution, and are conditionally independent of other individuals' values given the individual's underlying state D_r (drawn from a different multivalued discrete distribution). As shown in the Bayesian network representation in Figure 11.5, each individual's state D_r is conditionally independent given the event type E_k, the spatial region of effect S, and the fraction F of the population that has been affected.

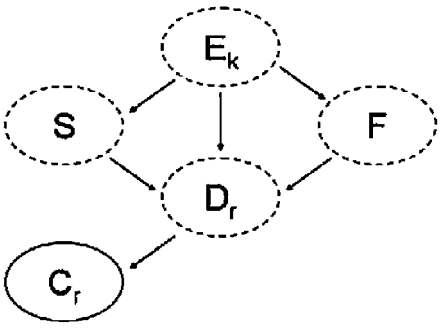


Figure 11.5. Bayesian network representation of the ABSS method. Solid oval represents observed quantities, and dashed ovals represent hidden quantities that are modeled. Each agent's value of C_r is directly observed.

Jiang *et al.* (2008) apply their agent-based approach to the detection of disease outbreaks using ED chief complaint data. The chosen Bayesian network representation is an extension of the Bayesian network used in PANDA-CDCA [Cooper *et al.* (2007)]. PANDA-CDCA does not incorporate spatial information, but ABSS adds an extra node to the Bayesian network representing the spatial region of effect S. In the ABSS framework, the event type E_k is assumed to take on one of 14 values: the 13 different outbreak diseases considered in PANDA-CDCA (influenza, anthrax, etc.) or H_0 (no outbreak occurring). Each individual's underlying state D_r represents two quantities: whether or not the individual goes to the ED, and in the event of an ED visit, what disease is responsible for the visit. Thus D_r can take on 15 different values: the 13 different outbreak types, "other" (i.e. the individual goes to the ED for another reason, such as an accident or broken bone), or "no ED" (i.e. the individual does not visit the ED). The observed values C_r represent the chief complaints for each ED patient (or "no ED" for individuals who did not visit the ED). As in PANDA-CDCA, chief complaints were classified into 54 different categories, and thus each C_r can take on 55 different values including "no ED".

In Jiang *et al.* (2008), the conditional probability table for each node of the Bayesian network in Figure 11.5 is pre-specified based on expert knowledge of the domain. The prior distribution $\Pr(E_k)$ assumes $\Pr(H_0) = 0.95$, $\Pr(\text{influenza}) = 0.04$, and small priors on the 12 other outbreak types (for example, $\Pr(\text{botulism}) = 0.0005$). As in MBSS, the events are assumed to be mutually exclusive, and thus $\Pr(H_0) + \sum_k \Pr(E_k) = 1$. Each event type E_k is assumed to have a uniform region prior, $\Pr(H_1(S, E_k) \mid E_k) = \frac{1}{N_{regions}}$, where $N_{regions}$ is the total number of spatial regions considered. More generally, each event type could have a different spatial prior distribution over regions, and these distributions could be either pre-specified by expert knowledge or learned from labeled training data (e.g. known outbreaks). The variable F is assumed to represent the fraction of the population that is affected by the outbreak and goes to the ED. In the current implementation of ABSS, Jiang *et al.* (2008) assume a fixed, discrete distribution for F. However, different outbreak types might tend to affect different fractions of the population, or be more or less likely to send affected individuals to the ED. The dependence of F on the event type E_k in Figure 11.5 allows this information to be incorporated as well.

The distribution of D_r depends on whether any outbreak is occurring, and if so, on whether individual r is in the affected spatial region S. In the event of no outbreak, or for individuals outside S, D_r is assigned the values "other" or "no ED", where the probability of an individual visiting the ED is estimated using historical data. For individuals inside region S when an outbreak is occurring, D_r is assigned either the outbreak disease (with probability F), "other", or "no ED". Finally, each outbreak disease D_r (including "other") has its own probability distribution over chief complaints C_r, and these distributions were specified by a domain expert.

We now consider how to compute the likelihood of the data for a given event type E_k, affected region S, and fraction F. For the null hypothesis H_0, the same inference can be performed, assuming that $S = \emptyset$. Given the observed value C_r for each individual $r = 1, \ldots, R$, Jiang *et al.* (2008) perform inference on the Bayesian network, marginalizing over the values of the hidden nodes D_r:

$$\Pr(D \mid H_1(S, E_k, F)) = \prod_r \sum_{D_r} \Pr(C_r \mid D_r) \Pr(D_r \mid H_1(S, E_k, F))$$

$$= \prod_{r \in S} \sum_{D_r} \Pr(C_r \mid D_r) \Pr(D_r \mid H_1(E_k, F)) \times \prod_{r \notin S} \sum_{D_r} \Pr(C_r \mid D_r) \Pr(D_r \mid H_0).$$

The total likelihood of the data given each hypothesis can be calculated by marginalizing over the distribution of F, and the posterior probabilities can be computed from the likelihoods and priors using Bayes's theorem as above.

11.3.2 Evaluation

Jiang *et al.* (2008) evaluated ABSS on simulated outbreaks of influenza and illness caused by *Cryptosporidium*, injected into real-world ED data from Allegheny County, Pennsylvania. Detection power (average days to detection, as a function of the allowable false positive rate) and spatial detection accuracy (average overlap between true and detected clusters) were compared to two previously proposed methods, PANDA-CDCA [Cooper *et al.* (2007)] and Kulldorff's original (univariate) spatial scan statistic [Kulldorff (1997)]. Their comparisons demonstrate that ABSS outperformed both PANDA-CDCA and spatial scan by a substantial margin for both datasets and according to both performance measures. The improvement over PANDA-CDCA, which does not use spatial information, demonstrates that incorporation of spatial information into the agent-based Bayesian network framework substantially improves detection power. The improvement over spatial scan, which only uses the aggregate case count in each spatial area rather than the counts for each individual symptom, demonstrates that incorporation of multivariate information (and modeling of the underlying causal structure) also enables improved detection.

11.3.3 Discussion

The ABSS model can be considered a variant of standard scan statistic approaches where data is provided for each individual in the population rather than for a set of data streams. This model is particularly appropriate when we have individual-level data, but can be used for aggregate count data as well. Using individual-level data, if it is available, has several advantages. Though the current ABSS model assumes that each individual r has the same probability distribution for their underlying state D_r and observed variable C_r, the model

can be easily extended to the case where these distributions are conditioned on individual-level covariates such as age, gender, and occupation. Additionally, the agent-based model can be extended to the case where each individual has a joint distribution over multiple observable variables. Observing a single individual with multiple indicators of an event (for example, an ED patient who has both a fever and a rash) may enable faster and more accurate detection than separately considering the number of individuals with each indicator.

On the other hand, if only the aggregate counts are provided, then either the agent-based (multinomial) or the stream-based (multivariate Poisson) method may be more appropriate. For example, we may observe only the number of ED patients with each chief complaint type, or the total sales of each category of OTC medication. If the number of individuals in the population is known, and each individual can take only one action (such as visiting the ED with a specific chief complaint type) out of a predefined set of actions, then the ABSS model may be preferable. If individuals can take multiple actions, and the population size is not known, we might prefer to infer the expected counts from historical data and compare actual to expected counts, as in MBSS.

Comparison to prior methods

The Bayesian network shown in Figure 11.5 is a special case of the general agent-based scan statistic in Figure 11.6. In the general case, each individual r has an observed value C_r. The joint distribution of the C_r is conditional on the event type E_k and region S. However, different individuals' values of C_r may be correlated by the addition of hidden nodes to the Bayesian network. For example, in Figure 11.5, observing an individual with disease symptoms increases the likelihood that F (the fraction of the population affected) is large, and thus increases the probability that another individual has disease symptoms.

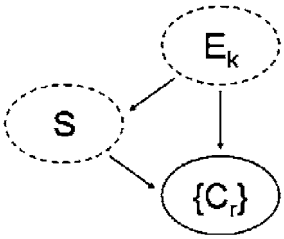


Figure 11.6. General Bayesian network representation of agent-based scan approaches. Solid oval represents observed quantities, and dashed ovals represent hidden quantities that are modeled. Each agent's value of C_r is conditioned on the event type E_k and region S, and these values may be correlated by additional hidden nodes.

The Bernoulli spatial scan statistic [Kulldorff (1997)] can also be considered a special case of the Bayesian network in Figure 11.6, with one event type ($E_k = H_1$ or H_0) and binary variables C_r. In this case, we assume three additional nodes in the Bayesian network (q_{in}, q_{out}, q_{all}). Under the null hypothesis H_0, $\Pr(C_r = 1) = q_{all}$ everywhere, and under the alternative hypothesis $H_1(S)$, $\Pr(C_r = 1) = q_{in}$ inside region S and $\Pr(C_r = 1) = q_{out}$ outside region S. However, rather than marginalizing over q_{in}, q_{out}, and q_{all}, the Bernoulli spatial scan assumes the maximum likelihood values for each node.

Future work

Future work by Jiang *et al.* will compare the agent-based approach to other multivariate spatial detection methods, including MBSS and Kulldorff's multivariate spatial scan statistic. Additionally, the current implementations of ABSS and PANDA-CDCA used a "specific" detector with pre-specified models of 13 outbreak diseases (including influenza, *Cryptosporidium*-caused illness, and the CDC Category A diseases), and the simulated outbreaks were generated assuming a distribution of chief complaints that is identical to these models. Future work will evaluate ABSS on disease outbreaks generated according to different chief complaint distributions (i.e. measuring performance as a function of the difference between true and assumed distributions), and thus test the robustness of this method to model misspecification. While the current implementation of ABSS is specific to ED disease surveillance, ABSS can be extended to other application domains using more general definitions of the individual's underlying state D_r, observed behavior C_r, and the fraction of the population affected F. Finally, future versions of ABSS will include many of the current features of MBSS, such as incorporation of temporal information, visualization of outputs, and learning of event models from labeled data.

11.4 The Anomalous Group Detection Method

We now consider how the scan statistic framework can be extended from the specific case of event surveillance to more general multivariate datasets. This extension poses several challenges. Since many datasets have no explicit space or time component, we cannot simply search over geographical regions, and thus it is not clear which subsets of the data should be considered. Additionally, while other scan statistic methods assume a fixed parametric model for the effects of different types of patterns on the data, we may wish to detect anomalous patterns in more general datasets where no such model is known. One solution to these challenges is provided by the anomalous group detection (AGD) method, recently proposed by Das *et al.* (2008). Rather than relying on a fixed parametric

model, AGD *learns* the structure and parameters of a Bayesian network from the data, and searches over self-similar subsets of the data to find anomalous patterns.

The AGD method can be used to detect anomalous groups in arbitrary, non-spatial datasets with discrete-valued attributes. For typical stream-based scan statistic approaches, each data point consists of a set of real-valued "location" attributes as well as real-valued "count" data. The set of search regions is defined by the location attributes (e.g. spatial scan searches over geographically contiguous subsets of the data) while the likelihood under each hypothesis $H_1(S)$ is a function of the counts inside and outside region S. In the more general pattern detection problem, there may be no defined set of location attributes, and thus we can no longer predefine a set of search regions based on geographical attributes such as size, shape, or contiguity. Nevertheless, we want to formulate a measure of how well a subset of data points fits as a *group* based on the similarity between the data points. We must then perform a search over all possible subsets of the data in order to find the most anomalous groups.

Another difference between the AGD method and other scan statistic approaches is in the definition of anomalousness for a data point or a group of points. Scan statistics are usually applied to detect overdensities of records in a given space: individual records are aggregated into counts, and clusters with anomalously high counts are detected. In the AGD framework, however, each record has many discrete-valued attributes, and can have an inherent degree of anomalousness depending on its features. Most records are generated from the "normal" distribution of data and hence are not relevant. Instead, the goal of AGD is to detect groups of records that are both anomalous and also self-similar in some respect.

11.4.1 Methods

The AGD framework assumes a multivariate dataset D, where each data record $R_i \in D$ has values for a set of discrete-valued attributes $X_1 \ldots X_M$. As in the original spatial scan statistic approach [Kulldorff (1997)], AGD finds the set of records that maximizes the likelihood ratio statistic $F(S) = \frac{\Pr(D \mid H_1(S))}{\Pr(D \mid H_0)}$, where H_0 is the null hypothesis that there are no anomalies present, and $H_1(S)$ is the alternative hypotheses specifying that the set S is an anomalous group. AGD assumes Bayesian network models for both the null and alternative hypotheses, and computes the data likelihoods given these models. For the null hypothesis H_0, a Bayesian network model is inferred from a separate training dataset (e.g. historical data), which is assumed to contain no anomalies, and all data records are assumed to have been drawn independently from this model. Under the alternative hypothesis $H_1(S)$, the records contained in subset S are assumed to have been drawn from a different Bayes net model, while the rest of the

data records are generated from the null model. The Bayesian network model parameters for the alternative hypothesis $H_1(S)$ are learned directly from the records in subset S, as discussed below.

This scoring metric gives a higher score to anomalous records, as well as setting a constraint of similarity between the records in a group. If the records in S are similar to each other, then $H_1(S)$ will be able to model them tightly. This will result in a high value of the data likelihood under the alternative hypotheses $H_1(S)$, thus increasing the score $F(S)$. Also, records that are poorly modeled by the training data will have low likelihoods under the null hypothesis H_0, again increasing the group score $F(S)$. Hence, maximizing this score leads to grouping of similar records and at the same time it prefers records that are anomalous (i.e. that have low likelihoods under the null hypothesis).

As discussed by Das *et al.* (2008), the AGD algorithm consists of three steps:

1. Learn the Bayesian network model for the null hypothesis H_0 from the training data.
2. For all subsets of the data S:
   (a) Fit the alternate hypothesis Bayesian network ($H_1(S)$) parameters using data from subset S.
   (b) Compute the group likelihood ratio score $F(S)$.
3. Output the groups with highest score.

Step 1 is to learn the Bayesian network corresponding to the null hypothesis. The network structure is learned automatically from the training dataset using the optimal reinsertion algorithm [Moore and Wong (2003)], and this structure is assumed for the null hypothesis H_0 and for all alternative hypotheses $H_1(S)$. The probability table parameters of H_0 are then learned from the training dataset using smoothed maximum likelihood estimation. For a given node corresponding to the variable X_i in the Bayes net, let X_{Π_i} denote the set of variables corresponding to the parent nodes of X_i. The conditional probability table of X_i has parameters corresponding to the conditional probability values $\theta_{ijk} = \Pr(X_i = j \mid X_{\Pi_i} = k)$. Here we must estimate θ_{ijk} for each variable X_i, value j, and set of parent values k. The maximum likelihood parameter estimates are given by $\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_{j'} N_{ij'k}}$, where N_{ijk} denotes the number of instances in the training dataset with $X_i = j$ and $X_{\Pi_i} = k$. To deal with sparsity of the training data, Das *et al.* (2008) apply Laplace smoothing to adjust the estimate of each model parameter.

Steps 2–3 find groups of records S that maximize the likelihood ratio score $F(S) = \frac{\Pr(D \mid H_1(S))}{\Pr(D \mid H_0)}$, where the alternative hypothesis $H_1(S)$ assumes that the records in subset S form an anomalous group, and the null hypothesis H_0 assumes that no anomalous groups are present. The optimal group can be found by

searching over all subsets of the test data, but this exhaustive search would require exponential time. Thus Das *et al.* (2008) propose a greedy heuristic search method which starts from each record as an initial seed and iteratively adds the record that most improves the likelihood ratio score. This search method can find high-scoring groups in a computationally efficient manner, but does not guarantee that the optimal group will be found.

Step 2a fits the parameters of the Bayesian network for the alternative hypothesis $H_1(S)$. Das *et al.* (2008) use an empirical Bayes approach in which these parameters are estimated from the counts in the subset of the test dataset represented by S, following an approach of smoothed maximum likelihood estimation similar to Step 1 above. In this case, N_{ijk} denotes the corresponding counts in region S. Since the number of records in group S may be small and this data is used to fit a large number of Bayesian network parameters, data sparsity is a serious problem, and computing the likelihood $\Pr(D \mid H_1(S))$ using this model risks overfitting of the data.

Step 2b computes the group likelihood ratio score $F(S)$, performing inference on the Bayesian networks corresponding to $H_1(S)$ and H_0 to compute the data likelihoods under each hypothesis. Since data points are assumed to be conditionally independent given the model, and records not contained in subset S have identical likelihoods given $H_1(S)$ and H_0, the likelihood ratio statistic simplifies to

$$F(S) = \frac{\prod_{R_i \in S} \Pr(R_i \mid H_1(S))}{\prod_{R_i \in S} \Pr(R_i \mid H_0)}. \tag{11.6}$$

Das *et al.* (2008) deal with the overfitting problem mentioned above by using a "leave-one-out" method based on the pseudo-likelihood of each record R_i in S. In this case, the numerator of Equation (11.6) becomes $\prod_{R_i \in S} \Pr(R_i \mid H_1(S - \{R_i\}))$. To compute the likelihood of each record R_i, assuming the alternative hypothesis $H_1(S)$, a Bayesian network model is learned from all the records in S except for R_i, and this model is used to compute the likelihood of R_i. Since the likelihood of each record is computed without using that record to estimate the model parameters, this reduces the risk of overfitting.

Step 3 outputs the highest scoring groups found in step 2. Additionally, Das *et al.* (2008) compute an anomalousness score for each individual record R in the test data by finding the highest scoring group $S^*(R)$ that contains R. The score of record R can then be computed in one of two ways. In the "group likelihood ratio" approach, $Score(R)$ is set equal to the group score $F(S^*(R))$. This approach gives a high score to any record that is contained in a highly anomalous group, regardless of whether the record is itself anomalous or just similar to other anomalous records. Alternatively, we can consider only the contribution of record R to the score of $S^*(R)$. In this "single record likelihood ratio" approach, $Score(R)$ is set equal to the partial record pseudo-likelihood ratio, $\frac{\Pr(R \mid H_1(S^*(R) - \{R\}))}{\Pr(R \mid H_0)}$.

11.4.2 Evaluation

Das *et al.* (2008) compare the performance of their method to the baseline method described above, which detects individual records with low likelihoods given the null Bayes net model. Synthetic anomalies were injected into two real-world datasets: a dataset of ED records from Allegheny County, Pennsylvania, and the PIERS dataset of container shipping data. The former dataset contains records of patients visiting Allegheny County EDs. Each record consists of six categorical attributes (hospital ID, prodrome, age decile, home zip code, and chief complaint class), and the goal is to detect anomalous groups of records (e.g. spatial disease clusters, age/gender clusters, and increases in different symptom types) that correspond to emerging disease outbreaks. The second dataset consists of records describing containers imported into the country. Each record consists of 10 attributes: country of origin, departing and arriving ports, shipping line, shipper name, vessel name, commodity being shipped, and the size, weight, and value of the container. In this case, the goal is to detect anomalous groups of records corresponding to patterns of smuggling, terrorist activity, or other illicit shipments.

Das *et al.* (2008) evaluated the performance of the algorithms in two different ways. The first evaluation criterion was the ability of each algorithm to identify each individual anomaly correctly. Figure 11.7 plots the detection precision, i.e. the ratio of number of true positives to the total number of predicted positives, against the detection rate, i.e. the proportion of total true anomalies that were detected. For both the "group likelihood ratio" and "single record likelihood ratio" methods, AGD performed significantly better than the baseline method without grouping. Similar results were obtained when examining the ability of the algorithms to identify and distinguish between entire datasets which have anomalous groups against ones which do not have any anomalies, e.g. distinguishing datasets containing outbreaks from datasets with no outbreaks. For these experiments, the grouping method again achieved significantly higher performance than the baseline anomaly detection method. While

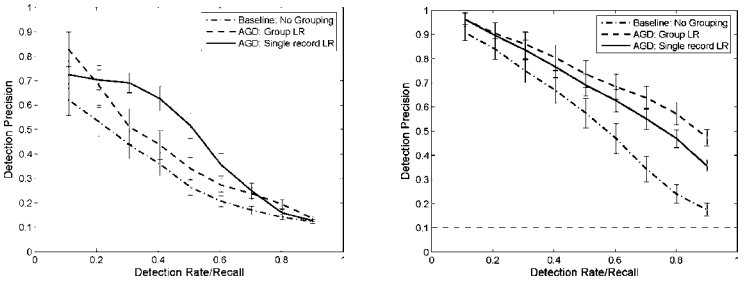


Figure 11.7. Plot of detection precision vs. recall for (left) ED dataset and (right) PIERS dataset, from Das *et al.* (2008).

the set of anomalies was synthetically generated, current work by Das *et al.* includes evaluation on real anomalies, e.g. retrospective analysis of known disease outbreaks.

11.4.3 Discussion

The primary advantage of the AGD method is its generality: unlike the MBSS and ABSS methods, AGD can be directly applied to arbitrary multivariate datasets without the need for a pre-specified Bayesian network model of how the data is generated. Instead, the structure of the network and the parameters for each node are learned from a training dataset, and the learned model is used for detection. Although Das *et al.* (2008) exclusively deal with categorical valued datasets, AGD can be generalized to handle datasets containing real-valued attributes as well, using Bayesian network models that incorporate both categorical and real-valued nodes. However, AGD does have several disadvantages. It cannot model and distinguish between multiple event types, since the parameters for the alternative hypothesis $H_1(S)$ are fitted directly from that subset of the test data. Learning a model using the test data and then computing the likelihood of the test data given that model can result in overfitting, and the proposed solution (use of the pseudo-likelihood) gives outputs that cannot be interpreted as posterior probabilities.

Comparison to prior methods

The AGD algorithm can be thought of as a generalization of scan statistic methods such as MBSS and ABSS to arbitrary multivariate datasets without predefined location or count attributes. All attributes of the data are used to determine both the self-similarity of the group and the anomalousness of its component records, as opposed to previous methods that determine the anomalousness of the count attributes and use the location attributes for grouping. While standard scan statistics implicitly or explicitly assume a fixed Bayesian network model relating the observed variables (i.e. aggregate counts in stream-based approaches, and individual-level variables in agent-based approaches) to the underlying event and affected region, AGD *learns* the underlying model from the training dataset. Additionally, standard scan statistics are geared toward the event detection problem, searching over a set of contiguous spatial regions that are predefined based on the location attributes of the data, while AGD performs a heuristic search over arbitrary subsets of the data.

Future work

Future work by Das *et al.* will extend the AGD approach to incorporate multiple pattern types E_k, model the effects of each pattern type on the data, and

distinguish between multiple pattern types (by computing the posterior probability that each pattern type E_k affects each subset of the data S). Each pattern type can have a different prior probability $\Pr(E_k)$ and a different distribution over subsets of the data. Models of how each pattern type will affect a given subset of the data can be defined, allowing computation of the data likelihood given each hypothesis $H_1(S, E_k)$. Different pattern types can have a different distribution over Bayesian network structures and parameters, and the data can be represented as a "mixture of Bayes nets." Each alternative Bayes net model can be related to the null Bayes net by changing the conditional distributions of the output attributes based on the event model E_k. Finally, future work will develop methods which can *learn* these models for each pattern type. These extensions could be valuable for finding groups in new datasets that match specific patterns of anomalous activity learned from earlier data.

Acknowledgments

This work was partially supported by NSF grant IIS-0325581 and CDC grant 8 R01 HK000020 02. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NSF or CDC.

References

1. Bronstein, A., Das, J., Duro, M., Friedrich, R., Kleyner, G., Mueller, M., Singhal, S., and Cohen, I. (2001). Bayesian networks for detecting anomalies in internet-based services, In *Intl. Symposium on Integrated Network Mgmt.*

2. Clayton, D., and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671–681.

3. Cooper, G. F., Dash, D. H., Levander, J. D., Wong, W.-K., Hogan, W. R., and Wagner, M. M. (2004). Bayesian biosurveillance of disease outbreaks, In *Proc. Conference on Uncertainty in Artificial Intelligence.*

4. Cooper, G. F., Dowling, J. N., Levander, J. D., and Sutovsky, P. (2007). A Bayesian algorithm for detecting CDC Category A outbreak diseases from emergency department chief complaints, *Advances in Disease Surveillance*, **2**, 45.

5. Das, K., Schneider, J., and Neill, D. B. (2008). Detecting anomalous groups in categorical datasets, submitted for publication, Carnegie Mellon University, School of Computer Science.

6. Dong-Her, S., Hsiu-Sen, C., Chun-Yuan, C., and Lin, B. (2004). Internet security: malicious e-mails detection and protection, *Industrial Mgmt. and Data Sys.*, **104**, 613–623.

7. Duczmal, L., and Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrary shaped spatial clusters, *Computational Statistics and Data Analysis*, **45**, 269–286.

8. Heckerman, D., Geiger, D., and Chickering, M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data, *Machine Learning*, **20**, 197–243.

9. Hjalmars, U., Kulldorff, M., Gustafsson, G., and Nagarwalla, N. (1996). Childhood leukemia in Sweden: using GIS and a spatial scan statistic for cluster detection, *Statistics in Medicine*, **15**, 707–715.

10. Jiang, X., Neill, D. B., and Cooper, G. F. (2008). A Bayesian network model for spatial event surveillance, *International Journal of Approximate Reasoning*, in press.

11. Kleinman, K., Abrams, A., Kulldorff, M., and Platt, R. (2005). A model-adjusted space-time scan statistic with an application to syndromic surveillance, *Epidemiology and Infection*, **133**(3), 409–419.

12. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**(6), 1481–1496.

13. Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society A*, **164**, 61–72.

14. Kulldorff, M., Athas, W., Feuer, E., Miller, B., and Key, C. (1998). Evaluating cluster alarms: a space-time scan statistic and cluster alarms in Los Alamos, *American Journal of Public Health*, **88**, 1377–1380.

15. Kulldorff, M., Feuer, E. J., Miller, B. A., and Freedman, L. S. (1997). Breast cancer clusters in the northeast United States: a geographic analysis, *American Journal of Epidemiology*, **146**(2), 161–170.

16. Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic, *Statistics in Medicine*, **25**, 3929–3943.

17. Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K., and Platt, R. (2007). Multivariate scan statistics for disease surveillance, *Statistics in Medicine*, **26**, 1824–1833.

18. Kulldorff, M., and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine*, **14**, 799–810.

19. Mollié, A. (1999). Bayesian and empirical Bayes approaches to disease mapping, In Lawson, A. B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., and Bertollini, R., *Disease Mapping and Risk Assessment for Public Health*, Wiley, New York.

20. Moore, A., and Wong, W.-K. (2003). Optimal reinsertion: a new search operator for accelerated and more accurate Bayesian network structure learning, In *Proceedings of the 20th Intl. Conf. on Machine Learning*, 552–559.

21. Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R., and Kulasekera, V. (2003). Dead bird clustering: a potential early warning system for West Nile virus activity, *Emerging Infectious Diseases*, **9**, 641–646.

22. Neill, D. B. (2006). Detection of spatial and spatio-temporal clusters, CMU-CS-06-142, Ph.D. thesis, Carnegie Mellon University, School of Computer Science.

23. Neill, D. B. (2007). Incorporating learning into disease surveillance systems, *Advances in Disease Surveillance*, **4**, 107.

24. Neill, D. B., and Cooper, G. F. (2008). A multivariate Bayesian scan statistic for early event detection and characterization, *Machine Learning*, in press.

25. Neill, D. B., and Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance, *Advances in Disease Surveillance*, **4**, 106.

26. Neill, D. B., and Moore, A. W. (2004). Rapid detection of significant spatial clusters, In *Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 256–265.

27. Neill, D. B., Moore, A. W., and Cooper, G. F. (2006). A Bayesian spatial scan statistic, In *Advances in Neural Information Processing Systems 18*, 1003–1010.

28. Neill, D. B., Moore, A. W., and Cooper, G. F. (2007). A multivariate Bayesian scan statistic, *Advances in Disease Surveillance*, **2**, 60.

29. Neill, D. B., Moore, A. W., and Sabhnani, M. R. (2005a). Detecting elongated disease clusters, *Morbidity and Mortality Weekly Report*, **54 (Supplement on Syndromic Surveillance)**, 197.

30. Neill, D. B., Moore, A. W., Sabhnani, M. R., and Daniel, K. (2005b). Detection of emerging space-time clusters, In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*.

31. Neill, D. B., and Sabhnani, M. R. (2007). A robust expectation-based spatial scan statistic, *Advances in Disease Surveillance*, **2**, 61.

32. Patil, G. P., and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Envir. Ecol. Stat.*, **11**, 183–197.

33. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA.

34. Tango, T., and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11.

35. Wong, W.-K., Moore, A. W., Cooper, G. F., and Wagner, M. M. (2003a). Bayesian network anomaly pattern detection for disease outbreaks, In *Proc. 20th International Conference on Machine Learning.*

36. Wong, W.-K., Moore, A. W., Cooper, G. F., and Wagner, M. M. (2003b). WSARE: What's strange about recent events? *Journal of Urban Health*, **80**(2 Suppl. 1), i66–i75.

37. Ye, N., and Xu, M. (2000). Probabilistic networks with undirected links for anomaly detection, In *IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, 175–179.

12

ULS Scan Statistic for Hotspot Detection with Continuous Gamma Response

Ganapati P. Patil,[1] **Sharadchandra W. Joshi,**[2] **Wayne L. Myers,**[3] **and Rajesh E. Koli**[4]

[1] *Center for Statistical Ecology and Environmental Statistics, Department of Statistics, The Pennsylvania State University, University Park, PA, USA*
[2] *Department of Computer Science, Slippery Rock University of Pennsylvania, Slippery Rock, PA, USA*
[3] *School of Forest Resources, The Pennsylvania State University, University Park, PA, USA*
[4] *Watershed Surveillance and Research Institute, JalaSRI, M.J. College, Jalgaon, India*

Abstract: An approach using the upper level set (ULS) scan statistic to detect geospatial hotspots along with its software implementation is presented for continuous response. The ULS scan statistic is based on the ULS scan tree. A ULS scan tree is a data structure constructed from response data over a geographic region partitioned into cells. Candidates for hotspots are zones in the region. Each such candidate zone consists of cells that are connected geographically. A ULS scan tree is used to identify candidate zones systematically. Nodes of the ULS scan tree are connected zones. The root (the bottom level) of the ULS scan tree is a zone consisting of the entire region. Zones at the top level (leaf zones) consist of cells with maximal response values. For in-between levels, zones at a given level consist of connected cells with higher response values than zones at a lower level. A suitable likelihood statistic and Monte Carlo analysis are used to determine the significance of zonal nodes as hotspots. The gamma response model is studied in detail. A case study illustrating application of the gamma response model is presented.

Keywords and phrases: Upper level set scan statistic, ULS tree, hotspot detection, continuous response model

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_12,

12.1 Introduction

The one-dimensional scan statistic has been exhaustively covered in two books [Glaz and Balakrishnan (1999), Glaz *et al.*, (2001)]. A wide variety of methods has been proposed for modeling and analyzing geospatial data [Cressie (1991)]. More recently, the spatial scan statistic proposed by Kulldorff and Nagarwala (1995) and Kulldorff (1997) has provided a popular tool in the form of the SatScan software system developed by Kulldorff *et al.* (1998) for detection and evaluation of disease clusters for discrete response data. It is available on the web free of charge. A commercial software system [Biomedware (2001)] is also available. With suitable modifications, the scan statistic approach can be used for critical area analysis in fields other than the health sciences, and also for continuous response data.

Basic components of the scan statistic are the topological structure under investigation, the probability distribution used to model responses and the shapes and sizes of the scanning window. In this paper, we present an approach to the scan statistic: the upper level set (ULS) tree scan statistic, as well as its software implementation, with characteristics that are different from a typical spatial scan statistic software in the following ways.

- The ULS scan statistic uses an irregularly shaped scanning window, unlike most other scan statistics, which are based on some regularly (circularly or elliptically) shaped windows.
- Applicability of the ULS scan statistic is not limited to geospatial regions. It can be conveniently used to detect hotspots in any structure with the network topology.
- The software provides an option of the use of the gamma distribution to model response data that are of a continuous nature in addition to the binomial and the Poisson models.

In Sections 12.2, 12.3 and 12.4 we introduce basic ideas behind the ULS scan statistic based on Patil and Taillie (2003, 2004). In Section 12.5 we discuss some computational aspects. The gamma response model is presented in Section 12.6. Section 12.7 contains a fairly detailed account of software implementation of the ULS scan statistic. We conclude with an environmental application of the software using the gamma response model.

12.2 Basic Ideas

We consider the following scenario: A geospatial region R is partitioned or tessellated into N cells. Response data on $y_1, y_2, \ldots, y_N$ are available for the N cells, y_a being the response for cell a. $y_1, y_2, \ldots, y_N$ are regarded as observed values of independently distributed response variates $Y_1, Y_2, \ldots, Y_N$. Also known is the "size" A_a of cell $a, a = 1, 2, \ldots, N$. Interpretation of size depends on the context in which the data are collected. Thus, in a situation where response data are counts of incidences of a certain disease in R, A_a is the size of the exposed population of cell a. If y_a is arable acreage, then A_a can be the geographic area of the cell. Of essential interest are the response rates or response intensities, $G_a = y_a/A_a, a = 1, 2, \ldots, N$.

The spatial scan statistic seeks to identify "hotspots," which are clusters of cells in R that have elevated response rates compared with the rest of the region. A cluster of cells in R must satisfy two properties before it can be considered as a hotspot candidate:

1. The cluster must be geographically connected. Such a cluster will be referred to as a zone. The set of all zones is denoted by Ω.

2. The zone should not be excessively large; otherwise, the zone rather than its exterior would constitute background. Generally, we limit the search for hotspots to zones that do not comprise more than, say, fifty percent of the region.

To detect a hotspot, the circle-based scan statistic due to Kulldorff adopts a hypothesis testing model. In order to illustrate the concept, let us consider the case when each $Y_a \sim \text{Binomial}(n_a, p_a)$ where $0 < p_a < 1$ is an unknown parameter and n_a is the cell size. With this, the following is a statement of the null and the alternative hypotheses:

$$
\begin{aligned}
H_0 &: p_a \text{ is the same for all cells } a \text{ in } R \\
H1 &: \text{there is a non-empty zone } Z \in \Omega \text{ and parameter values} \\
&\quad 0 < p_0,\ p_1 < 1 \text{ such that} \\
&\quad p_a = p_1 \text{ for all cells } a \text{ in } Z \\
&\quad p_a = p_0 \text{ for all cells } a \text{ in } R - Z \text{ and} \\
&\quad p_1 > p_0
\end{aligned}
$$

H_0 asserts that there is no hotspot. Z occurs in H_1 as an unknown parameter so that the full model $H_0 \cup H_1$ involves three parameters, Z, p_0, and p_1.

Under H_1 we need to compute the likelihood $L(Z, p_0, p_1)$ maximized over $Z \in \Omega$, and $0 < p_0, p_1 < 1$. For a given Z, the profile likelihood

$$L(Z) = \max\{L(Z, p_0, p_1) : 0 < p_0, p_1 < 1\}$$

is readily determined with maximum likelihood estimations (MLEs) of p_0 and p_1. The difficult part is to maximize $L(Z)$ over $Z \in \Omega$ since usually Ω is extremely large, making exhaustive, search for the maximum impractical. One common approach to obtain at least an approximately optimal solution is to use reduced parameter space, that is, to maximize $L(Z)$ over a suitable subset Ω_0 of Ω. The success of this approach depends on whether Ω_0 contains the MLE of Z over Ω or at least a satisfactorily close approximation to it. The traditional circular scan statistic uses expanding circles with centers in each cell to determine Ω_0. This strategy tends to produce compact candidate zones and may do a poor job of approximating actual clusters of arbitrary shapes. The reduced parameter space is determined by the geometry of tessellation without involving the response data.

The ULS scan statistic described below and implemented as a software package described later also uses the approach of parameter space reduction. Its central idea lies in the concept of upper level sets. This approach takes an adaptive view so that the resulting reduced parameter space, Ω_{ULS}, depends on data.

12.3 ULS Scan Statistic

The ULS approach views the response data as a surface in three dimensions. With the region R in the xy-plane, the surface is constructed by erecting a solid cylinder along the z-axis over each cell. The height of the cylinder over cell a is proportional to the response rate of the cell.

To begin, we construct zones at different levels. A zone at level g is a connected component of the upper level set

$$U_g = \{a \in R : G_a \geq g\},$$

where $g \in G = \{G_a : a \in R\}$.

The reduced set of candidate zones, Ω_{ULS}, is the collection of all connected components of all upper level sets. Graphically, the upper level set at level g is the projection on R of the cross section of the response surface with the horizontal plane $z = g$.

Ω_{ULS} can also be thought of as a data structure in the form of a tree. All members of Ω_{ULS} are nodes of the ULS tree. To further describe the tree

structure, let us assume the set G has m elements: $g_1 > g_2 > \ldots > g_m$ and define the sets

$$T_i = \{a \in R : G_a = g_i\}, i = 1, 2, \ldots, m.$$

Also, for brevity, denote the set U_g by U_i when $g = g_i$. Then

$$U_i = T_1 \cup T_2 \cup \ldots \cup T_i, i = 1, 2, \ldots, m.$$

With this notation, connected components of U_i are level i nodes. The root of the ULS tree is $U_m = R$, the lowest level node. Connected components of U_1, the highest level nodes, are leaf nodes. Given $T_i, 1 < i < m$, consider a fixed connected component C of T_i. If C has no cell adjacent to any of the higher level nodes, then C is also a leaf node. (Such a zone is a local peak of the response surface.) On the other hand, if C has cells that are adjacent to higher level nodes, say $Z_1, Z_2, \ldots, Z_k$, then we have a connected component $C \cup Z_1 \cup Z_2 \cup \ldots \cup Z_k$ of U_i as a level i node and this node is the parent node of $Z_1, Z_2, \ldots, Z_k$. Figures 12.1, 12.2, and 12.3 illustrate the ULS tree building process.

As implied in the discussion above, it is convenient for our purpose to orient the ULS tree with the leaf nodes at the top and the root node at the bottom. As we trace the ULS tree from the top node towards the root, each cell in R makes its entry in the tree in a uniquely determined node. This implies that the cardinality of Ω_{ULS} is less than or equal to N and is equal to N if $m = N$. Thus, our search for the maximized $L(Z)$ over Ω_{ULS} is at most N evaluations, but actually substantially less than N, since we stipulate that a hotspot not be more than fifty percent of the size of R.

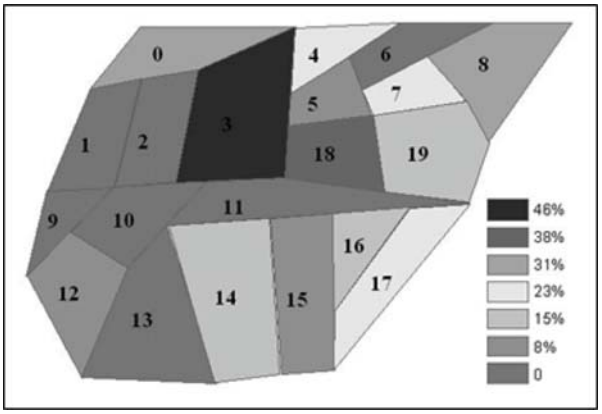


Figure 12.1. Illustrative data.

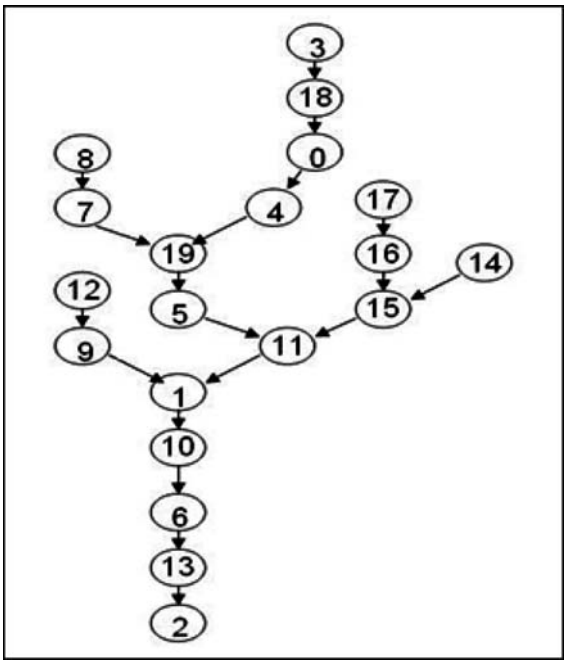


Figure 12.2. Cells topologically sorted.

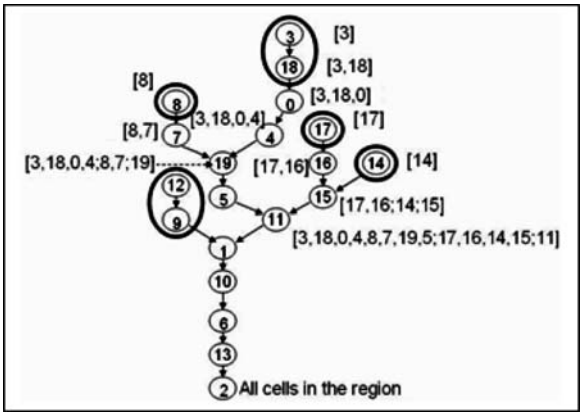


Figure 12.3. The ULS tree.

12.4 Computational Aspects

A consequence of the adaptive approach of the ULS scan statistic is that Ω_{ULS} must be computed fresh for each simulation run. Hence, it is important that the algorithm to construct the ULS tree be efficient, especially for a large

Table 12.1. Computational time for selected datasets.

Response model	Number of cells	Total population size	Time in seconds to do the task		
			Data simulation	ULS tree construction	Likelihood computation
Gamma	211	N/A	1	84	138
Binomial	211	21,100	18	14	<1
Binomial	12	3,067,740	232	<1	<1

tessellation. At the same time, realize that construction of the ULS tree is only a part of the overall computational effort. We can identify three main tasks involved in the whole process: Construction of the ULS tree, generation of simulated data and calculation of $L(Z)$ for each Z in the (reduced) parameter space. Major factors contributing to the execution time can be the type of the response model (discrete or continuous), population size, and complexity of the likelihood equations. These factors have effects on the three tasks in varying degrees. Table 12.1 illustrates the point. The numbers shown in the table are derived from 999 iterations of simulation runs with actual datasets. The results were obtained on a Dell Dimension® 8200 Series computer with Intel® Pentium® 4 2.40GHz CPU and 2.39GHz, 1.12GB RAM, running a Windows XP® operating system. The program was compiled using MicroSoft® Visual Studio® 2005.

Of the three datasets, the one with the gamma response model is the subject matter of the case study presented in Section 12.9. It is a part of a Pennsylvania biodiversity research project [Joly (1996), Myers *et al.* (2000)]. The second dataset is also a part of the same project. It consists of the percentage of the land under forest in each cell. All 211 cells are identical in shape and size. We processed the data to identify significantly forested parts of the state assuming the binomial response model with a population of 100 units of area for each cell. Details of the finding are not presented in this chapter. Only the processing time statistics are included in the table to underscore some contrasts between a continuous response model and a discrete response model with respect to the three computational tasks. The third dataset has only 12 cells. In none of the three cases presented in the table is construction of the ULS tree the most time-consuming task, but of all the three it is most so for the gamma distribution. Samples with the most distinct values are expected for a continuous distribution, resulting in more levels for the ULS tree than for a discrete distribution. The complexity of the likelihood equations for the gamma distribution is clearly reflected by the time it takes to compute likelihoods. The effect of the large population size in the case of the binomial response model is clear from the third dataset. We point out that the sampling involved for the binomial model

is actually from the multivariate hypergeometric distribution. To generate a vector $(y'_1, y'_2, \ldots, y'_N)$ from the $(N-1)$ dimensional hypergeometric distribution one needs to generate $t = y_1+y_2+\cdots+y_N$ random numbers, and t is potentially quite large. On the other hand, to generate a similar vector from the Dirichlet distribution for the gamma model, the generation of only N random numbers is required.

12.5 Testing Significance of the Scan Statistic

We will be primarily interested in determining the significance of the likelihood of a candidate zone with the maximum likelihood. The distribution of the scan statistic under the null hypothesis is intractable mathematically. Traditionally, the p-value of the statistic is determined using Monte Carlo methods. The process involves obtaining the conditional distribution of $Y_1, Y2, \ldots, Y_N$ under the null hypothesis conditioned on a suitable statistic. For binomial and Poisson response models, it is obtained by holding $Y_1 + Y_2 + \cdots + Y_N$ fixed at $y_1, +y_2 + \cdots + y_N$. This sum being sufficient for the respective parameter under investigation, the conditional distribution (multivariate hypergeometric and multinomial, respectively) is independent of the respective parameter. Simulated samples from the conditional distribution are used to construct the scan statistic for comparison with the observed scan statistic. The entire process for binomial and Poisson response models is straightforward. In some cases a sufficient statistic may not exist or may not be suitable, as will be seen with the gamma distribution in the next section.

12.6 Gamma Response Model

Binomial and Poisson response models have been studied extensively in hotspotting because of their wide applicability to epidemiology. Relatively, continuous distributions have received less attention. Here we use the gamma model to illustrate application of the ULS scan statistic to continuous distributions.

The gamma distribution has two parameters, k and β, where k is the index parameter and β is the scale parameter. Thus, if Y is a gamma variate,

$$E[Y] = k\beta \text{ and } Var\,[Y] = k\beta.$$

Here both k and β can vary from cell to cell, but additivity of the family of gamma distributions with respect to the index parameter suggests that we take

k to be proportional to the size A_a of the cell:

$$k_a = A_a/c,$$

where c is an unknown but whose value is the same for all cells in R. Thus, we have

$$E[Y_a] = \beta_a A_a/c,$$

and given a candidate zone Z, the null hypothesis to test absence of a hotspot becomes

$$H_0 : \beta_a \text{ are the same, say } \beta_0 \text{ for all cells in } R$$

against the alternative hypothesis

$$H_1 : \beta_a = \begin{cases} \beta_1' & \text{for all cells a in Z} \\ \beta_0' & \text{for all cells a outside Z and } \beta_1' > \beta_0'. \end{cases}$$

Incidentally, for the reparametrized gamma response model, the coefficient of variation square is

$$CV^2[Y_a] = c/A_a,$$

which says that the relative variability of the response decreases as the cell size increases and is a desirable property of the model.

The likelihood equation for estimating c_0 (c under H_0), β_0, c_1 (c under H_1), β_1', and β_0' take the form

$$\begin{aligned} &\sum_R A_a \left[\log(A_a/c_0) - \psi(A_a/c_0)\right] \\ =&(\sum_R A_a) \log(\sum_R y_a / \sum_R A_a) - \sum_R \left[A_a \log(y_a/A_a)\right] \end{aligned} \tag{12.1}$$

$$\beta_0 = c_0 \sum_R y_a / \sum_R A_a \tag{12.2}$$

$$\begin{aligned} &\sum_R \left[\log(A_a/c_1) - \psi(A_a/c_1)\right] \\ =&(\sum_{NZ} A_a) \log(\sum_{NZ} y_a / \sum_{NZ} A_a) + (\sum_Z A_a) \log(\sum_Z y_a / \sum_Z A_a) \\ &- \sum_R \left[A_a \log(y_a/A_a)\right] \end{aligned} \tag{12.3}$$

$$\beta_0' = c_1(\sum_{NZ} y_a / \sum_{NZ} A_a) \tag{12.4}$$

and

$$\beta_1' = c_1(\sum_{Z} y_a / \sum_{Z} A_a), \tag{12.5}$$

where $\sum_R$, $\sum_Z$ and $\sum_{NZ}$ denote summation of summands for all cells belonging to R, all cells inside Z, and all cells outside Z, respectively, and $\psi(\cdot)$ is the digamma function.

It is known that

$$g(t) = \log(t) - \psi(t),\ t \geq 0,$$

is strictly increasing with $g(0) = 0$ and $g(\infty) = \infty$. Further analysis shows that Equations (12.1) and (12.3) give unique solutions for c_0 and c_1, respectively. It has been verified that the Newton–Raphson algorithm gives rapid convergence. In the software implementation discussed in the next section, starting with moment estimates as initial guesses, satisfactory convergence never took more than ten iterations, and frequently took much fewer.

12.6.1 Monte Carlo simulation

As noted above, the gamma model is additive with respect to the index parameter so that, under the null hypothesis, $\sum_R Y_a$ is a gamma variable with parameters $(\beta, \sum_R A_a/c)$ and the conditional distribution of $(Z_1, Z_2, \ldots, Z_N), Z_a = Y_a / \sum Y_a$, given $\sum_R Y_a = t$ is Dirichlet with parameters $(k_1, k_2, \ldots, k_N)$. Thus, to generate simulated $y_1, y_2, \ldots, y_N$ we simulate generation of $Z_1, Z_2, \ldots, Z_N$ from the Dirichlet distribution with parameters $(k_1, k_2, \ldots, k_N)$ and compute $y_a = tZ_a$. To generate simulated $Z_1, Z_2, \ldots, Z_N$ it is enough to generate $x_1, x_2, \ldots, x_N$ from independent gamma distributions with $(\hat{\beta}_0, A_a/\hat{c}_0)$ as their respective parameters. Here $\hat{c}_0$ and $\hat{\beta}_0$ are MLEs of c and β under the null hypothesis that there is no hotspot. Once $x_1, x_2, \ldots, x_N$ are generated, one computes Z_a as $x_a/(x_1 + x_2 + \cdots + x_N)$ and finally, simulated response y_a as $y_a = tZ_a$.

12.7 Details of Software Implementation

The program was written in C++ using Microsoft Visual Studio 2005 on the Windows platform. While the software can still be considered as a prototype, consideration was given to two important objectives so that the current version

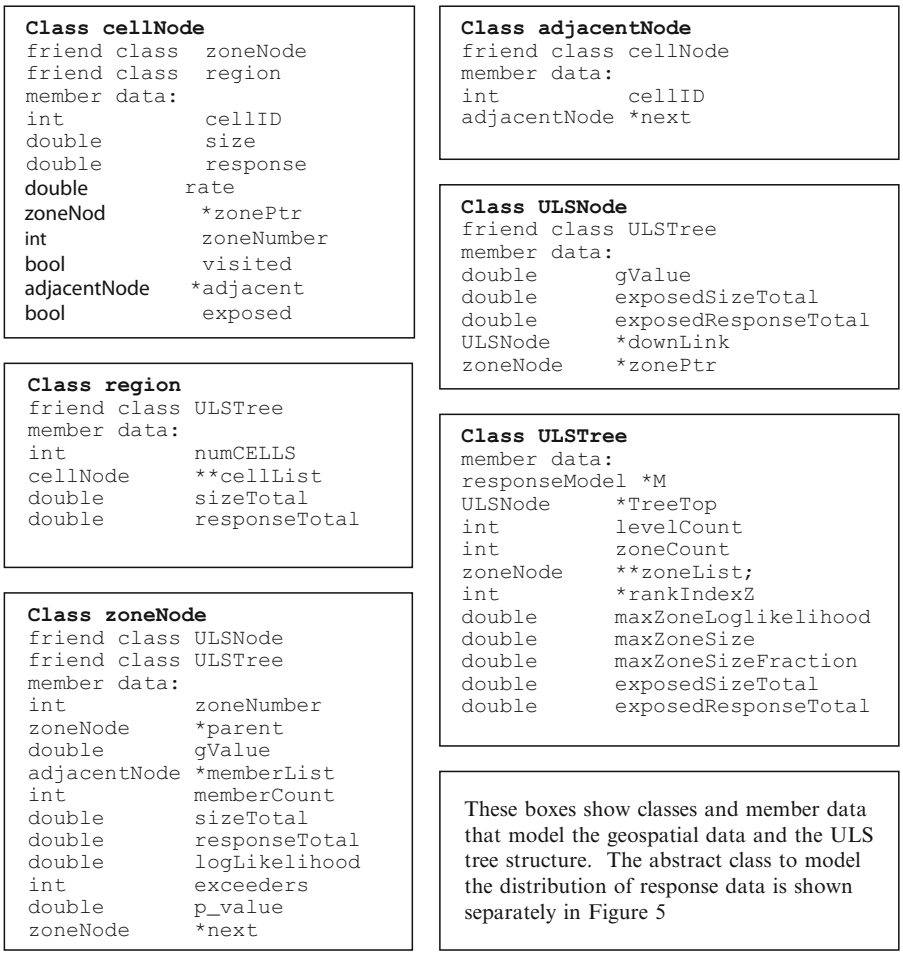


Figure 12.4. Overall data structure.

may form a basis for developing a production model. The first objective was to model the data structure to closely match the geostatistical model while using the computer memory economically. The second was to make it easily extensible if one wishes to add a new distribution to model responses or to deal with multi-response data or to construct confidence sets for hotspots. Figure 12.4 shows how the objectives were met. The figure shows essential data definitions. We suppress details of data, input/output and utility functions/methods used for debugging or that do trivial things.

The most basic object is cellNode, which is used to store the cell response value y (identifier named response in the program), area or N (represented by the identifier *size*), rate (y/N), pointer to the list of cells adjacent to the

given cell, and a link to the zone containing the cell. The Boolean data member visited is used to construct connected components during the building of the ULS tree. The exposed flag is set to true when the cell becomes a member of a connected zone during the ULS tree construction. The object region is an array of cellNode's. The object zoneNode represents a set of connected cells in the region and is a linked list of adjacentNodes, one adjacentNode for each cell in the zone, such that the y/N (that is, response/size) value for the cell is greater than or equal to a given g value. Each zoneNode except the root zone has a link to its parent. It also stores other attribute values of the zone. For each level of the ULS tree there is one instance of the object ULSNode. It points to a linked list of connected components/zones making up the level. Each instance of ULSNode has a link to the ULSNode instance representing the next level down (towards the root level) except for the root level ULSNode instance. Each ULSNode instance stores the corresponding g value. This linked list of ULSNodes is a ULSTree that we construct. Finally, one instance of the object ULSTree points to the linked list of ULSNodes making up one ULSTree. There are two ULSTree node instances, one pointing to the ULSTree constructed from the observed responses and the other pointing to the ULSTree constructed from a simulated copy of responses. For every simulation run we destroy the linked list of ULSNodes that makes up the tree and create a new list for the new tree. Both trees (observed and simulated) share the same storage to store observed and simulated responses and adjacency data. This is possible since all the information necessary for processing observed data is saved into the corresponding ULSTree structure consisting of ULSNodes and zoneNodes. The second objective of making the software flexible enough so that a new response model can be included in the program is achieved by means of an abstract class responseModel, as shown in Figures 12.4 and 12.5. In order to include a new response

```
Abstract class responseModel
friend   class ULSTree;
virtual void   computeMLE (void) = 0;
virtual void   computeMLE (zoneNode *zone)=0;
virtual void   computeLogLikelihoodNull(void)=0;
//             computes loglikelihood under H0
virtual double getLogLikelihoodNull(void)=0;
virtual void   computeZoneLogLikelihoodRatio(zoneNode* zone)=0;
virtual void   SimulateData(void)=0;
member data:
int            numCELLS
cellNode       **cellList
int            *V              // array to sort response rates
double         sizeTotal       // for the region
double         responseTotal   // for the region
```

Figure 12.5. Abstract response model class.

model one needs to create a new concrete class derived from the base abstract class responseModel and instantiate an object of the new concrete class in the main program on the lines of the currently available concrete classes for the binomial, Poisson and gamma models. The main program and the algorithm used to construct the ULSTree are outlined next.

12.8 Construction of the ULS Scan Tree

Our algorithm to construct the ULS tree begins with sorting the array of n cells representing the region in descending order by the g value (rate) using a sort index V, that is, $V[i]$ is the cellID with the i-th largest g value, for $i = 0, 1, 2, \ldots, n-1$. Here n is the number of cells in the region. The following algorithm expressed in pseudocode returns a pointer TreeTop to a linked list of ULSNode's. The number of nodes in this linked list will be the number of distinct g values obtained from the data plus 1. The first node is only a header node. Each of the remaining nodes in this list will point to the list of connected zones of the ULSTree occurring at one particular level corresponding to one distinct g value.

Algorithm construct ULSTree

```
oldgvalue = infinity
TreeTop = a new ULSNode with g value set to infinity.
    //points to an empty list of zones
    //serves as the header node for list of ULSNode
currentU = TreeTop
zoneCount = 0                  // count of the zones created
create an empty stack          // used in computation of connected
                               //component below
for i = 0 to n − 1 {
    currentcellID = the cellID whose rank is i; call it currentCell
    newgvalue = gvalue of currentcellID
    if currentcellID is exposed
            // do nothing, the cell is already exposed, continue with next
            // i value
    else {   // we have either a new level or we continue with the same
            // level in either case we have new connected zone
        if ( newgvalue < oldgvalue ) {   // we have a new level
        newU = new ULSNode
        set down link of currentU to newU
        currentU = newU
```

```
            clear visited tag of all cellNodes
            }
            // we have new zone
            Z = new zoneNode; initialize member data of Z
            Increase zoneCount by one
            Make currentcellID a member of Z - this also sets exposed tag
            to true
            append Z to the linked list of zones belonging to currentU
            ULS Node
            //at this point we do the standard depth-first traversal of all
            //cells  reachable from currentcellID and build up Z
            //as a connected zone that contains currentcellID
            // and all cells that are reachable from currentcellID
            // whose gvalue is greater than or equal to newgvalue
            set visited tag of currentCell
            push currentCell
            while (stack is not empty) {
                cellC = pop()
                for each neighbor neighborCell p of cellC do
                    if (visited tag of neighborCell is clear)
                        if (g value of neighborCell < newgvalue)
                            set visited tag of neighborCell
                        else if (g value of neighborCell is equal newgvalue) {
                            augment current zone Z with neighborCell
                            update all stats of the current zone Z
                            set visited tag of neighborCell
                            push neighborCell
                        }
                        else if (the neighborCell is not already in Z)
                            // case g value of neighborCell > newgvalue
                            set parent link of zone of neighborCell to Z
                            augment Z with all cells in the child zone
                        }
            }   // end of while stack is not empty
        }
        oldgvalue = newgvalue
    }   // end of for i = 0 to n - 1
    Update totals for the root level, current.
    // Finally, we construct an array of pointers pointing to each zone in the
    //tree in the order in which zones were created for an easy access
    // to the zones zoneList is an array of pointers to zoneNode
    i = 0;
```

```
currentU = TreeTop
while (curentU is not null) {
    Z = current'zonePtr
    while (Z is not null) {
        zoneList[i] = Z
        i = i + 1
        Z = Z → next
    }
    currentU = currentU→downlink
} // end of algorithm constructULSTree
```

12.9 A Case Study

In this section we present an application of the gamma response model to data collected to study biodiversity in the state of Pennsylvania. The section also illustrates input data and its format.

12.9.1 Description of Pennsylvania hexagonal biodiversity data

For the study, hexagonal tessellation of the state was used. The total number of hexagons covering the state is 211. The area of each hexagon is 635 sq km. The entire dataset consists of measurements, for each hexagon, of four different variables reflecting biodiversity or characteristics favorable to biodiversity. The four variables are bird species count, mammal species count, standard deviation of elevation, and percentage of the area covered by forest. Out of the 211 hexagonal areas in Pennsylvania, Table 12.2 shows the first five rows of the data for all the four variables. We will use the elevation data to locate highly rough terrain. For the purpose of measuring the elevation standard deviation a uniform grid of points was overlaid on the hexagonal tessellation. The elevation standard deviation is based on elevation measurements at these grid points.

Table 12.2. Biodiversity data for Pennsylvania hexagonal tessellates.

HexID	BirdSp	MamlSp	ElevSD	PctForst
1714	55	34	11	35.4
1827	58	37	32	84.3
1828	116	37	27	50.3
1829	96	34	17	25.3
1941	86	37	51	100.0

12.9.2 Pennsylvania elevation hotspot and illustrative data items and format

The gamma distribution appears to be an appropriate model to treat the elevation data. First we shall square each standard deviation to obtain the variance of the elevation measurements. Under the assumption of normality, the chi-square distribution is ideally suited for the transformed data. Even if basic measurements deviate from normality, the gamma distribution seems to be an acceptable model. Figure 12.6 shows only the first five lines of the data file actually used as input to the program. The input text file needs one line for each cell in the region. The first entry in each line is the cell identification number (cellID). The current version of the program requires that the cellID's be sequentially numbered starting with 0. For the current dataset, HexID's had to be translated sequentially into $0, 1, 2, \ldots, 210$. The second entry in each line is the "size" of the cell. The actual area of each hexagon is 635 sq km, but since the unit of measurement of size is irrelevant, we use 1 as the area of each cell. The third entry in each line is the value of the response variable for the cell. For Figure 12.6, it is the square of the elevation standard deviation so that the gamma model can be applied. The subsequent entries in each line are identification numbers of cells that are adjacent to the cell. Entries in each line are to be separated by one or more blank spaces or tabs. The end of line marks the end of data for the current cell. The format of the input data file described here remains the same irrespective of the response model used.

In addition to the basic data file in the form as shown in Figure 12.6, the user needs to specify the threshold, the maximal size that a potential hotspot could have. The threshold is a proper fraction relative to the size of the entire region. For the Pennsylvania data, we specified it as 0.50 for the elevation hotspot (as well as for the forest cover hotspot). In addition to the program run to detect the hotspot with respect to the high elevation standard deviation, the program was also run separately to detect the "coldspot," that is, the hotspot with respect to the low values of the elevation standard deviation, again with the threshold fraction of 0.50. The idea is to see if certain marginal hexagons qualify according to the program to be included in a hotspot as well as in a coldspot. An occurrence of one or more cells of this type could present a dilemma to decision makers. In our case three such cells were detected. The program outputs all

0	1	121	1	2			
1	1	1024	0	2	4	5	
2	1	729	0	1	3	5	6
3	1	289	2	6	7		
4	1	2601	1	5	12	11	

Figure 12.6. Input data file for elevation hotspot. The size is 1 here since all cells have the same area.

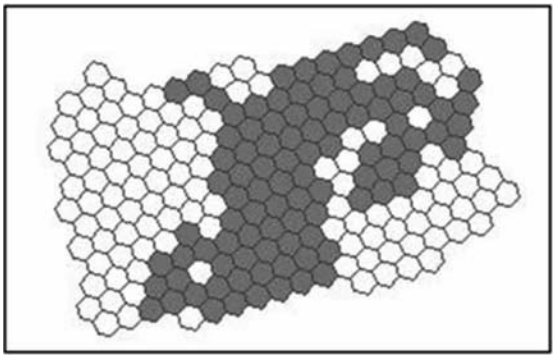


Figure 12.7. Elevation hotspot is in gray.

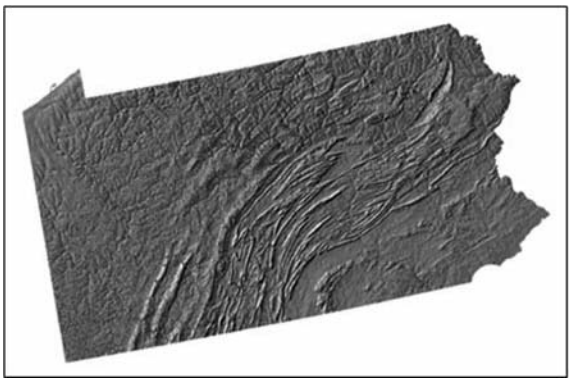


Figure 12.8. Topographical map of Pennsylvania.

hotspots, that is, the candidate zones with a p-value of 0.05 or less. With a little manual processing and inspection, by working towards the leaf nodes of the ULS tree, a maximal hotspot with no intersection with the coldspot was discovered. This hotspot is shown in Figure 12.7.

We show in Figure 12.8 a topographical map of Pennsylvania to facilitate comparison between the actual central high ridge terrain where rougher landscape is expected and the ULS hotspot.

12.10 Conclusions

We have presented the ULS scan statistic for geospatial hotspot detection and its object-oriented software implementation. The ULS scan statistic provides an effective means to handle arbitrarily shaped hotspots with significant reduction

of the parameter space. The software implementation contains an object representing the gamma response model, which is a continuous model, in addition to objects representing the more traditional discrete response models, binomial and Poisson. The flexibility of the software makes it convenient to introduce objects representing additional response models. A comparison between the gamma and the binomial response models with respect to the computational activity shows that for the gamma model construction of the ULS tree and likelihood calculations are more computer intensive, while Monte Carlo simulation is more so for the latter. Finally, a case study illustrating application of the gamma response model has been presented.

Acknowledgments

The first author acknowledges that this material is based upon work supported by (1) The National Science Foundation under Grant No. 0307010, and (ii) The United States Environmental Protection Agency under Grant No. CR-83059301 and No. R-828684-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

References

1. Biomedware (2001). *Software for the Environmental and Health Sciences*, Biomedware, Ann Arbor, MI.

2. Cressie, N. (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

3. Glaz, J., and Balakrishnan, N. (1999). *Scan Statistics and Applications*, Springer Publications, Netherlands.

4. Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan Statistics*, Springer Publications, New York.

5. Joly, K. (1996). Mammalian biodiversity in Pennsylvania at the USEPA 635 square kilometer hexagonal scale, *Master of Science thesis*, Pennsylvania State University, University Park, PA.

6. Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**(6), 1481–1496.

7. Kulldorff, M. (2006). *SaTScan v 7.0: Software for the spatial and space-time scan statistics*, Information Management Services Inc., Silver Spring, MD.

8. Kulldorff, M., and Nagarwala, N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine*, **14**, 799–810.

9. Kulldorff, M., Rand, K., Gherman, G., Williams, G., and DeFrancesco, D. (1998). *SaTScan v 2.1: Software for the spatial and space-time scan statistics*, National Cancer Institute, Bethesda, MD.

10. Myers, W., Bishop, J., Brooks, R., O'Connell, T., Argent, D., Storm, G., Stauffer, J., and Carline, R. (2000). Pennsylvania Gap Analysis Project; leading landscapes for collaborative conservation: Final report. School of Forest Resources, Cooperative Fish and Wildlife Research Unit, and Environmental Resources Research Institute. Pennsylvania State University, University Park, PA.

11. Patil, G.P. (2007). Statistical geoinformatics of geographic hotspot detection and multicriteria prioritization for monitoring, etiology, early warning and sustainable management for digital governance in agriculture, environment, and ecohealth, *Journal of Indian Society of Agricultural Statistics*, **61**, 132–146.

12. Patil, G.P., and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection, *Statistical Science*, **18**(4), 457–465.

13. Patil, G.P., and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental Ecological Statistics*, **11**, 183–197.

14. Patil, G.P., Acharya, R., Glasmier, A., Myers, W., Phoha, S., and Rathbun, S. (2006). Hotspot detection and prioritization geoinformatics for digital governance, In *Digital Government: Advanced Research and Case Studies* (Eds., H. Chen, L. Brandt, V. Gregg, R. Traunmüller, S. Dawes, E. Hovy, A. Macintosh, C. Larson), Springer, New York.

15. Patil, G.P., Acharya, R., Myers, W., Phoha, S., and Zambre, R. (2007). Hotspot geoinformatics for detection, prioritization, and security, In *Encyclopedia of Geographical Information Science* (Eds., S. Shekhar and H. Xiong), Springer, New York.

16. Patil, G.P., Acharya, R., and Phoha, S. (2007). Digital governance, hotspot detection, and homeland security, In *Encyclopedia of Quantitative Risk Analysis*, Wiley, New York.

17. Patil, G.P., Acharya, R., Modarres, R., Myers, W.L., and Rathbun, S.L. (2007). Hotspot geoinformatics for digital government. In *Encyclopedia of Digital Government*, Volume II (Eds. Ari-Veikko Anttiroiko and Matti Malkia), Idea Group Publishing, Hershey, PA, 919.

18. Patil, G.P., Joshi, S.W., and Rathbun, S.L. (2007). Hotspot geoinformatics, environmental risk, and digital governance, In *Encyclopedia of Quantitative Risk Analysis*, Wiley, New York.–927, Idea Group Reference, Hershey, PA.

19. Patil, G.P., Patil, V.D., Pawde, S.P., Phoha, S., Singhal, V., and Zambre, R. (2008). Digital governance, hotspot geoinformatics, and sensor networks for monitoring, etiology, early warning, and sustainable management, In *Geoinformatics for Natural Resource Management* (Ed. P.K. Joshi), Nova Science Publishers, New York (in press).

13

False Discovery Control for Scan Clustering

Marco Perone-Pacifico[1] **and Isabella Verdinelli**[1,2]

[1] *Department of Statistics, Sapienza University of Rome, Rome, Italy*
[2] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA*

Abstract: This chapter describes and summarizes methods for identifying the presence of clusters in a random field. The approach is based on controlling the fraction of false discoveries and considers a density estimator as the test statistic. A procedure called shaving is adopted for correcting the bias of the density estimator. This type of scanning for cluster identification does not use a window of fixed size; the role of the window size is played by the bandwidth of the kernel estimator. Clusters obtained using different bandwidths are combined in order to increase the detection power of the procedure.

In this chapter we stress some more intuitive aspects of these procedures and present some applications.

Keywords and phrases: False discovery control, false discovery rates, multiple hypothesis testing, scan clustering, kernel density estimators

13.1 Introduction

Identifying unusual clusters, among events scattered over space or time, is a problem that arises in a wide variety of applications. Typical examples include localization of centers of infection in epidemiology or, in the analysis of magnetic resonance images, detection of activity in different regions of the brain.

What constitutes an event and a cluster depends on each application, but from a statistical perspective, events are data points drawn from a spatial or temporal point process, and clusters are regions where the points are more dense.

An important scan statistics method for cluster detection [Glaz, Naus, and Wallenstein (2001), Patil and Taillie (2003)] counts the number of events N_s observed in a fixed window (such as a rectangle or circle) centered at each $s \in S$, where S is the domain of the point process. The null hypothesis that there are no clusters is tested with the statistic $T = \sup_{s \in S} N_s$, and the p-value for T

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_13,

is computed under the uniform distribution on S. The null is rejected if the p-value is not greater than the significance level.

The window scan clustering presented above gives information about the existence of clusters, but it neither identifies their number nor their location. For finding and localizing clusters, Perone-Pacifico *et al.* (2004, 2007) presented a method based on performing multiple tests in a random field's domain. The problem of controlling the type I error, the main task in multiple testing procedures, in those papers is approached through bounds on the fraction of rejection errors.

As in standard multiple testing problems, controlling the fraction of errors is an alternative to the traditional approach of controlling the overall probability of type I errors, also denoted as the family-wise error rate. In fact, controlling the family-wise error rate provides a strong guarantee, but it can be conservative in the sense of low power.

In Section 13.2 we present different approaches to controlling false rejections in multiple testing and discuss their advantages and disadvantages. Section 13.3 describes a method proposed by Perone-Pacifico *et al.* (2004, 2007) for identifying and localizing clusters in a point process. Section 13.4 shows how to correct the clusters' distortion due to the kernel estimator's bias, and Section 13.5 illustrates how to combine information across different bandwidths while maintaining type I error control. Section 13.6 provides a few numerical examples.

13.2 The Basics of Multiple Testing

Single test procedures are usually based on rejecting the null hypothesis when the p-value is smaller than or equal to the desired significance level α. This guarantees that the probability of a type I error is not greater than α.

Suppose now that m null hypotheses $H_{0,1}, \ldots, H_{0,m}$ need to be tested and, for $s = 1, \ldots, m$, denote by p_s the p-value relative to the single test of $H_{0,s}$.

A multiple test procedure that rejects all nulls corresponding to p-values not greater than α would guarantee that each null hypothesis is tested at level α, but it could result in a very high family-wise error rate (overall probability of false rejections)

$$\mathbb{P}(\text{at least 1 false rejection}) = \mathbb{P}(\bigcup_{s=1}^{m} \text{false rejection of } H_{0,s}) > \alpha.$$

Hence, to ensure control on type I errors, it is necessary to reduce the significance level of the single tests. This can be achieved by rejecting all null hypotheses whose corresponding p-values are either not greater than α/m

(*Bonferroni correction*) or not greater than the α-quantile of the distribution, under all the null hypotheses, of the minimum p-value $\min_{s=1,\cdots,m} p_s$.

The Bonferroni correction is simpler but, being based on the Bonferroni inequality, usually results in family-wise error rates much smaller than their target value. The second correction gives sharper type I error control but, since it requires the null distribution of the minimum p-value, it can be difficult to implement in the case of dependent test statistics. Both these procedures guarantee strong control of type I errors, namely that the resulting family-wise error rate is bounded by α. Conversely, they are not of practical relevance for a large number m of null hypotheses, since the low significance level of the single tests gives too little power to the whole procedure.

Based on these and other considerations, Benjamini and Hochberg (1995) suggested a different point of view on the problem of multiplicity and introduced a multiple testing error measure alternative to the family-wise error rate. This quantity, called the *false discovery rate*, is the expected proportion of type I errors among all rejected null hypotheses. Benjamini and Hochberg (1995) also proposed a multiple testing procedure that rejects all null hypotheses whose corresponding p-values are not greater than a data-dependent threshold. This procedure is proved to control the false discovery rate below the target level and to be often more powerful than traditional methods based on the family-wise error rate.

Since the pioneering paper of Benjamini and Hochberg, many contributors have proposed alternative error criteria and multiple testing procedures for different situations. We briefly review here only the aspects that will be useful to present our clustering procedure.

Genovese and Wasserman (2004) extended the theory by introducing the *realized false discovery rate* (called *false discovery proportion* in later papers): in a test procedure that rejects $H_{0,s}$ for all $s \in R$, the false discovery proportion is defined as the fraction of type I errors

$$\Gamma(R) = \frac{\#(R \cap S_0)}{\#R},$$

where S_0 is the unknown set of indices whose corresponding null hypotheses are true and $\#$ is the counting measure. The false discovery proportion Γ is conventionally set to be zero when $\#R = 0$.

The false discovery rate of Benjamini and Hochberg (1995) is the expected value of Γ,

$$FDR(R) = \mathbb{E}(\Gamma(R)).$$

As an alternative type I error measure, Genovese and Wasserman (2004) proposed the *false discovery exceedance* as the tail probability of false discovery proportion

$$FDX(R) = \mathbb{P}(\Gamma(R) > \gamma),$$

for given $\gamma \in (0, 1)$. Clearly, controlling the false discovery exceedance instead of the false discovery rate gives a stronger prevention from large fractions of type I error.

The same paper also gave a unified formulation of the multiple testing theory in terms of a threshold for p-values: a multiple test is equivalent to choosing a threshold t and rejecting $H_{0,s}$ for all $s \in L_t$, where

$$L_t = \{s : p_s \leq t\} \tag{13.1}$$

is the set of indices whose corresponding p-values are not greater than t. For example, the uncorrected multiple testing corresponds to choosing $t = \alpha$, and the Bonferroni method corresponds to $t = \alpha/n$.

In this framework, choosing a multiple testing procedure is equivalent to controlling features of the random process $\Gamma(L_t)$, and choosing the (possibly data-driven) threshold T that satisfies the desired criterion. Genovese and Wasserman (2004) proposed several procedures for choosing *confidence thresholds* to bound the false discovery exceedance below a target level α.

A simple method for controlling false discovery exceedance is presented in van der Laan, Dudoit, and Pollard (2006). This method, called *augmentation*, is based on multiple testing procedures that control the family-wise error: let R be a rejection set that controls the family-wise error rate at level α and let A be any set of indices, disjoint from R, such that

$$\#A \leq \frac{\gamma}{1-\gamma}\#R.$$

Rejecting all the nulls $H_{0,s}$ with $s \in A \cup R$ controls the false discovery exceedance at level α. Besides its operational relevance, this shows that testing procedures based on family-wise error rate are more conservative than those based on false discovery exceedance.

In the following sections we show how multiple testing can be extended to the continuous case (*i.e.* with uncountably many null hypotheses) and can be used to localize clusters in a field while keeping type I errors under control.

13.3 The Method

Perone-Pacifico *et al.* (2004, 2007) formalized a typical scan clustering problem as follows: the observations $(X_1, \ldots, X_n)$ are drawn from a point process on the space $S \subset \mathbb{R}^d$ with some intensity function ν. Clusters are regions in S with higher intensity, hence ν is assumed to be constant $\nu(s) = \nu_0$ for all s in an

unknown subset $S_0 \subset S$, while $\nu(s) > \nu_0$ for all $s \in S_1 = S_0^c$. Each connected component of S_1 is a cluster.

In order to detect and localize clusters, one can test for each $s \in S$ the local hypothesis

$$H_{0,s} : s \in S_0 \qquad \text{versus} \qquad H_{1,s} : s \notin S_0.$$

This rephrases the clustering problem in terms of testing simultaneously uncountably many null hypotheses.

Denoting by $f(s) = \frac{\nu(s)}{\int_S \nu(s)\, ds}$ the density of $X_1, \ldots, X_n$, the hypotheses can also be formulated as

$$H_{0,s} : f(s) = \frac{\nu_0}{\int_S \nu(s) ds} \qquad \text{versus} \qquad H_{1,s} : f(s) > \frac{\nu_0}{\int_S \nu(s) ds}. \tag{13.2}$$

The value of the integral $\int_S \nu(s) ds$ is not known, but it is not less than $\nu_0 \cdot \lambda(S)$, where λ denotes the Lebesgue measure. This consideration allows us to construct a conservative clustering procedure, testing

$$H_{0,s} : f(s) \leq \frac{1}{\lambda(S)} \qquad \text{versus} \qquad H_{1,s} : f(s) > \frac{1}{\lambda(S)}. \tag{13.3}$$

Note that testing the hypotheses in (13.3) can be considerably more conservative than testing (13.2) since, when there is an abundance of clusters, $\frac{\nu_0}{\int_S \nu(s) ds}$ can be much smaller than $\frac{1}{\lambda(S)}$.

13.3.1 False discovery control for uncountably many tests

The extension to the continuous case of multiple testing theory based on false discovery control is, at least in theory, almost straightforward: in a test procedure that rejects $H_{0,s}$ for all $s \in R$, the false discovery proportion is defined as

$$\Gamma(R) = \frac{\lambda(R \cap S_0)}{\lambda(R)}.$$

In the continuous case the test statistic Z is a random process over S. When the null distributions of $Z(s)$ are the same for all $s \in S$, rejection regions can be defined either through thresholds for p-values as in (13.1) or in terms of the test statistics

$$L_t = \{s \in S : Z(s) \geq t\}. \tag{13.4}$$

Perone-Pacifico *et al.* (2004) proposed a testing procedure that controls false discovery exceedance and determined a confidence threshold T such that, for a given $0 < \alpha < 1$ and $0 < \gamma < 1$,

$$\mathbb{P}(\Gamma(L_T) \geq \gamma) \leq \alpha. \tag{13.5}$$

The method consists in finding a confidence superset U that contains S_0 with probability at least $1-\alpha$. The confidence superset U is obtained through a sequence of tests over subsets of the field's domain

$$U = \bigcup \left\{ A \subset S : \mathbb{P}_0 \left(\sup_{s \in A} Z(s) > \sup_{s \in A} z(s) \right) \geq \alpha \right\}, \tag{13.6}$$

where z denotes the observed value of the test statistic Z and $\mathbb{P}_0$ denotes the distribution of Z under the global null hypotheses that $f(s) = \frac{1}{\lambda(S)}$ for all $s \in S$.

From (13.6) it seems that the determination of U requires considering every subset of S. Perone-Pacifico *et al.* (2004, Sections 2.1 and 2.2) presented an algorithm that reduces the number of sets A for which p-values must be computed to a level that is feasible in practice.

The set U permits us to define a confidence upper envelope for the false discovery proportion

$$\overline{\Gamma}(L_t) = \frac{\lambda(U \cap L_t)}{\lambda(L_t)},$$

with the property that

$$\mathbb{P}\left(\Gamma(L_t) \leq \overline{\Gamma}(L_t) \text{ for all } t\right) \geq 1-\alpha,$$

so that a threshold that satisfies (13.5) can be obtained as

$$T = \inf\{t \in \mathbb{R} : \overline{\Gamma}(L_t) \leq \gamma\}. \tag{13.7}$$

The same authors, in Perone-Pacifico *et al.* (2007), extended to the continuous case the augmentation procedure, initially proposed in van der Laan, Dudoit, and Pollard (2006) for finite S. They proved that if R controls the family-wise error rate at level α and A is any set, disjoint from R, with $\lambda(A) \leq \frac{\gamma}{1-\gamma}\lambda(R)$, then rejecting all null hypotheses in

$$\text{aug}_\gamma(R) = R \cup A$$

controls the false discovery exceedance at the same level α. Even if the result is valid for any set A with the proper size, it is advisable to let the augmentation set A contain the highest possible values of the test statistic.

In the same paper it is proved that the threshold in (13.7) can be obtained through augmentation, considering as the initial R the complement of the superset U defined in (13.6). This observation is crucial for dealing with the bias problem.

13.3.2 The test statistic

The scan clustering problem has been formalized in (13.3) as a multiple test on the value of a density, hence the most natural test statistic is a density estimator. We consider the kernel density estimator $\widehat{f}_H$,

$$\widehat{f}_H(s) = \frac{1}{n}\sum_{i=1}^{n} K_H(s - X_i), \tag{13.8}$$

where we dropped n from $\widehat{f}_H$ for ease of notation. The kernel K_H in (13.8) is based on a symmetric density φ, and it is defined for any $s \in S$ and for any non-singular diagonal bandwidth matrix H as

$$K_H(s) = \frac{1}{\det H}\varphi(H^{-1}s). \tag{13.9}$$

In order to allow combining density estimators over many bandwidths, an asymptotic approximation to the distribution of $\widehat{f}_H$ is needed that holds uniformly over H and s. Theorem 4 in Perone-Pacifico *et al.* (2007) provides such a result, extending the work of Chaudhuri and Marron (2000) to the case $\det H \to 0$ as $n \to \infty$.

For a fixed bandwidth matrix H, the asymptotic distribution of $\widehat{f}_H$ is Gaussian with expected value and covariance given by

$$\begin{aligned}
\mathbb{E}(\widehat{f}_H(s)) &= \int K_H(s-x)f(x)dx \\
\mathbb{C}(\widehat{f}_H(s), \widehat{f}_H(r)) &= \frac{1}{n}\Big(\int K_H(s-x)K_H(r-x)f(x)dx \\
&\qquad -\mathbb{E}(\widehat{f}_H(s))\mathbb{E}(\widehat{f}_H(r))\Big).
\end{aligned}$$

Note that the $\widehat{f}_H$ is biased since $\mathbb{E}(\widehat{f}_H) \neq f$.

The *test statistic process* used for cluster detection, for a given bandwidth matrix H, is

$$Z_H(s) = \frac{\widehat{f}_H(s) - \frac{1}{\lambda(S)}}{\sigma_H(s)},$$

where $\sigma_H(s)$ is the standard deviation of $\widehat{f}_H(s)$.

From the theorem quoted above, one can obtain the asymptotic null distribution of the test statistic: under the null hypothesis $Z_H(s)$ is approximately a normal random variable with mean less than or equal to 0. The variance $\sigma^2_H(s)$ of the process depends on the unknown density, but it can be either estimated from the data or approximated by the variance under the global null hypothesis,

$$\sigma^2_H(s) \approx \frac{1}{n}\left(\frac{1}{\lambda(S)}\int K_H(s-x)^2dx - \frac{1}{\lambda(S)^2}\right).$$

In order to determine the set U in (13.6), it is necessary to evaluate the null distribution of $\sup_{s\in A} Z_H(s)$. The tail probability $\mathbb{P}_0\left(\sup_{s\in A} Z_H(s) > z\right)$ can be approximated with the formulas in Adler (2000) or Worsley (1994, 1995) based on the expected Euler characteristic. Perone-Pacifico *et al.* (2004) used instead an approximation based on Piterbarg (1996, Theorem 7.1) that seems to be more accurate when the set A is not convex. For normal kernels, under the global null hypothesis $f(s) = \frac{1}{\lambda(S)}$, the approximation to the tail probability is

$$\mathbb{P}_0\left(\sup_{s\in A} Z_H(s) > z\right) \simeq \frac{\lambda(A)}{(4\pi)^{\frac{d}{2}} \det H} \left(\frac{\lambda(S)}{\lambda(S) - (4\pi)^{\frac{d}{2}} \det H}\right)^{\frac{d}{2}} \left(\frac{z}{\sigma}\right)^d \left(1 - \Phi\left(\frac{z}{\sigma}\right)\right),$$

where Φ is the univariate standard normal cumulative distribution function.

13.4 Clusters Shaving for Bias Correction

As mentioned, the kernel density estimate (13.8) is biased, thus a test based on $\widehat{f}_H$ (or, equivalently, on Z_H) does not really test (13.3), but it tests the biased null hypotheses

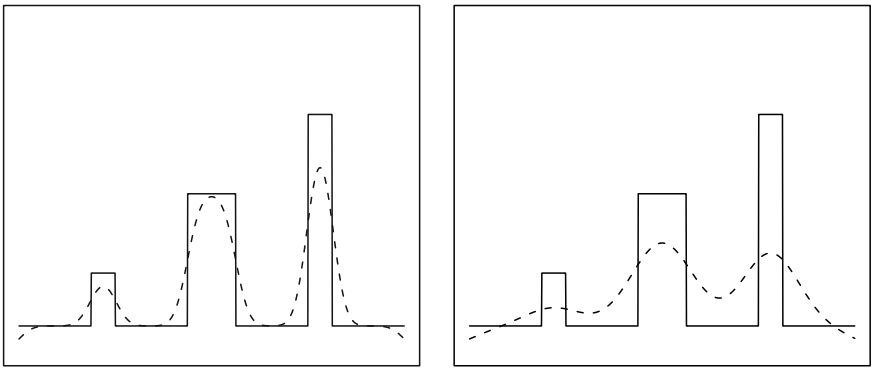
$$H_{0,s} : f_H(s) \leq \frac{1}{\lambda(S)} \qquad \text{versus} \qquad H_{1,s} : f_H(s) > \frac{1}{\lambda(S)}, \tag{13.10}$$

where

$$f_H(s) = \mathbb{E}[\widehat{f}_H(s)] = \int K_H(s-x) f(x)\, dx \neq f(s).$$

Figure 13.1 illustrates the effect of bias in cluster detection: the expected kernel density estimator (dashed lines) distorts the clusters, and the distortion increases with the bandwidth H (compare panels A and B). In particular, the clusters defined with respect to f_H are larger than those defined with respect to f. This could lead to an excess of false discoveries. Thus, some adjustment is needed to guarantee control of false discoveries at the prescribed level.

Since the goal here is to test for identifying locations of high intensity, bias correction is more feasible than in density estimation, because we only need to adjust the bias at the edges of level sets. Moreover, the types of errors one can make due to bias are asymmetric in nature: a bias that shrinks clusters does not increase the false discovery proportion, only biases that enlarge clusters can increase type I errors.



A: small bandwidth B: large bandwidth

Figure 13.1. Bias in kernel density estimation: The solid line is the true density f. The dashed line is the expected kernel density estimator f_H, for small (A) and large (B) bandwidths.

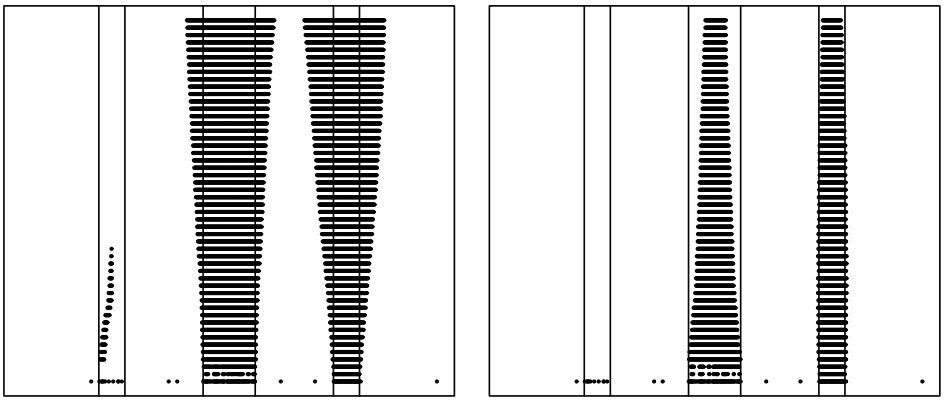
The bias correction procedure proposed in Perone-Pacifico *et al.* (2007), called *shaving*, is based on the augmentation procedure presented in Section 13.3. For a fixed bandwidth matrix H, the bias correction can be schematically summarized as follows:

1. let R_H be a rejection set controlling the family-wise error rate at level α for the biased null hypotheses in (13.10) (for instance, one can take R_H to be the complement of the superset U in (13.6));
2. let C_H be the support of the kernel K_H;
3. consider the *shaved* version of R_H as the Minkowsky difference between R_H and C_H
$$\mathrm{sh}(R_H) = \{s \in R_H : s + C_H \subset R_H\};$$
4. augment $\mathrm{sh}(R_H)$ as described in Section 13.3, the resulting clusters are the connected regions of the set $\mathrm{aug}_\gamma(\mathrm{sh}(R_H))$.

Perone-Pacifico *et al.* (2007, Theorem 5) proved that under mild separation conditions on the clusters, if the kernel has compact support, then $\mathrm{sh}(R_H)$ controls, family-wise error rate at level α and $\mathrm{aug}_\gamma(\mathrm{sh}(R_H))$ controls the false discovery exceedance, *i.e.*

$$\mathbb{P}(\Gamma(\mathrm{aug}_\gamma(\mathrm{sh}(R_H))) \geq \gamma) \leq \alpha.$$

The condition on compactness of the kernel's support does not seem to be crucial. Perone-Pacifico *et al.* (2007) used Gaussian kernels, which have unbounded support, taking C_H in step 2 to be an ellipse with radii equal to the bandwidths.



A: unshaved clusters B: shaved clusters

Figure 13.2. Bias correction: Vertical lines delimit the true clusters. Horizontal lines show not bias-adjusted (A) and bias-adjusted (B) rejection regions for different bandwidths.

Note that the rejection set $\text{aug}_\gamma(\text{sh}(R_H))$ is not necessarily a level set defined through some threshold as in (13.4). This set might contain some points whose corresponding value of the test statistic is lower than other points that are not in the set.

For the example of Figure 13.1, Figure 13.2 shows the rejection regions, both bias-adjusted (panel B) and not bias-adjusted (panel A), as a function of the bandwidth (y-axis). Due to the increasing bias of the density estimates, the size of the unshaved rejection region (A) increases with the bandwidth. This results in extra false discoveries. In the shaved regions (B) most extra false discoveries are eliminated although, especially for large bandwidths, bias adjustment yields low power. We will deal with this problem in the following section.

13.5 Power Increase Through Multiple Bandwidths

Bandwidth selection is always one of the main concerns when dealing with kernel density estimates. In density estimation, one usually tries to choose H to trade off bias and variance to obtain optimal squared error rates of convergence. In the clustering context this might not be practical.

Instead of choosing a bandwidth, Perone-Pacifico *et al.* (2007) suggested combining information across different bandwidths. Their procedure is based on repeating steps 1–3 in Section 13.4 for k different bandwidths $H_1, \ldots, H_k$ starting with sets R_H which controls the family-wise error rate at level α/k

instead of α. The final rejection region is obtained through augmentation of the union of all the shaved regions

$$R = \mathrm{aug}_\gamma \left(\bigcup_{j=1}^{k} \mathrm{sh}(R_{H_j}) \right).$$

Perone-Pacifico *et al.* (2007, Theorem 6) proved that the resulting rejection set R controls false discovery exceedance and has power close to the optimal with high probability, where the measure of power for a rejection region R is

$$\pi(R) = \frac{\lambda(R \cap S_1)}{\lambda(S)}.$$

In their Remark 6, they also give some hints on how to choose the set of bandwidths.

13.6 Examples

This section shows the results obtained by applying the methods described in the previous sections to two-dimensional data sets. All the data sets considered consist of points scattered in the unit square $[0,1]^2$ (in the cosmological data example, the data were normalized). Smoothing was performed using Gaussian kernels with diagonal bandwidth matrix

$$H = h \begin{pmatrix} \widehat{\sigma}_1 & 0 \\ 0 & \widehat{\sigma}_2 \end{pmatrix}.$$

For the parameter h, $k = 20$ equally spaced values were considered, ranging between the pixel size $1/256$ and the oversmoothing bandwidth $h_{OS} = 1.1 \times n^{-1/6}$.

In all cases, the goal was to control false discovery exceedance with $\alpha = 0.05$ and $\gamma = 0.1$.

In the two simulated examples the detected clusters are compared with the true ones, and the actual false discovery proportion and power are computed.

13.6.1 Mixture of uniforms

The first data set consists of $n = 15{,}000$ points from a mixture of uniform densities over $[0,1]^2$,

$$f(s) = \frac{256}{466} \times \begin{cases} 3 & s \in \text{ clusters 1 and 6} \\ 6 & s \in \text{ clusters 2 and 5} \\ 9 & s \in \text{ clusters 3 and 4} \\ 1 & \text{elsewhere.} \end{cases} \tag{13.11}$$

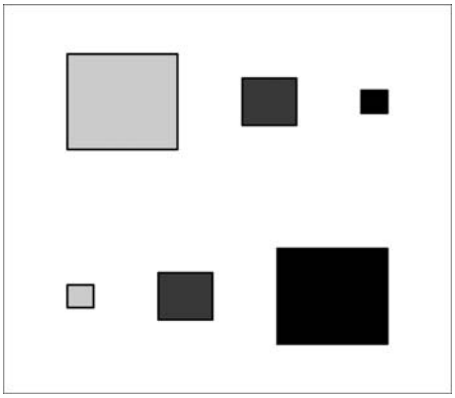


Figure 13.3. Contour plot of density in (13.11).

The generating density is shown in Figure 13.3, where the clusters are enumerated clockwise from the top left.

The left panels (A, C, and E) in Figure 13.4 show the clusters $\text{aug}(R_H)$ detected with small, intermediate, and large bandwidths, respectively, without bias correction. The shaving procedure removes most false discoveries, as shown in panels B, D, and F, which display the sets $\text{aug}(\text{sh}(R_H))$. These plots confirm that small bandwidths produce tests with low power due to the high variance of the kernel estimate, while for large bandwidths the amount of shaving reduces power again.

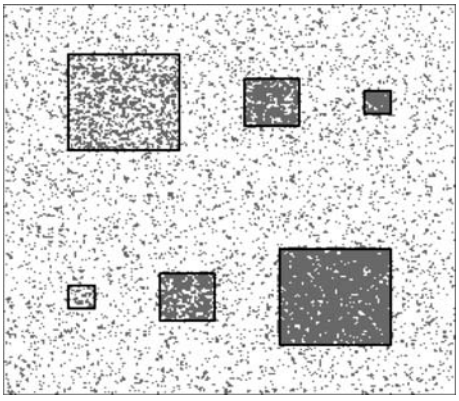
Figure 13.5 A shows the behavior of the false discovery proportion for the unshaved and shaved rejection regions. Clearly, the unshaved regions have far too many false discoveries, while the shaved ones always keep that measure under control. Plot B in the same figure shows that the loss of power due to shaving is small with respect to the reduction of the false discovery proportion.

Finally, Figure 13.6 shows the clusters detected using the procedure described in Section 13.5. The final set has no false discoveries, and its power is higher than the power obtained at each single bandwidth.

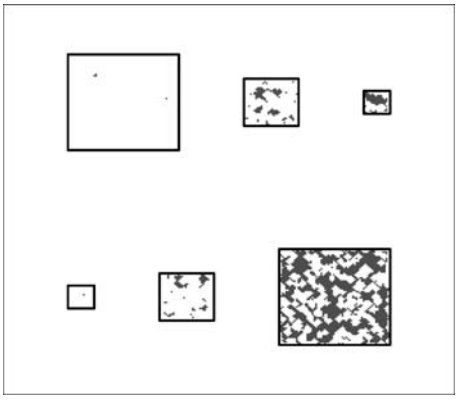
13.6.2 Smooth density with diagonal contours

The second simulated data set consists of $n = 15{,}000$ observations generated from two bivariate normal densities over a uniform background,

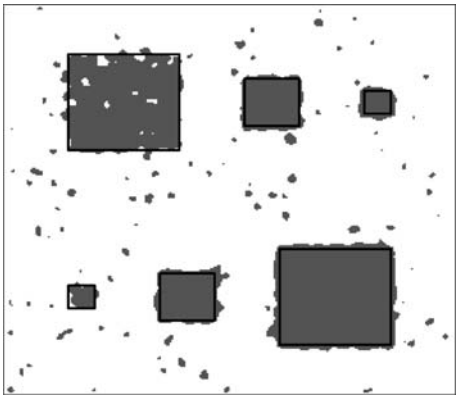
$$f(s) \propto \begin{cases} \varphi_2(s, \mu_1, \Sigma_1) & s \in \text{cluster 1} \\ \varphi_2(s, \mu_2, \Sigma_2) & s \in \text{cluster 2}, \\ 1 & \text{elsewhere} \end{cases}$$



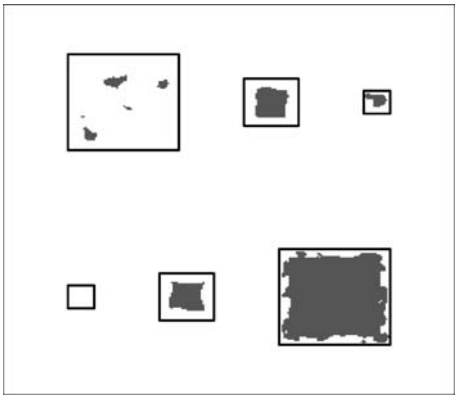
A: unshaved clusters, small h



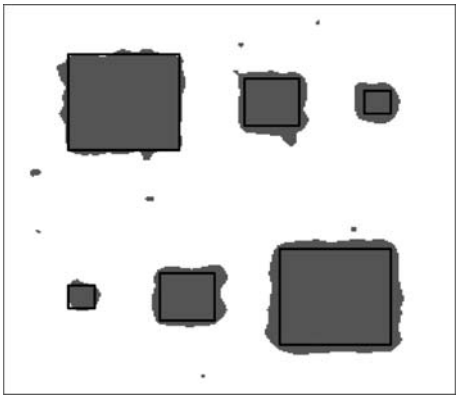
B: shaved clusters, small h



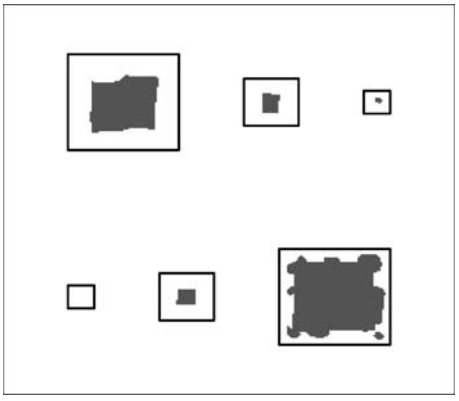
C: unshaved clusters, intermediate h



D: shaved clusters, intermediate h

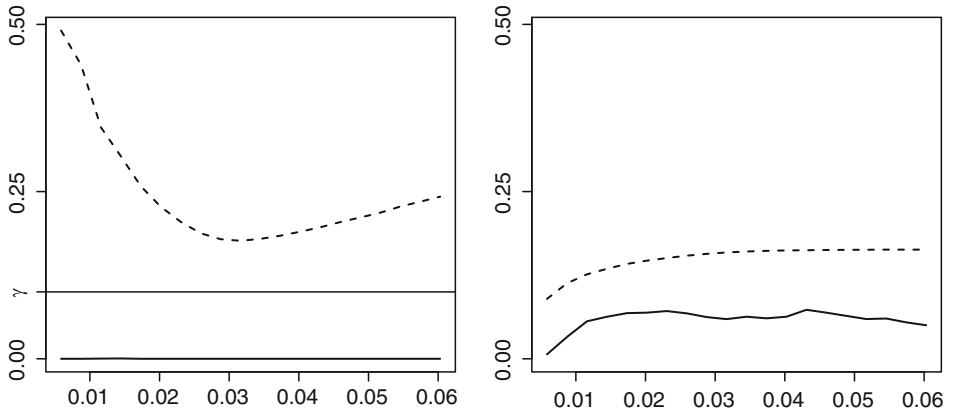


E: unshaved clusters, large h



F: shaved clusters, large h

Figure 13.4. Unshaved (left panels) and shaved (right panels) rejection regions for small, intermediate, and large bandwidths.



A: false discovery proportion B: power

Figure 13.5. False discovery proportion (panel A) and power (panel B) for unshaved (dashed) and shaved (solid) rejection regions as functions of bandwidth.

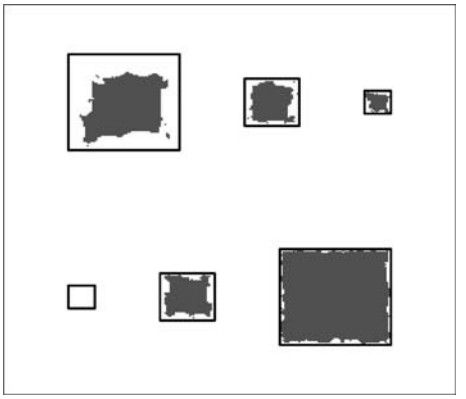


Figure 13.6. Clusters detected combining different bandwidths.

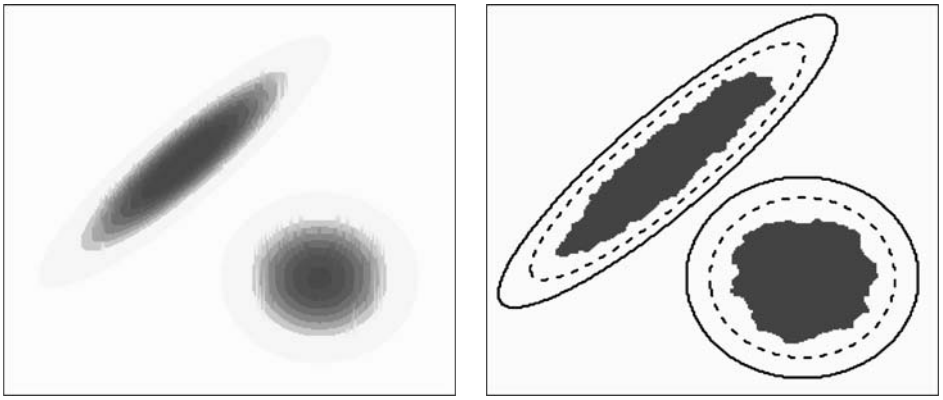
where $\varphi_2(\cdot, \mu, \Sigma)$ denotes the bivariate normal density with mean μ and covariance matrix Σ. In this example the parameters are

$$\mu_1 = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} \quad \Sigma_1 = \frac{1}{36} \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

and

$$\mu_2 = \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix} \quad \Sigma_2 = \frac{1}{72} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and the resulting density is shown in Figure 13.7 A.



A: true density B: detected clusters

Figure 13.7. True density (A) and detected clusters (B). In plot B, the solid line represents the conservative null hypothesis in (13.3), the dashed line the null in (13.2).

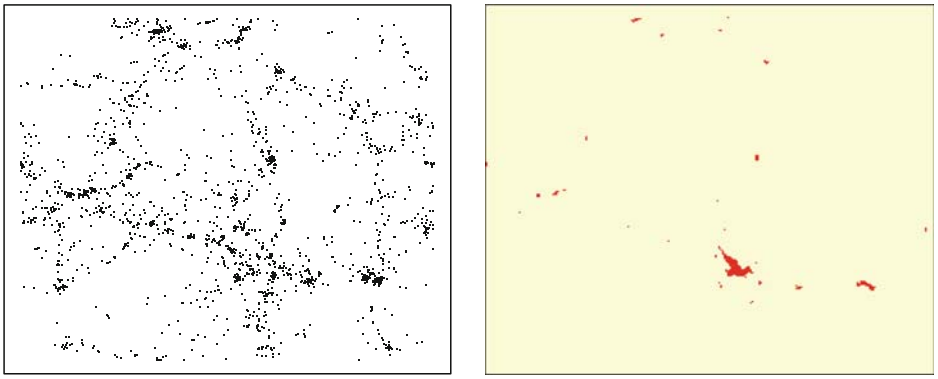
Also in this case, the final set, shown in Figure 13.7 B, does not have false discoveries, and its power is higher than the power obtained at each single bandwidth.

The clusters detected are much smaller than the true ones. This is partially due to the smoothness of the density, which makes the clusters less pronounced, but also to the fact that we are actually testing the conservative hypotheses (13.3) instead of the true nulls (13.2). In fact, looking at the level set of the density at $\frac{\nu_0}{\int_S \nu(s)ds}$ (dashed line in Figure 13.7 B), the clusters identified are still smaller than the true ones, but the difference is less relevant.

13.6.3 Cosmological data

Galaxy maps present a network of filaments of various sizes, called the cosmic web, with relatively empty voids between them and with clusters of galaxies located at the intersection of filaments. Astronomers are interested in studying the cosmic web, as it still retains a direct link to the matter distribution in the primordial universe and thus contains a wealth of direct information on the cosmic structure formation process.

Figure 13.8 A shows galaxies in a two-dimensional "slice" of the *Millennium run semi-analytic galaxy catalogue*, publicly available online at `http://www.mpa-garching.mpg.de/galform/agnpaper`. Figure 13.8 B shows the clusters detected using the method presented in the previous sections. Clearly, in this case there is no way to compare the clusters detected with the true ones, and



A: observed data points B: detected clusters

Figure 13.8. Observed data points (A) and detected clusters (B).

the evaluation of neither false discoveries nor power is possible, but it seems that the procedure has caught many of the clusters present.

References

1. Adler, R.J. (2000). On excursion sets, tube formulas and maxima of random fields. *The Annals of Applied Probability*, **10**, 1–74.

2. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

3. Chaudhuri, P. and Marron, J.S. (2000). Scale space view of curve estimation. *The Annals of Statistics*, **28**, 408–428.

4. Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, **32**, 1035–1061.

5. Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer, New York.

6. Patil, G.P. and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science*, **18**, 457–465.

7. Perone-Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, **99**, 1002–1014.

8. Perone-Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2007). Scan clustering: a false discovery approach. *Journal of Multivariate Analysis*, **98**, 1441–1469

9. Piterbarg, V.I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields*, American Mathematical Society. Providence, RI.

10. van der Laan, M., Dudoit, S. and Pollard, K. (2006). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, **3**.

11. Worsley, K.J. (1994). Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. *Advances in Applied Probability*, **26**, 13–42.

12. Worsley, K.J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *The Annals of Statistics*, **23**, 640–669.

14

Martingale Methods for Patterns and Scan Statistics

Vladimir Pozdnyakov[1] **and J. Michael Steele**[2]

[1] *Department of Statistics, University of Connecticut, Storrs, CT, USA*
[2] *Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA*

Abstract: We show how martingale techniques (both old and new) can be used to obtain otherwise hard-to-get information for the moments and distributions of waiting times for patterns in independent or Markov sequences. In particular, we show how these methods provide moments and distribution approximations for certain scan statistics, including variable length scan statistics. Each general problem that is considered is also illustrated with a concrete example confirming the computational tractability of the method.

Keywords and phrases: Scan, pattern, martingale

14.1 Introduction

The martingale method for waiting times for patterns in an independent sequence was pioneered in Li (1980), and in the intervening time many variations on the original idea have been developed. Our first aim here is to survey these developments using the unifying language of gambling teams. We further show how the martingale method can be extended to cover a great variety of problems in applied probability, including the occurrence of patterns in Markov sequences. One of the key intermediate steps is the development of a clear understanding of the distribution of the first time of occurrence of a pattern from a finite set of patterns. It is this general problem that leads to methods that are applicable to the theory of scan statistics.

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_14,

14.2 Patterns in an Independent Sequence

By $\{Z_n, n \geq 1\}$ we denote a sequence of independent and identically distributed (i.i.d.) random variables with values from a finite set $\Omega = \{1, 2, \ldots, M\}$, which we call the process alphabet. To specify the distribution of Z_n, we then set

$$p_1 = \mathbf{P}(Z_n = 1) > 0,\ p_2 = \mathbf{P}(Z_n = 2) > 0, \ldots,\ p_M = \mathbf{P}(Z_n = M) > 0.$$

By a *pattern* A we mean a finite ordered sequence of letters $a_1 a_2 \cdots a_m$ over the alphabet Ω. The random variable that is of most interest here is τ_{A}, the first time that one observes the pattern A as a run in the sequence $\{Z_n, n \geq 1\}$. Our main goal is to provide methods—and often explicit formulas—for the expected value, the higher moments, and the probability generating function of τ_{A}.

14.2.1 A gambling approach to the expected value

We begin with a construction that originates with Li (1980) and that we frame as a gambling scheme. Consider a casino game that generates the sequence $\{Z_n,\ n \geq 1\}$, say, as the output of a biased roulette wheel. Next consider a sequence of gamblers who arrive sequentially so that the nth gambler arrives right before the nth round when Z_n is generated. We also assume that this casino pays fair odds, so that a dollar bet on an event that has probability p would pay $1/p$ dollars to a winner (and zero to a loser).

Now we consider the strategy that is followed by the nth gambler, the one who arrives just before the nth round of play. For specificity, we first consider the gambler who enters just before the first round. This gambler bets one dollar that $Z_1 = a_1$. If Z_1 is not a_1, the gambler stops betting after having lost one dollar. If Z_1 yields a_1, the gambler wins $1/\mathbf{P}(Z_1 = a_1)$. He then continues to play, now betting his entire capital on $Z_2 = a_2$. If he loses, he stops gambling; otherwise, he increases his bet by the factor $1/\mathbf{P}(Z_2 = a_2)$. The gambler then continues in the same fashion until the entire pattern A is exhausted or until he has lost his original dollar, whichever comes first.

If the first gambler is very lucky and pattern A is observed after m rounds, the gambler stops and has total winnings of

$$\Big(\mathbf{P}(Z_1 = a_1)\mathbf{P}(Z_2 = a_2) \times \cdots \times \mathbf{P}(Z_m = a_m)\Big)^{-1}$$

dollars. Otherwise, the first gambler simply loses his initial bet of \$1.

In the meanwhile, additional gamblers enter the casino at successive times $2, 3, \ldots$ and each of these gamblers uses the same strategy that was used by the first gambler. That is, he bets successively on the letters of the pattern, each time "letting his stake ride." We then let X_n denote the total net gain of the

casino at the end of the nth round of play. The game was fair at each stage, so the stochastic process $\{X_n, \sigma(Z_1, ..., Z_n)\}$ is a martingale.

Now consider the random variable X_{τ_A}; this is the casino's net gain at the time when the pattern A is first observed. This random variable is well defined since τ_A is finite with probability one. In fact, it is easy to show that τ_A is bounded by a geometrically distributed random variable, so by Wald's lemma [or the optional stopping theorem, Williams (1991, p. 100)], we have the basic relation

$$\mathbf{E}(X_{\tau_A}) = 0.$$

Fortunately, we know more about X_{τ_A}. Specifically, we know that

$$X_{\tau_A} = \tau_A - W,$$

where W is the total amount of money that has been won by gamblers by time τ_A. The key observation is that W is not a random variable. The value of W is fully determined by the way in which the pattern A overlaps with itself.

Moreover, it is reasonably easy to calculate W. For a gambler to have any capital left when pattern A is first observed, that gambler needs to still be gambling, so in particular the gamblers who entered the game before $\tau_A - m + 1$ must have all lost their dollar. The gambler who enters the game at time $\tau_A - m + 1$ is the lucky guy who wins the most, but also some of those gamblers who entered after him may have some amount in their pockets.

The total amount of money that these few players have is represented by a certain measure of the overlapping of pattern A with itself. To describe this measure, we first consider $0 \le i, j \le m$ and set

$$\delta_{ij} = \begin{cases} 1/\mathbf{P}(Z_1 = a_i), & \text{if } a_i = a_j, \\ 0, & \text{otherwise.} \end{cases}$$

With this notation we then find the explicit formula

$$W = \delta_{11}\delta_{22}\cdots\delta_{mm} + \delta_{21}\delta_{32}\cdots\delta_{mm-1} + \cdots + \delta_{m1}, \tag{14.1}$$

so from our earlier observation that $\mathbf{E}(X_{\tau_A}) = 0$, we find

$$\mathbf{E}(\tau_A) = \delta_{11}\delta_{22}\cdots\delta_{mm} + \delta_{21}\delta_{32}\cdots\delta_{mm-1} + \cdots + \delta_{m1}. \tag{14.2}$$

The relation (14.2) really is quite explicit, and it provides an easily applied answer to our first question, say, as one sees in the following example.

Example 14.2.1 Let $\Omega = \{1, 2\}$ and consider the pattern 1121 of length 4. We then have

$$\mathbf{E}(\tau_A) = W = (p_1 \times p_1 \times p_2 \times p_1)^{-1} + (p_1)^{-1}.$$

14.2.2 Gambling on a generating function

By a natural modification of the preceding method, one can obtain a formula for the generating function of τ_{A}. The trick is to change the initial bet for each gambler. Now instead of \$1, the nth gambler starts his betting by placing a bet of size α^n, where $0 < \alpha < 1$. Let $\alpha^{\tau_{\mathrm{A}}} W(\alpha)$ be the total winnings of all the gamblers by time τ_{A}. As before, we let X_n denote the casino's gain at the end of the nth round and $\{X_n\}$ is a martingale. For convenience, we denote the total accumulated winnings of the gamblers when the pattern A is first observed by $\alpha^{\tau_{\mathrm{A}}} W(\alpha)$. Again, the key is that we have a nice relation for the casino's net gain $X_{\tau_{\mathrm{A}}}$. Specifically, we have

$$\begin{aligned} X_{\tau_{\mathrm{A}}} &= \alpha^1 + \alpha^2 + \cdots + \alpha^{\tau_{\mathrm{A}}} - \alpha^{\tau_{\mathrm{A}}} W(\alpha) \\ &= \alpha \frac{\alpha^{\tau_{\mathrm{A}}} - 1}{\alpha - 1} - \alpha^{\tau_{\mathrm{A}}} W(\alpha) \\ &= \alpha^{\tau_{\mathrm{A}}} \left(\frac{\alpha}{\alpha - 1} - W(\alpha) \right) - \frac{\alpha}{\alpha - 1}. \end{aligned}$$

As in the previous subsection, $W(\alpha)$ is not a random variable, and it has an explicit representation:

$$W(\alpha) = \delta_{11}\delta_{22}\cdots\delta_{mm-1}/\alpha^{m-1} + \delta_{21}\delta_{32}\cdots\delta_{mm-1}/\alpha^{m-2} + \cdots + \delta_{m1}/1.$$

The optional stopping theorem implies

$$0 = \mathbf{E}(X_{\tau_{\mathrm{A}}}) = \mathbf{E}(\alpha^{\tau_{\mathrm{A}}}) \left(\frac{\alpha}{\alpha - 1} - W(\alpha) \right) - \frac{\alpha}{\alpha - 1}.$$

When we solve this relation for $\mathbf{E}(\alpha^{\tau_{\mathrm{A}}})$, we obtain

$$\mathbf{E}(\alpha^{\tau_{\mathrm{A}}}) = \left(1 + \frac{1 - \alpha}{\alpha} W(\alpha) \right)^{-1}.$$

Again, this is an explicit usable formula, as one sees in the following example.

Example 14.2.2 Let $\Omega = \{1, 2\}$ and again consider the pattern 1121. One then has

$$W(\alpha) = \frac{\alpha^{-3}}{p_1^3 p_2} + \frac{1}{p_1},$$

so by substitution one has

$$\begin{aligned} \mathbf{E}(\alpha^{\tau_{\mathrm{A}}}) &= \frac{p_1^3 p_2 \alpha^4}{1 - \alpha + \alpha^3 (1 - p_2 \alpha) p_1^2 p_2} \\ &= p_1^3 p_2 \alpha^4 + p_1^3 p_2 \alpha^5 + p_1^3 p_1 \alpha^6 + p_1^3 p_2 (1 - p_1^2 p_2) \alpha^7 + o(\alpha^7). \end{aligned}$$

As a check, one should note that this formula can be used to confirm the calculation of the mean from our first example:

$$\frac{\partial \mathbf{E}(\alpha^{\tau_{\mathrm{A}}})}{\partial \alpha}\Big|_{\alpha=1} = \frac{1}{p_1^3 p_2} + \frac{1}{p_1} = E\tau_{\mathrm{A}}.$$

14.2.3 Second and higher moments

In theory, the ability to compute the probability generating function also gives one the higher moments, but in practice it is often useful to have an alternative method. Here it also seems instructive to show how the method of sequential gamblers can be used to find $\mathbf{E}(\tau_{\mathrm{A}}^2)$.

This time the trick is that the gambler who joins the game in the nth round will bet n dollars. If, as always, we let X_n denote the casino's net gain after n rounds, then X_n is again a martingale. Moreover, in this case one can check that at the stopping time τ_{A} we have

$$\begin{aligned} X_{\tau_{\mathrm{A}}} &= 1 + 2 + \cdots + \tau_{\mathrm{A}} \\ &\quad -(\tau_{\mathrm{A}} - m + 1)\delta_{11}\delta_{22}\cdots\delta_{mm} \\ &\quad -(\tau_{\mathrm{A}} - m + 2)\delta_{21}\delta_{32}\cdots\delta_{mm-1} \\ &\quad \cdots \\ &\quad -(\tau_{\mathrm{A}} - m + m)\delta_{m1} \\ &= 1 + 2 + \cdots + \tau_{\mathrm{A}} - \tau_{\mathrm{A}} W - N \\ &= \frac{\tau_{\mathrm{A}}^2 + \tau_{\mathrm{A}}}{2} - \tau_{\mathrm{A}} W - N, \end{aligned}$$

where

$$N = -\delta_{11}\delta_{22}\cdots\delta_{mm}(m-1) - \delta_{21}\delta_{32}\cdots\delta_{mm-1}(m-2) - \cdots - \delta_{m1}0.$$

It is now time to apply the optional stopping theorem, but in this case the increments of X_n are no longer uniformly bounded, so a more refined version of Doob's optional stopping theorem is needed. Here we can use the stopping time theorem of Shiryaev (1995, p. 485) since we have $X_n = O(n^2)$ and since $\mathbf{P}(\tau_{\mathrm{A}} > n)$ decays at an exponential rate. The application of this optional stopping theorem leads us to

$$0 = \mathbf{E}(X_{\tau_{\mathrm{A}}}) = \mathbf{E}(\tau_{\mathrm{A}}^2)/2 + \mathbf{E}(\tau_{\mathrm{A}})/2 - W\mathbf{E}(\tau_{\mathrm{A}}) - N.$$

Solving this equation for $\mathbf{E}(\tau_{\mathrm{A}}^2)$ gives us

$$\mathbf{E}(\tau_{\mathrm{A}}^2) = (2W-1)\mathbf{E}(\tau_{\mathrm{A}}) + 2N = 2W^2 - W + 2N,$$

and as a corollary we have the nice formula

$$\mathbf{Var}(\tau_{\mathrm{A}}) = W^2 - W + 2N.$$

Naturally, variations of this technique can be applied to obtain formulas for any moment. For example, to find an expression for the third moment, the nth gambler's bet should now be taken to be n^2.

Example 14.2.3 For the traditional sample space $\Omega = \{1, 2\}$ and the pattern 1121 we now find

$$N = -\frac{3}{p_1 \times p_1 \times p_2 \times p_1},$$

and

$$\mathbf{Var}(\tau_{\mathrm{A}}) = \left(\frac{1}{p_1} + \frac{1}{p_1^3 p_2}\right)^2 - \frac{1}{p_1} - \frac{7}{p_1^3 p_2}.$$

Here it is interesting to note that when either $p_1 \to 0$ or $p_2 \to 0$ one has the limit relation

$$\frac{\mathbf{E}(\tau_{\mathrm{A}})}{\mathbf{Var}(\tau_{\mathrm{A}})^{1/2}} \to 1.$$

Moreover, there is an intuitive explanation for this limit. When either $p_1 \to 0$ or $p_2 \to 0$ the occurrence of the pattern 1121 becomes a rare event. By the clumping heuristic [c.f. Aldous (1989)], one then expects the distribution of τ_{A} to be well approximated by an exponential distribution, and for an exponential X we have the equality $\mathbf{E}(X) = \sqrt{\mathbf{Var}(X)}$.

14.3 Compound Patterns and Gambling Teams

In many important applications—such as scans—one is concerned about the waiting time until the first occurrence of one out of *many* patterns from a *finite list of patterns*. Here we call a finite collection of K patterns $\{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_K\}$ a *compound pattern* and denote it simply by $\mathcal{A}$. Now, if τ_{A_i} denotes the first time until the pattern A_i has been observed as a completed run in the i.i.d. series $Z_1, Z_2, \ldots$, then the new random variable of interest is

$$\tau_{\mathcal{A}} = \min\{\tau_{\mathrm{A}_1}, \ldots, \tau_{\mathrm{A}_K}\}.$$

In words, $\tau_{\mathcal{A}}$ is the first time when we observe a pattern from $\mathcal{A}$, and one should note that without loss of generality we assume that in $\mathcal{A}$ no pattern is a subblock of another.

Gerber and Li (1981) studied compound patterns with the help of an appropriate Markov chain imbedding. We use an alternative method that has several benefits. In particular, the new method gives us clear hints on how we should extend the martingale approach to the case of Markov dependent trials. It also guides us when we consider the case of highly regular patterns, such as those associated with scans or structured motifs.

14.3.1 Expected time

It seems natural in the case of compound pattern $\mathcal{A}$ to introduce K *gambling teams*. The gamblers from each gambling team will bet on a pattern from the list $\mathcal{A}$. But now the problem is that the total amount of winnings of all the gamblers at time $\tau_{\mathcal{A}}$ is a random variable. It depends on how the game is stopped.

However, if one knows which simple pattern from $\mathcal{A}$ triggered the stop, then the winnings of a gambling team are not random. This amount is fully determined by the overlapping of two patterns: (1) the pattern associated with the gambling team, and (2) the pattern associated with the *ending scenario*. An explicit expression for this amount will be given a bit later.

As we will demonstrate in a moment, it is beneficial to allow every gambling team to have their own size for an initial bet. More specifically, let y_j be an amount with which the gambler from the jth gambling team (the team that bets on A_j) starts his betting. Let $W_{ij}y_j$ be total winnings of the jth gambling team in the case when the game was ended by the ith scenario (i.e., the pattern A_i is observed at time $\tau_{\mathcal{A}}$). If X_n is, as before, the net casino gain, then it is clear that it forms a martingale, because a weighted sum of martingales is a martingale. The stopped martingale $X_{\tau_{\mathcal{A}}}$ is given by

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{K} y_j \tau_{\mathcal{A}} - \sum_{i=1}^{K} \sum_{j=1}^{K} W_{ij} y_j 1_{E_i},$$

where 1_{E_i} is the indicator that the game is ended by the ith scenario.

There is an analogy between the way gambling teams are used here and the notion of hedging in finance. The trick, analogous to arbitrage constructions, is to choose weights y_j in such a way that the total winnings of all the teams $\sum_{j=1}^{K} W_{ij}y_j$ is equal to 1 regardless of an ending scenario. Now, if the vector $\{y_j\}_{1\le j\le K}$ is a solution of the linear system

$$\sum_{j=1}^{K} W_{ij} y_j = 1, \quad 1 \le i \le K, \tag{14.3}$$

then the stopped martingale is given by

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{K} y_j \tau_{\mathcal{A}} - 1.$$

This puts us on familiar ground. By another application of the optional stopping theorem, we obtain a computationally effective representation of $\mathbf{E}(\tau_{\mathcal{A}})$.

Theorem 14.3.1 *If vector $\{y_j\}_{1\le j\le K}$ solves the linear system (14.3), then the expected value of $\tau_{\mathcal{A}}$ is given by*

$$\mathbf{E}(\tau_{\mathcal{A}}) = \frac{1}{\sum_{j=1}^{K} y_j}.$$

Here we should make two technical comments. First, in the course of their Markov imbedding method, Gerber and Li (1981) showed that the matrix W_{ij} is nonsingular if no pattern from $\mathcal{A}$ is a subpattern of another. Consequently, the solution $\{y_j\}_{1\le j\le K}$ always exists.

Second, there is an explicit formula for W_{ij}. For example, consider two patterns $\mathrm{A} = a_1a_2\cdots a_m$ and $\mathrm{B} = b_1b_2\cdots b_l$. Next, we consider the measure of t-overlap of a suffix of A with a prefix of B that is given by the formula

$$\delta_t(A,B) = \begin{cases} \dfrac{1}{\prod_{s=1}^{t}\mathbf{P}(Z_1=b_s)}, & \text{if } b_1=a_{m-t+1}, b_2=a_{m-t+2}, ..., b_t=a_m, \\ 0, & \text{otherwise.} \end{cases}$$

Now, if the jth gambling team bets on A, and the ith ending scenario is associated with pattern B, then

$$W_{ij} = \sum_{t=1}^{\min(m,l)} \delta_t(A,B).$$

14.3.2 The generating function and the second moment

The method of gambling teams can be used to obtain a formula for the probability generating function $\mathbf{E}(\alpha^{\tau_{\mathcal{A}}}), 0<\alpha<1$, for any compound pattern $\mathcal{A}$. The solution is a little more complicated, but it mainly calls for the systematic elaboration of ideas that we have already seen. Here we consider the same number of gambling teams and ending scenarios that we used before, but now the gambler from the jth team who joins the game in the nth round will place an initial bet of size $y_j\alpha^n$, where the weights $\{y_j\}_{1\le j\le K}$ will be chosen later.

Let $W_{ij}(\alpha)y_j\alpha^{\tau_{\mathcal{A}}}$ denote the winnings of the jth gambling team when the game ends by the ith ending scenario. If X_n denotes the martingale that gives us the casino's net gain at time n, then the stopped martingale $X_{\tau_{\mathcal{A}}}$ is given by

$$X_{\tau_{\mathcal{A}}} = \alpha\frac{\alpha^{\tau_{\mathcal{A}}}-1}{\alpha-1}\sum_{j=1}^{K} y_j - \sum_{i=1}^{K}\sum_{j=1}^{K} W_{ij}(\alpha)y_j\alpha^{\tau_{\mathcal{A}}}1_{E_i},$$

where, as before, 1_{E_i} is the indicator of the ith ending scenario.

Again, the key fact is that $W_{ij}(\alpha)$ is not a random variable. If the jth gambling team bets on pattern A, and the ith ending scenario is linked with pattern B, then

$$W_{ij}(\alpha) = \sum_{t=1}^{\min(m,l)} \delta_t(A,B)\alpha^{1-t}, \tag{14.4}$$

where $\delta_t(A, B)$ is defined as in the preceding section. If weights $\{y_j(\alpha)\}_{1\le j\le K}$ are chosen such that

$$\sum_{j=1}^{K} W_{ij}(\alpha) y_j(\alpha) = 1, \quad 1 \le i \le K, \tag{14.5}$$

then the stopped martingale $X_{\tau_{\mathcal{A}}}$ is given by

$$X_{\tau_{\mathcal{A}}} = \alpha \frac{\alpha^{\tau_{\mathcal{A}}} - 1}{\alpha - 1} \sum_{j=1}^{K} y_j(\alpha) - \alpha^{\tau_{\mathcal{A}}}.$$

After taking the expectation, a little algebra leads one to a strikingly simple formula for the generating function for $\tau_{\mathcal{A}}$.

Theorem 14.3.2 *If the vector $\{y_j(\alpha)\}_{1\le j\le K}$ solves the linear system (14.5), then*

$$\mathbf{E}(\alpha^{\tau_{\mathcal{A}}}) = 1 - \frac{1}{1 + \sum_{j=1}^{K} y_j(\alpha)\alpha/(1-\alpha)}.$$

We can use Theorem 14.3.2 to obtain the higher moments of $\tau_{\mathcal{A}}$, but it is also possible to use the method of gambling teams more directly. For example, to compute the second moment of $\tau_{\mathcal{A}}$ we ask the gambler from the jth team that starts gambling in the nth round to place an initial bet of $y_j + nz_j$ dollars on the first letter of A_j and to continue betting his fortune on the subsequent letters of A_j until he either loses or until some gambler observes a pattern from $\mathcal{A}$.

This time we write the winnings of the jth team in the case of the ith ending scenario by the sum

$$W_{ij} y_j + \tau_{\mathcal{A}} W_{ij} z_j + N_{ij} z_j,$$

where W_{ij} is as before, but where N_{ij} is a new quantity for which we will give an explicit formula shortly. The casino's net gain at time $X_{\tau_{\mathcal{A}}}$ then is given by

$$\begin{aligned} X_{\tau_{\mathcal{A}}} = & \sum_{j=1}^{K} y_j \frac{\tau_{\mathcal{A}}(\tau_{\mathcal{A}}+1)}{2} + \sum_{j=1}^{K} z_j \tau_{\mathcal{A}} \\ & - \sum_{i=1}^{K} \left(\sum_{j=1}^{K} W_{ij} y_j \tau_{\mathcal{A}} + \sum_{j=1}^{K} N_{ij} y_j + \sum_{j=1}^{K} W_{ij} z_j \right) 1_{E_i}. \end{aligned}$$

Now, if weights $\{y_j\}_{1\le j\le K}$ and $\{z_j\}_{1\le j\le K}$ are such that

$$\begin{aligned} \sum_{j=1}^{K} W_{ij} y_j &= 1, \quad 1 \le i \le K, \\ \sum_{j=1}^{K} (N_{ij} y_j + W_{ij} z_j) &= 1, \quad 1 \le i \le K, \end{aligned} \tag{14.6}$$

then the stopped martingale is equal to

$$\sum_{j=1}^{K} y_j \frac{\tau_{\mathcal{A}}(\tau_{\mathcal{A}}+1)}{2} + \sum_{j=1}^{K} z_j \tau_{\mathcal{A}} - \tau_{\mathcal{A}} - 1.$$

After the application of the optional stopping theorem we obtain a formula for the second moment.

Theorem 14.3.3 *If* $\{y_j\}_{1\le j\le K}$ *and* $\{z_j\}_{1\le j\le K}$ *solve the linear system (14.6), then*

$$\mathbf{E}(\tau_{\mathcal{A}}^2) = \frac{1 + (1 - \sum_{j=1}^{K} z_j - \sum_{j=1}^{K} y_j/2)\mathbf{E}(\tau_{\mathcal{A}})}{\sum_{j=1}^{K} y_j/2}.$$

As we mentioned above, N_{ij} is just another measure of the overlap of two patterns. Specifically, if the jth gambling team bets on pattern A and the ith ending scenario corresponds to pattern B, then we have the explicit recipe:

$$N_{ij} = \sum_{t=1}^{\min(m,l)} \delta_t(A,B)(1-t).$$

Also note that from the representation (14.4) for $W_{ij}(\alpha)$ we also have the nice alternative formulas

$$W_{ij}(1) = W_{ij}, \quad \frac{\partial W_{ij}(\alpha)}{\partial \alpha}\Big|_{\alpha=1} = N_{ij}.$$

As before, an example shows that these representations are all quite explicit.

Example 14.3.1 As usual we take $\Omega = \{1,2\}$, but now we consider the compound pattern $\mathcal{A} = \{11, 121\}$. If we further assume that

$$\mathbf{P}(Z_1 = 1) = \mathbf{P}(Z_1 = 2) = 1/2,$$

then we find

$$W_{ij} = \begin{bmatrix} 6 & 2 \\ 2 & 10 \end{bmatrix},$$

$$W_{ij}(\alpha) = \begin{bmatrix} 4\alpha^{-1} + 2 & 2 \\ 2 & 8\alpha^{-2} + 2 \end{bmatrix},$$

and

$$N_{ij} = \begin{bmatrix} -4 & 0 \\ 0 & -16 \end{bmatrix}.$$

The theorems of this section then give us the concrete answers:

$$\mathbf{E}(\tau_{\mathcal{A}}) = \frac{8}{3}, \quad \text{and} \quad \mathbf{Var}(\tau_{\mathcal{A}}) = 10,$$

and

$$\mathbf{E}(\alpha^{\tau_{\mathcal{A}}}) = \frac{\alpha^2(\alpha+2)}{8-4\alpha-\alpha^3} = \frac{\alpha^2}{4} + \frac{\alpha^3}{4} + \frac{\alpha^4}{8} + \frac{3\alpha^5}{32} + \frac{5\alpha^6}{64} + \frac{7\alpha^7}{128} + o(\alpha^7).$$

14.4 Patterns in Markov Dependent Trials

Gambling teams provide a handy way to deal with many questions about sequences of independent symbols, but, when the symbols are generated by a Markov chain, one finds that the method of gambling teams is especially powerful—even though some new subtleties are introduced. For example, in the Markov case one typically needs to introduce multiple teams of gamblers who gamble according to different rules. To illustrate the basic ideas in the simplest nontrivial case, we first apply the gambling team method to the calculation of the expected time until one observes a specified pattern in a sequence generated by a two-state Markov chain.

14.4.1 Two-state Markov chains and a single pattern

In next two sections we take $\{Z_n, n \geq 1\}$ to be a Markov chain with state space $\Omega = \{1, 2\}$. We suppose the chain has the initial distribution $\mathbf{P}(Z_1 = 1) = p_1$, $\mathbf{P}(Z_1 = 2) = p_2$ and the transition matrix

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}.$$

Here, as usual, p_{ij} is shorthand for $\mathbf{P}(Z_{n+1} = j \,|\, Z_n = i)$. Given a pattern $\mathrm{A} = a_1 a_2 \cdots a_m$, we then let τ_{A} denote the first time that the pattern is observed in the sequence generated by the Markov chain.

Next, we need to make explicit the Markov version of a fair casino where a gambler who bets on the event $\{Z_{n+1} = a\}$ is assumed to have first observed Z_n. Here, if one first observes $Z_n = 1$, then the bettor of one dollar on the event $\{Z_{n+1} = a\}$ receives p_{1a}^{-1} dollars if $Z_{n+1} = a$ occurs; otherwise, the bettor receives 0. Similarly, if one first observes $Z_n = 2$ and then bets that $Z_{n+1} = a$, the payoffs are p_{2a}^{-1} and 0, respectively.

There are now three distinct scenarios under which the pattern A can be observed. Either

- the pattern A occurs at the beginning of the sequence $\{Z_n, n \geq 1\}$, or
- the pattern 1A occurs at the end of the sequence, or
- the pattern 2A occurs at the end of the sequence.

The probability of the first scenario is easy to find, but to determine the individual probabilities of the last two scenarios would be more subtle. Instead we

will use another gambling team trick to avoid such calculations. The new trick is to consider *two* gambling teams and to allow the teams to bet *differently* on the pattern A. The added flexibility will permit us to set things up so that the teams' total winnings are known if we know how the game ended.

For each time n two new gamblers are ready to take action, one from each team. The gamblers now follow two rules:

1. For each n a gambler from the first team arrives before round n and watches the result of the nth trial. He then bets y_1 dollars on the first letter of the sequence A and continues to bet his accumulated winnings on the successive letters in the successive rounds until either he loses or the patten A is observed, either by himself or by some other gambler from one of the two teams. We call the gamblers on this team *straightforward gamblers*.

2. Gamblers from the second team bet differently. If $Z_n \neq a_1$ then the nth gambler from the second team bets y_2 dollars on the round $n+1$ on the first letter of the pattern A. This gambler then continues to "let his fortune roll" until either he loses or until A is observed, either by himself or by some other gambler. On the other hand, if $Z_n = a_1$ then this gambler (intelligently!) bets y_2 dollars on a_2 on round $n+1$ and then he continues to bet on the remaining letters of the pattern $a_3 \cdots a_m$ until he loses or until the pattern A is observed by himself or by some other gambler. We call the gamblers of the second team *smart gamblers*.

Now, we let $W_{ij}y_j$, $i = 1, 2, 3$, $j = 1, 2$ be the amount of money that the jth team wins if the game ends in the ith scenario. It is vital to note that the *deterministic* quantities W_{ij} are easy to compute. The stopped martingale $X_{\tau_{\mathrm{A}}}$ that represents the net casino gain at time τ_{A} is given by

$$X_{\tau_{\mathrm{A}}} = (y_1 + y_2)(\tau_{\mathrm{A}} - 1) - \sum_{i=1}^{3} \sum_{j=1}^{2} W_{ij} y_j 1_{E_i},$$

where 1_{E_i} is the indicator of the ith ending scenario. To see this, note that no money was bet on the first round, and $y_1 + y_2$ was the amount bet by each of the first-time bettors at each of the subsequent rounds.

Now, we assume that we can find $\{y_j\}_{1 \le j \le 2}$ such that

$$\sum_{j=1}^{2} W_{ij} y_j = 1, \quad 2 \le i \le 3.$$

The existence of y_1 and y_2 depends on the computed values $\{W_{ij}\}$, but they will exist except in isolated, degenerate cases. The stopped martingale is then given by the simpler formula

$$X_{\tau_{\mathrm{A}}} = (y_1 + y_2)(\tau_{\mathrm{A}} - 1) - (W_{11}y_1 + W_{12}y_2)1_{E_1} - 1_{E_1^c},$$

where $1_{E_1^c}$ is the indicator of the complement of the first ending scenario. Taking the expectation and employing the optional stopping theorem, we obtain

$$0 = (y_1 + y_2)(\mathbf{E}(\tau_{\mathrm{A}}) - 1) - \pi_1(W_{11}y_1 + W_{12}y_2) - (1 - \pi_1),$$

where π_1 is the probability of the first scenario. As we noted earlier, it is always easy to compute π_1, so at the end of the day one just solves for $\mathbf{E}(\tau_{\mathrm{A}})$ to find

$$\mathbf{E}(\tau_{\mathrm{A}}) = 1 + \frac{\pi_1(W_{11}y_1 + W_{12}y_2) + (1 - \pi_1)}{y_1 + y_2}.$$

Example 14.4.1 To see that this is indeed an explicitly computable formula, consider the pattern A = 121. The straightforward gamblers start with a fortune of y_1 dollars and successively bet their accumulated fortunes on the successive values of 121. On the other hand, the smart gamblers start with y_2 dollars and bet their accumulated fortune on the successive values of 121 if they observed 2 before placing their first bet, but they bet their money on the successive values of 21 if they observed 1 before placing their first bet. The three scenarios are (1) the game ends with 121 at the beginning, or (2) the game ends with 2121 at the end of some indeterminate number of rounds, or (3) the game ends with 1121 at the end of some indeterminate number of rounds. The 3×2 (scenarios by teams) matrix $\{W_{ij}\}$ is then given by

$$\begin{bmatrix} & \frac{1}{p_{21}} & & \frac{1}{p_{12}p_{21}} + \frac{1}{p_{21}} \\ \\ \frac{1}{p_{21}p_{12}p_{21}} + \frac{1}{p_{21}} & & \frac{1}{p_{21}p_{12}p_{21}} + \frac{1}{p_{12}p_{21}} + \frac{1}{p_{21}} \\ \\ \frac{1}{p_{11}p_{12}p_{21}} + \frac{1}{p_{21}} & & \frac{1}{p_{12}p_{21}} + \frac{1}{p_{21}} \end{bmatrix}.$$

To determine the initial bet sizes y_1 and y_2, we then just solve the relations

$$y_1\Big(\frac{1}{p_{21}p_{12}p_{21}} + \frac{1}{p_{21}}\Big) + y_2\Big(\frac{1}{p_{21}p_{12}p_{21}} + \frac{1}{p_{12}p_{21}} + \frac{1}{p_{21}}\Big) = 1,$$

$$y_1\Big(\frac{1}{p_{11}p_{12}p_{21}} + \frac{1}{p_{21}}\Big) + y_2\Big(\frac{1}{p_{12}p_{21}} + \frac{1}{p_{21}}\Big) = 1,$$

to find

$$y_1 = \frac{p_{11}p_{12}p_{21}}{p_{12} + p_{21} + p_{12}p_{21}} \quad \text{and} \quad y_2 = \frac{p_{12}p_{21}(p_{21} - p_{11})}{p_{12} + p_{21} + p_{12}p_{21}}.$$

The probability π_1 of the first scenario is just $p_1p_{12}p_{21}$, so after substitution and simplification we obtain the pleasingly succinct formula

$$\mathbf{E}(\tau_{\mathrm{A}}) = 1 + \frac{p_2}{p_{21}} + \frac{1}{p_{21}^2} + \frac{1}{p_{12}p_{21}}.$$

14.4.2 Two-state Markov chains and compound patterns

The next natural challenge is to compute the expected value of $\tau_{\mathcal{A}}$ the first time that one observes a pattern from the set $\mathcal{A} = \{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_K\}$. The gambling teams method again applies, but one more nuance emerges. In particular, it is useful to refine the split notion of ending scenarios into *initial-ending scenarios* and *later-ending scenarios*. Specifically, we consider K *initial-ending scenarios*, where in the ith initial-ending scenario the pattern $\mathrm{A}_i, 1 \le i \le K$ occurs in the beginning of the sequence $\{Z_n, n \ge 1\}$, and we consider $2K$ *later-ending scenarios*, where either the pattern $1\mathrm{A}_i$ for some $1 \le i \le K$ occurs or else the pattern $2\mathrm{A}_i$ for some $1 \le i \le K$ occurs after some indeterminate number of rounds.

This gives us complete coverage of how one of the patterns from $\mathcal{A}$ can appear; in fact the coverage is over complete since it is possible that some of the later-ending scenarios need not be achievable as final blocks of the Markov sequence at time $\tau_{\mathcal{A}}$. For example, if $\mathcal{A} = \{212, 22\}$, then the *doubling step* formally gives us four later-ending scenarios: $\{\cdots 1212, \cdots 2212, \cdots 122, \cdots 222\}$, but 221 and 222 cannot occur as a substring of the string $Z_1, Z_2, ..., Z_{\tau_{\mathcal{A}}}$. Similarly, if the initial collection is $\mathcal{A} = \{21, 111\}$, then the only observable later-scenarios are $\{\cdots 121, \cdots 221\}$.

Thus, one typically needs to do some cleaning of the initial list of later-ending scenarios, and, if a later-ending scenario cannot be observed in a sequence that ends at time $\tau_{\mathcal{A}}$, then the scenario is eliminated from the original list of $2K$ later-ending scenarios. The *final list* of ending scenarios is then the set of initial-ending scenarios and later-ending scenarios that have not been eliminated. We let N' denote the number of later-ending scenarios in the final list.

Now we introduce N' gambling teams, one for each of the later-ending scenarios. The rule is simple. If in the final list of scenarios there are *two* later-ending scenarios associated with the pattern A_i, then we introduce *two* gambling teams. One team bets on A_i in a straightforward way, and one team bets on A_i in the smart way of the previous section. On the other hand, if in the final list we have only one later-ending scenario associated with the pattern A_i we will use only one gambling team of straightforward gamblers. Finally, if there are no later-ending scenarios in the final list associated with A_i, no gambling teams linked with A_i are needed.

We let X_n denote the casino's net gain at time n. We take $y_j, 1 \le j \le N'$ to be the initial bet with which a gambler from the jth gambling team starts his betting, and we let $W_{ij}y_j$, $1 \le i \le K$ be the total winnings of the jth gambling team in the case of the ith initial-ending scenario. Finally, we let $y_j W_{ij}$, $K + 1 \le i \le K + N'$ be the total winnings of the jth gambling team in the case when the game is ended by the ith later-ending scenario. Then the stopped martingale $X_{\tau_{\mathcal{A}}}$ is given by

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{N'} y_j(\tau_{\mathcal{A}} - 1) - \sum_{i=1}^{K}\sum_{j=1}^{N'} W_{ij}y_j 1_{E_i} - \sum_{i=K+1}^{K+N'}\sum_{j=1}^{N'} W_{ij}y_j 1_{E_i},$$

where E_i is the event that the ith scenario occurs. Again, the W_{ij} are not random, and, parallel to our earlier calculations, we assume that one can find $\{y_j\}_{1\le j\le N'}$ such that

$$\sum_{j=1}^{N'} W_{ij}y_j = 1, \text{ for all } K+1 \le i \le K+N'. \tag{14.7}$$

We then have the representation

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{N'} y_j(\tau_{\mathcal{A}} - 1) - \sum_{i=1}^{K}\sum_{j=1}^{N'} W_{ij}y_j 1_{E_i} - \sum_{i=K+1}^{K+N'} 1_{E_i},$$

so the optional stopping theorem tells us that

$$0 = \mathbf{E}(X_{\tau_{\mathcal{A}}}) = \sum_{j=1}^{N'} y_j(\mathbf{E}(\tau_{\mathcal{A}}) - 1) - \sum_{i=1}^{K}\sum_{j=1}^{N'} W_{ij}y_j\pi_i - (1 - \sum_{i=1}^{K}\pi_i),$$

where π_i is the probability that the ith initial-ending scenario occurs. Solving this equation, we obtain a slightly untidy but still completely computable formula for $\mathbf{E}(\tau_{\mathcal{A}})$.

Theorem 14.4.1 *If $\{y_j\}_{1\le j\le N'}$ solves the linear system (14.7), then*

$$\mathbf{E}(\tau_{\mathcal{A}}) = 1 + \frac{(1 - \sum_{i=1}^{K}\pi_i) + \sum_{i=1}^{K}\pi_i \sum_{j=1}^{N'} y_j W_{ij}}{\sum_{j=1}^{N'} y_j}. \tag{14.8}$$

Example 14.4.2 For the collection of patterns $\mathcal{A} = \{11, 212\}$ we find after the doubling and cleaning steps that the final list of later-ending scenarios is $\{211, 1212, 2212\}$. Together with our initial-ending scenarios, we have a total of five ending scenarios which we order as

$$\{11, 212, 211, 1212, 2212\}.$$

The scenario-by-team win matrix $\{W_{ij}\}$ is then given by

$$\begin{bmatrix} \frac{1}{p_{11}} & 0 & 0 \\ 0 & \frac{1}{p_{12}} & \frac{1}{p_{21}p_{12}} + \frac{1}{p_{12}} \\ \frac{1}{p_{21}p_{11}} + \frac{1}{p_{11}} & 0 & 0 \\ 0 & \frac{1}{p_{12}p_{21}p_{12}} + \frac{1}{p_{12}} & \frac{1}{p_{12}p_{21}p_{12}} + \frac{1}{p_{21}p_{12}} + \frac{1}{p_{12}} \\ 0 & \frac{1}{p_{22}p_{21}p_{12}} + \frac{1}{p_{12}} & \frac{1}{p_{21}p_{12}} + \frac{1}{p_{12}} \end{bmatrix},$$

and, after solving the corresponding linear system, we find that the appropriate initial team bets are given by

$$y_1 = \frac{p_{21}p_{11}}{1+p_{21}}, \quad y_2 = \frac{p_{22}p_{21}p_{12}}{p_{21}+p_{12}+p_{21}p_{12}}, \quad y_3 = \frac{p_{21}p_{12}(p_{12}-p_{22})}{p_{21}+p_{12}+p_{21}p_{12}}.$$

The probabilities π_1 and π_2 that 11 and 212 are initial segments of the process $\{Z_n, n \geq 1\}$ are given by p_1p_{11} and $p_2p_{21}p_{12}$, respectively, so the formula (14.8) leads one to the following result:

$$\mathbf{E}(\tau_{\mathcal{A}}) = 2 + p_1p_{12} + \frac{1-p_1p_{11}}{p_{21}},$$

which we see was not so complicated after all.

Finally, one should note that when a martingale method for the expected waiting time is developed, it is usually straightforward to extend the method to obtain formulas for higher moments or generating functions. We have already seen how this can be done in the independent model, and Glaz et al. (2006) give a more detailed exposition that covers the case of the two-state Markov chains.

14.4.3 Finite state Markov chains

Now consider a temporally homogeneous Markov chain $\{Z_n, n \geq 1\}$ with a finite state space $\Omega = \{1, 2, ..., M\}$, initial distribution $\mathbf{P}(Z_1 = m) = p_m$, $1 \leq m \leq M$, and transition matrix $P = \{p_{ij}\}_{1 \leq i,j \leq M}$, where as always,

$$p_{ij} = \mathbf{P}(Z_{n+1} = j | Z_n = i).$$

We let $\mathcal{A} = \{\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_K\}$ denote a compound pattern, and let

$$\tau_{\mathcal{A}} = \min\{\tau_{\mathrm{A}_1}, ..., \tau_{\mathrm{A}_K}\}$$

denote the first time when we observe a pattern from $\mathcal{A}$ in the Markov sequence. We also assume that the Markov chain has the following normalization and regularization properties:

- We assume that no pattern of $\mathcal{A}$ contains another pattern of $\mathcal{A}$ as a subpattern. This property holds without loss of generality, since if one pattern is a subpattern of another, the longer one can be excluded from our list.

- We assume that $\mathbf{P}(\tau_{\mathcal{A}} = \tau_{\mathrm{A}_i}) > 0$ for all $1 \leq i \leq K$. If, on the contrary, one were to have $\mathbf{P}(\tau = \tau_{\mathrm{A}_i}) = 0$ for some i, then A_i could simply be

excluded from the list. This possibility is excluded by the first assumption for independent sequences, but for Markov sequences it often needs our attention. For example, if the pattern A_i contains subpattern km and $p_{km} = 0$, then A_i cannot happen as a run of $\{Z_n, n \geq 1\}$.

- We assume that $\mathbf{P}(\tau_{\mathcal{A}} < \infty) = 1$. If the patterns of $\mathcal{A}$ all contain transient states, this condition can easily fail even for a finite Markov chain. Here we should note that for finite Markov chains the basic finiteness condition $\mathbf{P}(\tau_{\mathcal{A}} < \infty) = 1$ already implies the formally stronger condition $\mathbf{E}[\tau_{\mathcal{A}}] < \infty$.

The Multi-state Chain Martingale Construction

When $M = |\Omega| > 2$ the critical martingales require a more elaborate description. We begin by decomposing the possible occurrence of a single pattern A_i into an *initial list* of $1 + M + M^2$ ending scenarios:

- Either the sequence A_i occurs as an initial segment of $\{Z_n, n \geq 1\}$, or
- for some $1 \leq k \leq M$, the pattern kA_i occurs as an initial segment of the sequence $\{Z_n, n \geq 1\}$, or
- for some pair (k, m), $1 \leq k, m \leq M$, the pattern kmA_i occurs after some indeterminant number of rounds.

The first $1 + M$ ending scenarios are called *initial* scenarios. The last M^2 scenarios are called *later* scenarios. Since we have K patterns, we have an initial list of $(1 + M + M^2)K$ scenarios.

For every later scenario associated with the pattern $km\mathrm{A}_i$ we introduce a team of gamblers that we call the *$km\mathrm{A}_i$-gambling team*. Gambler $n + 1$ from the $km\mathrm{A}_i$-gambling team arrives before round $n+1$ to observe the result of the nth trial, Z_n.

This gambler then starts his betting. If $Z_n = k$ he bets a certain amount of money (which is the same for all gamblers from the $km\mathrm{A}_i$-gambling team) on the pattern $m\mathrm{A}_i$. If $Z_n \neq k$ he bets on A_i. Here, of course, by "betting \$1 on the pattern $\mathrm{A} = a_1 a_2 \cdots a_m$, when $Z_n = a_0$" we mean the following:

- After observing Z_n the gambler bets a dollar that the next trial yields a_1. If $Z_{n+1} \neq a_1$, he loses his dollar and leaves the game. If $Z_{n+1} = a_1$, he gets $1/p_{a_0 a_1}$. Note that the odds are fair. If he wins he continues his betting.
- Now he bets his entire capital that the $n+2$ round yields a_2. If it is a_2 he increases his capital by factor $1/p_{a_1 a_2}$; otherwise, he leaves the game with nothing. He continues to bet his full fortune on the successive letters of the pattern A until either the pattern A is observed, or until some other gambler has succeeded.

Now recall that it is possible that some of the scenarios on our initial list simply cannot occur before the waiting time $\tau_{\mathcal{A}}$. Moreover, some ending scenarios are impossible simply because some new patterns associated with some ending scenarios cannot be observed at all in the Markov chain. Thus, we need to clean the initial list of ending scenarios.

Those scenarios that cannot occur at all and those that can occur only after the time τ_{A} must be eliminated. Let K' denote the number of initial scenarios, and let N' denote the number of later scenarios that we have in our list after cleaning. For each jth later scenario in the new list, we introduce the corresponding gambling team, and we assume that the inial amount with which the gamblers of the jth team start their betting is y_j. The values $\{y_j\}$ will be chosen later.

Let $y_j W_{ij}$, $1 \le i \le K' + N'$, $1 \le j \le N'$ be the amount of money that the jth team wins in the ith ending scenario. Let X_n denote the casino's net gain from all teams at time n. The sequence $\{X_n\}$ forms a martingale with respect to the filtration generated by the Markov chain $\{Z_n, n \ge 1\}$. Indeed, for every gambler in the game the bet size at a current round is fully determined by previous rounds, and odds—as we have seen—are fair. By bookkeeping, one finds for the stopped martingale $X_{\tau_{\mathcal{A}}}$ that

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{N'} y_j(\tau_{\mathcal{A}} - 1) - \sum_{i=1}^{K'} \sum_{j=1}^{N'} W_{ij} y_j 1_{E_i} - \sum_{i=K'+1}^{K'+N'} \sum_{j=1}^{N'} W_{ij} y_j 1_{E_i},$$

where E_i is the event that the ith scenario occurs. Here, again W_{ij} is not a random variable; it depends only on the overlap properties of the pattern associated with the ith scenario and the pattern associated with the jth gambling team.

If we now assume that we can find $\{y_j\}_{1 \le j \le N'}$ such that

$$\sum_{j=1}^{N'} W_{ij} y_j = 1, \text{ for all } K' + 1 \le i \le K' + N', \tag{14.9}$$

then $X_{\tau_{\mathcal{A}}}$ has the more tractable representation

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{N'} y_j(\tau_{\mathcal{A}} - 1) - \sum_{i=1}^{K'} \sum_{j=1}^{N'} W_{ij} y_j 1_{E_i} - \sum_{i=K'+1}^{K'+N'} 1_{E_i}.$$

Since $\{X_n\}_{n \ge 1}$ has bounded increments and $\mathbf{E}[\tau_{\mathcal{A}}] < \infty$, the Doob's optional stopping theorem gives us

$$0 = \mathbf{E}(X_{\tau_{\mathcal{A}}}) = \sum_{j=1}^{N'} y_j(\mathbf{E}(\tau_{\mathcal{A}}) - 1) - \sum_{i=1}^{K'} \sum_{j=1}^{N'} W_{ij} y_j \pi_i - \left(1 - \sum_{i=1}^{K'} \pi_i\right),$$

where π_i is the probability that the ith initial scenario occurs. Solving the equation with respect to $\mathbf{E}(\tau_{\mathcal{A}})$ we obtain the main result of this section.

Theorem 14.4.2 *If* $\{y_j\}_{1\le j\le N'}$ *solves the linear system (14.9), then*

$$\mathbf{E}(\tau_{\mathcal{A}}) = 1 + \frac{(1-\sum_{i=1}^{K'}\pi_i) + \sum_{i=1}^{K'}\pi_i \sum_{j=1}^{N'} y_j W_{ij}}{\sum_{j=1}^{N'} y_j}. \tag{14.10}$$

Example 14.4.3 Let $\Omega = \{1,2,3\}$ and $\mathcal{A} = \{323, 313, 33\}$. Let the initial distribution be given by

$$p_1 = 1/3, \quad p_2 = 1/3, \quad p_3 = 1/3,$$

and let the transition matrix P be given by

$$P = \begin{bmatrix} 3/4 & 0 & 1/4 \\ 0 & 3/4 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}.$$

After eliminating the impossible scenarios we get 9 initial scenarios:

$$323\cdots, 313\cdots, 33\cdots, 1323\cdots, 2323\cdots, 1313\cdots, 2313\cdots, 133\cdots, 233\cdots$$

and because transitions $1 \to 2$ and $2 \to 1$ are impossible we get just six later scenarios:

$$\cdots 11323, \cdots 22323, \cdots 11313, \cdots 22313, \cdots 1133, \cdots 2233.$$

Now we need to calculate the matrix W and we first consider some sample entries. For instance, the 11323-gambling team in the initial scenario $323\cdots$ wins $1/p_{23} = 4$. The same team in the later scenario $\cdots 11323$ wins $1/(p_{11}p_{13}p_{32}p_{23}) + 1/p_{23} = 268/3$, and in the later scenario $\cdots 22323$ it wins $1/(p_{23}p_{32}p_{23}) + 1/p_{23} = 68$. Finally, the entries of matrix W that correspond to the later scenarios—the ones that are needed for linear system (14.9)—are given by

$$\begin{bmatrix} 268/3 & 64 & 4 & 0 & 4 & 0 \\ 68 & 256/3 & 4 & 0 & 4 & 0 \\ 0 & 4 & 256/3 & 68 & 0 & 4 \\ 0 & 4 & 64 & 268/3 & 0 & 4 \\ 2 & 2 & 2 & 2 & 38/3 & 10 \\ 2 & 2 & 2 & 2 & 10 & 38/3 \end{bmatrix}.$$

Finally, from formula (14.10) we have the bottom line:

$$\mathbf{E}(\tau_{\mathcal{A}}) = 8\frac{7}{15}.$$

Higher Moments, the Generating Functions, and Efficiency

In parallel with our earlier examples, one can now take initial bets of size $y_j + z_j n$ to obtain a formula for the second moment, or take initial bets of size $y_j \alpha^n$ to obtain the corresponding generating function, see Pozdnyakov (2008).

Here we should note that while the method of this subsection is also applicable to two-state Markov chains, it is certainly less efficient than the one given in Subsection 14.4.2. Here, in the case of two-state Markov chains we would have $4K$ ending scenarios, but the method of Subsection 14.4.2 needs only $2K$.

Finally, one should note some of the computational differences between the martingale technique and the Markov chain imbedding method. To find the expected time $\mathbf{E}(\tau_{\mathcal{A}})$ via an appropriate Markov chain imbedding, one needs to solve a linear system associated with the transition matrix of the imbedded Markov chain; see Fu and Chang (2002, p. 73). The size of the matrix depends on the cardinality K of the compound pattern $\mathcal{A}$ and the lengths of single patterns in $\mathcal{A}$. Our matrix depends on K and the cardinality M of the alphabet. Thus, there are situations when the martingale approach is computationally more effective. For a very simple example, one can take $\mathcal{A}$ to consist of just one very long pattern.

14.5 Applications to Scans

In its simplest form [see Naus (1965)], the scan statistic is the largest number of "events that occur" in a window of a given fixed length when we scan the window over a realization of a temporally homogeneous process up to a specified terminal time. For a concrete example, consider a sequence of independent Bernoulli trials $\{Z_n, n \geq 1\}$ with

$$\mathbf{P}(Z_i = 1) = p = 1 - \mathbf{P}(Z_i = 0).$$

Now given $1 \leq w \leq T$ and $1 \leq i \leq T - w + 1$, we consider the sums

$$Y_{i,w} = \sum_{j=i}^{i+w-1} Z_j,$$

and we define the *scan statistic* $S_{w,T}$ to be the maximum of $Y_{i,w}$; that is,

$$S_{w,T} = \max_{1 \leq i \leq T-w+1} Y_{i,w}.$$

If $\tau_{k,w}$ denotes the first time when one first observes at least k occurrences of the value 1 in a window of length w, then $\tau_{k,w}$ is related to the scan statistic by

$$\mathbf{P}(S_{w,T} \geq k) = \mathbf{P}(\tau_{k,w} \leq T).$$

For us, the key observation is that the waiting time $\tau_{k,w}$ can be viewed as the waiting time $\tau_{\mathcal{A}}$ for an appropriate compound pattern $\mathcal{A}$. For example, for $k = 3$ and $w = 5$ the compound pattern $\mathcal{A}$ is given by

$$\{111, 1101, 1011, 11001, 10101, 10011\}.$$

The bottom line is that knowledge of the distribution of $\tau_{\mathcal{A}}$ gives us the distribution of the associated scan statistics. Moreover, this method of association goes well beyond the simple scan of this example. Analogous transformations permit one to treat the variable window scans of Glaz and Zhang (2006) or the double scans considered by Naus and Stefanov (2002) and Naus and Wartenberg (1997).

14.5.1 Second moments and distribution approximations

Since martingale methods yield effective computations of the moments of the waiting time $\tau_{\mathcal{A}}$, it is natural to ask if martingale methods also suggest *approximations* of the distribution of $\tau_{\mathcal{A}}$ that use the first two (or perhaps more) moments of the waiting time.

It is reasonable from the clumping heuristic that the stopping time $\tau_{\mathcal{A}}$ that one associates with a scan statistic should have tail probabilities $\mathbf{P}(\tau_{\mathcal{A}} \leq n)$ that are close to those of the exponential distribution. Still, when one considers the whole distribution, there are natural competitors to the exponential such as the gamma, the Weibull, and the shifted exponentials. The main finding in Pozdnyakov *et al.* (2005) was that in many natural situations it is the class of shifted exponential distributions that provides the most accurate approximation to the distribution of $\tau_{\mathcal{A}}$.

To make this approximation explicit, we first recall that X' is called a shifted exponential, provided that $X' = X + c$ where X has an exponential distribution. We take X' as our moment matching approximation to $\tau_{\mathcal{A}}$, provided c is chosen so that

$$\mathbf{E}(X + c) = \mathbf{E}(\tau_{\mathcal{A}}), \quad \mathbf{Var}(X + c) = \mathbf{Var}(\tau_{\mathcal{A}}).$$

For the tail probabilities this approximation gives us the relation

$$\mathbf{P}(\tau_{\mathcal{A}} \leq n) \approx 1 - \exp(-(n + 0.5 + \sigma - \mu)/\sigma), \tag{14.11}$$

where $\mu = \mathbf{E}(\tau_{\mathcal{A}})$, $\sigma = \mathbf{Var}(\tau_{\mathcal{A}})$, and the 0.5 term provides a continuity correction. As the following examples demonstrate, this approximation works remarkably well for a wide variety of scan statistics.

Example 14.5.1 (*Fixed window scans*). Here $\{Z_n, n \geq 1\}$ is a sequence of Bernoulli trials. We consider two scans: at-least-3-out-of-10 (Table 14.1) and at-least-4-out-of-20 (Table 14.2).

Table 14.1. Fixed window scans: at least 3 failures out of 10 consecutive trials, $\mathbf{P}(Z_n = 1) = .01$, $\mu = 30822$, $\sigma = 30815$.

n	exponential	shifted exponential	gamma	upper bound	lower bound
500	0.01600	0.01589	0.01597	0.01588	0.01589
1000	0.03183	0.03173	0.03179	0.03171	0.03174
1500	0.04741	0.04731	0.04736	0.04729	0.04733
2000	0.06274	0.06265	0.06267	0.06262	0.06267
2500	0.07782	0.07773	0.07775	0.07770	0.07776
3000	0.09266	0.09258	0.09258	0.09254	0.09261
4000	0.12162	0.12155	0.12154	0.12150	0.12169
5000	0.14966	0.14960	0.14957	0.14954	0.14965

Table 14.2. Fixed window scans: at least 4 failures out of 20 consecutive trials, $\mathbf{P}(Z_n = 1) = .05$, $\mu = 481.59$, $\sigma = 469.35$.

n	exponential	shifted exponential	gamma	upper bound	lower bound
50	0.09110	0.07827	0.08268	0.07713	0.07940
60	0.10977	0.09770	0.10059	0.09543	0.09989
70	0.12807	0.11672	0.11828	0.11337	0.11991
80	0.14599	0.13534	0.13573	0.13095	0.13949
90	0.16354	0.15357	0.15292	0.14819	0.15864
100	0.18073	0.17141	0.16985	0.16508	0.17736

For the fixed window scan statistics, Glaz and Naus (1991) developed tight lower and upper bounds which are provided in Tables 14.1 and 14.2 along with the approximations based on the exponential, shifted exponential, and gamma distributions. The Weibull distribution-based approximation is omitted, because the performances of Weibull approximations are significantly worse than those of the exponential and the gamma. As can be seen, the shifted exponential approximation does consistently well. In the easy case when μ is large and σ is close to μ, the differences between the various approximations are marginal, and all of the estimates are close to the true probability. On the other hand, if μ is relatively small and σ differs from μ, then the approximations based on the exponential and gamma distributions do not perform nearly as well as the shifted exponential approximations.

Example 14.5.2 (*Variable window scans*). Again we let $\{Z_n, n \geq 1\}$ be a sequence of Bernoulli trials, but this time we scan for the occurrence of either of two situations: either we observe at least 2 failures in 10 consecutive trials,

Table 14.3. Variable window: at least 2 failures out of 10 trials or at least 3 failures out of 50 trials, $\mathbf{P}(Z_n = 1) = .01$, $\mu = 795.33$, $\sigma = 785.85$.

n	exponential	shifted exponential	gamma	simulated $N = 100{,}000$
50	0.05857	0.05085	0.05542	0.05029
60	0.07033	0.06285	0.06685	0.06187
70	0.08195	0.07470	0.07817	0.07404
80	0.09342	0.08640	0.08939	0.08623
90	0.10474	0.09796	0.10050	0.09718
100	0.11593	0.10936	0.11150	0.11058

or we observe at least 3 failures in 50 consecutive trials. Here are interested in the approximation for the distribution of the waiting time τ until one of these two situations occurs. In this case we need a compound pattern $\mathcal{A}$ with 224 patterns in order for τ and $\tau_{\mathcal{A}}$ to have the same distribution.

The numerical results are given in Table 14.3. Since analytical bounds for this type of scan are not available, the performance of the approximation is judged by comparison with estimated probabilities based on $100,000$ replications. Here, again, we see that the shifted exponential distribution approximation that is calibrated by two moments performs quite well.

Example 14.5.3 (*Double scans*). Let $\{Z_n, n \geq 1\}$ be an i.i.d. sequence of random variables with the three-valued distribution specified by

$$\mathbf{P}(Z_n = 1) = .04, \quad \mathbf{P}(Z_n = 2) = .01, \quad \text{and} \quad \mathbf{P}(Z_n = 0).$$

Now we consider two types of "failures"; a type I failure corresponds to observing a 1 and a type II failure corresponds to observing a 2. Further, we assume that we scan with a window of length 10 until we observe at least 2 failures of type II or observe at least 3 failures (of any combination of kinds). Table 14.4 shows that the shifted exponential approximation works well even when μ and σ are relatively small and significantly different.

The initial arguments of Pozdnyakov *et al.* (2005) in favor of the shifted exponential approximation were predominantly empirical, but subsequently a more theoretical motivation has emerged from the work of Fu and Lou (2006, p. 307), which shows that for large n one has

$$\mathbf{P}(\tau_{\mathcal{A}} \geq n) \sim C^* \exp(-n\beta),$$

where the constants C^* and β are defined in terms of the largest eigenvalue (and corresponding eigenvector) of what Fu and Lou (2006) call the *essential transition probability matrix* of the imbedded finite Markov chain associated

Table 14.4. Double scans: at least 2 type II failures out of 10 trials or at least 3 failures of any kind out of 10 trials, $\mathbf{P}(Z_n = 1) = .04$, $\mathbf{P}(Z_n = 2) = .01$, $\mu = 324.09$, $\sigma = 318.34$.

n	exponential	shifted exponential	gamma	simulated $N = 100{,}000$
10	0.02438	0.01480	0.02175	0.01401
15	0.03932	0.03015	0.03568	0.03084
20	0.05403	0.04527	0.04959	0.04508
25	0.06851	0.06015	0.06342	0.06169
30	0.08277	0.07479	0.07714	0.07590
35	0.09681	0.08921	0.09074	0.09134
40	0.11064	0.10340	0.10419	0.10529
45	0.12425	0.11738	0.11749	0.11878
50	0.13766	0.13113	0.13063	0.13342

with compound pattern $\mathcal{A}$. One should note that this matrix is not a proper transition matrix; rather it is a restriction of a transition matrix.

Now, if we omit the continuity factor correction in our shifted exponential approximation (14.11), we have an approximation of exactly the same form:

$$\mathbf{P}(\tau_{\mathcal{A}} \geq n) \approx \exp(-(n + \sigma - \mu)/\sigma) = \exp((\mu - \sigma)/\sigma) \exp(-n/\sigma).$$

These relations suggest that there is a strong connection between the largest eigenvalue of the essential transition matrix of the imbedded Markov chain and the first and second moments of $\tau_{\mathcal{A}}$. In particular, we conjecture that (in the typical case at least) the largest eigenvalue $\lambda_{[1]}$ of the essential transition probability matrix of the imbedded finite Markov chain associated with compound pattern $\mathcal{A}$ will satisfy the approximation

$$\lambda_{[1]} \approx \exp(-1/\sigma). \tag{14.12}$$

14.5.2 Scan for clusters of a certain word

Let $\{Z_n, n \geq 1\}$ be a sequence of i.i.d. random variables that takes values over the alphabet $\Omega = \{1, 2, ..., M\}$ and let the distribution be given by

$$p_1 = \mathbf{P}(Z_n = 1) > 0,\ p_2 = \mathbf{P}(Z_n = 2) > 0,\ ...\ , p_M = \mathbf{P}(Z_n = M) > 0.$$

Given a pattern $\mathrm{A} = a_1 a_2 \cdots a_m$ over the alphabet Ω, we then take a window of length $w \geq m$ and scan the sequence until the time τ when in the window of width w we have k (possibly overlapping) occurrences of pattern A.

One can show that τ is equal to the waiting time until the occurrence of a certain compound pattern $\mathcal{A}$, so formally the moments of τ follow from our

previous results. Unfortunately, this approach runs into computational problems since the cardinality of compound pattern $\mathcal{A}$ grows exponentially as the window width w increases. There seems to be no way to circumvent this problem entirely, but given that $\mathcal{A}$ can be computed, we can greatly cut down on much of the other work.

A New Betting Scheme

The basic idea is to bet only on the pattern A and to *pause* the betting between nonoverlapping occurrences of A. To make this explicit, we first take $\mathcal{A}$ as given and consider a certain equivalence relation on the patterns from $\mathcal{A}$. Specifically, we say that elements A_i and A_j from $\mathcal{A}$ are *similar* provided that

- the lengths of A_i and A_j are the same,
- A_i and A_j have the same number of overlapping occurrences of A and,
- the patterns A_i and A_j have copies of A's at the same positions.

Now, to each equivalence class under this relation, we can associate a unique pattern over the *extended* alphabet $\bar{\Omega} = \{1, 2, ..., M, *\}$ by a simple rule. If the simple pattern $\mathrm{A}_i \in \mathcal{A}$ is a representative of an equivalence class, then to construct what we will call the "star pattern" for the class, we replace each symbol of A_i that is not part of a block equal to A by the symbol $*$. This recipe is made clear with an example.

Example 14.5.4 Let $\Omega = \{1, 2, 3\}$ and let A = 121. Suppose we want to scan until we find the occurrence of at least two copies of A's in a window of 8 symbols. The compound pattern $\mathcal{A}$ associated with this scan consist of 11 simple patterns, none of which is a subpattern of another):

1. exactly-2-in-5: 12121,
2. exactly-2-in-6: 121121,
3. exactly-2-in-7: 1211121 and 1213121,
4. exactly-2-in-8: 12111121, 12122121, 12113121, 12123121, 12131121, 12132121, and 12133121.

Now, although we have 11 simple patterns in $\mathcal{A}$, we have only 4 equivalence classes, which we can enumerate with their star patterns:

$$12121, 121121, 121 * 121, 121 * *121.$$

Here, it is important to note that $*$ does not mean just "any symbol", because, for example, 12121121 is not in $\mathcal{A}$, and, as a result, class $121 * *121$ does not include 12121121.

Given this reduction to equivalence classes, there are analogous reductions for the rest of our tools, such as the ending scenarios. Now, we introduce a list of ending scenarios associated with the list of equivalence classes (or star patterns). As before, we associate a gambling team with each element of the final list of ending scenarios.

The real key is the new betting rule. Now, a gambler from a gambling team associated with a star pattern bets on a symbol if it is a symbol from Ω, but he simply passes when it is a star. For example, a gambler from the gambling team that corresponds to $121**121$ first bets on 121 in the sequential fashion that should now be quite familiar. If he is successful after those three bets, he then pauses for two rounds. After the pause he bets then successively bets his entire capital on 121, the rest of the star pattern.

Assume that we have N' ending scenarios and N' gambling teams. A gambler from the jth gambling team that joins the game in the nth round will bet y_j dollars. Next let $y_j W_{ij}$, $1 \le i, j \le N'$ be the total winnings of the jth gambling team in the case that the game was ended by the ith scenario. As before, W_{ij} is not random; it is fully determined by the pattern of overlap of the star patterns associated with the given gambling team and ending scenario.

To make this explicit, we let $\mathrm{E} = e_1 e_2 \cdots e_m$ and $\mathrm{T} = t_1 t_2 \cdots t_l$ be two patterns over the extended alphabet $\bar{\Omega}$. We first define a measure of "two letters coincidence":

$$\delta(e_i, t_j) = \begin{cases} 1, & \text{if } t_j = * \\ 1/\mathbf{P}(Z_1 = t_j), & \text{if } t_j \neq *, e_i = t_j, \\ 0, & \text{if } t_j \neq *, t_j \neq e_i. \end{cases}$$

Next, we define a general measure of overlap for E and T:

$$W(\mathrm{E}, \mathrm{T}) = \sum_{i=1}^{\min(m,l)} \prod_{j=1}^{i} \delta(e_{m-i+j}, e_j).$$

Finally, if the ith ending scenario is associated with the pattern E and the jth gambling team bets on T, then we have the explicit (and deterministic) formula

$$W_{ij} = W(\mathrm{E}, \mathrm{T}).$$

If X_n is the casino's total net gain at the end of the nth round, then it is again a martingale, since taking a pause preserves a martingale property. At time $\tau_{\mathcal{A}}$ the stopped martingale is given by

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{N'} y_j \tau_{\mathcal{A}} - \sum_{i=1}^{N'} \sum_{j=1}^{N'} W_{ij} y_j 1_{E_i},$$

where 1_{E_i} is the indicator that the game is ended by the ith scenario, so if the vector $\{y_j\}_{1\leq j\leq N'}$ is a solution of the linear system

$$\sum_{j=1}^{N'} W_{ij}y_j = 1, \quad 1 \leq i \leq N', \tag{14.13}$$

then the stopped martingale has the tidy representation

$$X_{\tau_{\mathcal{A}}} = \sum_{j=1}^{N'} y_j \tau_{\mathcal{A}} - 1.$$

Since $\mathbf{E}(X_{\tau_{\mathcal{A}}})$, we come very quickly to our final formula.

Theorem 14.5.1 *If the vector $\{y_j\}_{1\leq j\leq N'}$ solves the linear system (14.13), then the expected value of $\tau_{\mathcal{A}}$ is given by*

$$\mathbf{E}(\tau_{\mathcal{A}}) = \frac{1}{\sum_{j=1}^{N'} y_j}.$$

Example 14.5.5 Let $\Omega = \{1, 2, 3\}$, and A = 121, and suppose we scan for at least two A's in a window of 8 symbols. As we have seen, there are only 4 equivalence classes:

$$12121, 121121, 121 * 121, 121 * *121.$$

The matrix W_{ij} in this case is

$$\begin{bmatrix} \frac{1}{p_1^3p_2^2} + \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^3p_2} + \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{2}{p_1^2p_2} + \frac{1}{p_1} \\ \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^4p_2^2} + \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^3p_2} + \frac{1}{p_1^2p_2} + \frac{1}{p_1} \\ \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^4p_2^2} + \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^2p_2} + \frac{1}{p_1} \\ \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^2p_2} + \frac{1}{p_1} & \frac{1}{p_1^4p_2^2} + \frac{1}{p_1^2p_2} + \frac{1}{p_1} \end{bmatrix}.$$

Theorem 14.5.1 gives us the following formula for the expected value:

$$\mathbf{E}(\tau_{\mathcal{A}}) = \frac{1 + p_1p_2(1 + p_1p_2)(1 + p_1(3 - p_2 - 2p_1p_2))}{p_1^3p_2^2(1 + p_1(3 - p_2 - 2p_1p_2))}.$$

One obviously can extend this technique to the case of the higher moments and generating function.

14.6 Concluding Remarks

The martingale method for studying the waiting time for a compound pattern is now well developed—even the stubborn Markovian case. Still, from the examples given here, one can see that successful application of the method requires some detailed combinatorial information. Specifically, one almost always needs to determine explicitly what we have called here the "final list of ending scenarios."

For problems, such as those that come from the theory of scan statistics, this final list can be large. Nevertheless, by the introduction of appropriate equivalence classes, one can still make steady progress. Explicit formulas for moments are possible more often than one might guess.

There are two problems that we believe deserve consideration: one general and one specific. The general problem is the identification of further problems like the one developed in Subsection 14.5.2 for clusters of words. Generically, the challenge is to identify the problems in which one can find a substantial simplification of what would otherwise be the waiting time problem for a very large class of patterns. Correspondingly, it would be useful to identify as many problems as possible where one has a firm combinatorial understanding of the final list of ending scenarios.

The more specific problem is the conjecture given in Equation (14.12). Historically, there has been considerable value in finding a good representation for the largest eigenvalue for even very special matrices. The class of matrices that are obtained as the essential (improper) transition probability matrix of the imbedded finite Markov chain associated with compound pattern $\mathcal{A}$ is indeed special, yet it is still reasonably large. For this class the conjecture (14.12) provides an explicit—and novel—approach to the analysis of the largest eigenvalue.

References

1. Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*, Springer Publishing, New York.

2. Fu, J.C. and Chang, Y. (2002). On probability generating functions for waiting time distribution of compound patterns in a sequence of multistate trials, *Journal of Applied Probability*, **39**, 70–80.

3. Fu, J.C. and Lou, W.Y.W. (2006). Waiting time distributions of simple and compound patterns in a sequence of r-th order Markov dependent multi-state trials, *Annals of the Institute of Statistical Mathematics*, **58**, 291–310.

4. Gerber, H. and Li, S. (1981). The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain, *Stochastic Processes and Their Applications*, **11**, 101–108.

5. Glaz, J., Kulldorff, M., Pozdnyakov, V. and Steele, J.M. (2006). Gambling teams and waiting times for patterns in two-state Markov chains, *Journal of Applied Probability*, **43**, 127–140.

6. Glaz, J. and Naus, J.I. (1991). Tight bounds for scan statistics probabilities for discrete data, *Annals of Applied Probability*, **1**, 306–318.

7. Glaz, J., Naus, J.I. and Wallenstein, S. (2001). *Scan Statistics*, Springer, New York.

8. Glaz, J. and Zhang, Z. (2006). Maximum scan score-type statistics, *Statistics and Probability Letters*, **76**, 1316–1322.

9. Li, S. (1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments, *The Annals of Probability*, **8**, 1171–1176.

10. Naus, J.I. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of The American Statistical Association*, **60**, 532–538.

11. Naus, J.I. and Stefanov, V.T. (2002). Double-scan statistics, *Methodology and Computing in Applied Probability*, **4**, 163–180.

12. Naus, J.I. and Wartenberg, D. A. (1997). A double-scan statistic for clusters of two types of events, *Journal of The American Statistical Association*, **92**, 1105–1113.

13. Pozdnyakov, V. (2008). On occurrence of patterns in Markov chains: method of gambling teams, to appear in *Statistics and Probability Letters*.

14. Pozdnyakov, V., Glaz, J., Kulldorff, M. and Steele, J.M. (2005). A martingale approach to scan statistics, *Annals of The Institute of Statistical Mathematics*, **57**, 21–37.

15. Pozdnyakov, V. and Kulldorff, M. (2006). Waiting times for patterns and a method of gambling teams, *The American Mathematical Monthly*, **113**, 134–143.

16. Shiryaev, A.N. (1995). *Probability*, Springer, New York.

17. Williams, D. (1991). *Probability with Martingales*, Cambridge University Press, Cambridge.

15

How Can Pattern Statistics Be Useful for DNA Motif Discovery?

Sophie Schbath[1] **and Stéphane Robin**[2]

[1] *INRA, UR1077 Mathématique, Informatique et Génome, F-78350, Jouy-en-Josas, France*
[2] *AgroParisTech/INRA, UMR518 Mathématique et Informatique Appliquées, F-75231, Paris, France*

Abstract: Statistics of motifs have been widely revisited in the last 15 years due to the increasing availability of genomic sequences. The identification of DNA motifs with biological functions is still a huge challenge of genome analysis. Many functional and essential motifs have the particularity to be very frequent all along the chromosome or to be concentrated in some particular regions (e.g. in front of genes) or to be co-oriented with the replication direction. The prediction of functional motifs is then mostly based on statistical properties of pattern occurrences in Markovian sequences. This chapter is primarily devoted to such properties with a special focus on pattern frequency. How does one compute or approximate the count distribution to assess motif exceptionality? How can we test if a motif is significantly unbalanced between two (sets of) sequences? How should one deal with degenerated patterns? How can we model occurrences to find regions significantly enriched with a given pattern? Examples of functional motifs will illustrate all these questions, and we will see how the Chi motif has been identified in *Staphylococcus aureus* because of its statistical properties.

Keywords and phrases: Pattern statistics, word count, Markov chain, DNA sequence, exceptional words, unexpected frequency, compound Poisson process

15.1 Introduction

For the last 15 years, genomic sequence analysis has probably offered the widest variety of problems on pattern statistics. This variety is due to the huge length of the sequences and to their heterogeneous composition and structure, but also

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_15,

to the complexity of the functional motifs. These motifs take place in fundamental molecular processes like chromosome maintenance or gene transcription, but few of them have been completely identified (i.e. their sequence of letters is known). Moreover, they are rarely conserved through species, leading to a very challenging area of DNA motif discovery. This chapter is related to the statistical approach used to predict candidate functional motifs. Indeed, many known functional motifs are characterized by an exceptional behavior of their occurrences. Some of them are extremely frequent along the entire genome (or along a particular DNA strand), others are avoided because their occurrences are lethal for the chromosome, and some are preferred in particular genomic regions. Thus, two main quantities have been widely studied from a probabilistic and statistical point of view: the number of occurrences of a motif in a random sequence and the distances (cumulated or not) between occurrences of a motif. To avoid a huge list of references, we recommend Chapters 6 and 7 from Lothaire (2005) for technical expositions and Robin *et al.* (2005) for a more applied exposition. In this chapter, we have chosen to present the main statistical results that are really used in practice to help identify functional DNA motifs. Many biological examples will then be given to illustrate the usefulness of the approaches. Most will be devoted to the question of detecting words with an exceptional frequency in a given sequence. Distribution of a word count in Markovian sequences will be studied in Section 15.2. We will also consider the related problem of comparing the exceptionality of a word frequency between two independent sequences. Functional motifs can indeed be specific from known parts of the chromosome (or from some particular chromosomes). In this case, the word occurrences themselves are modeled and a statistical test is derived from the two count processes (Section 15.3). However, when one look for regions significantly enriched with (or devoid of) a given word, the quantity of interest becomes the distance between occurrences. Section 15.3 also presents results on the distance distribution when the occurrences are modeled by a compound Poisson process. Other results on distances and waiting times can be found in Stefanov (2009) when the sequence is Markovian. Section 15.4 addresses the generalization to more complex patterns, namely degenerated patterns and structured motifs. Finally, we end with some ongoing works and open problems.

15.2 Words with Exceptional Frequency

Many functional DNA motifs are extremely over-represented in complete genomes, or in specific genomic regions, whichever compositional level of the biological sequence one takes into account. This statistical property reveals a strong constraint on the DNA sequence. For instance, if we look for the two

Table 15.1. Expected counts of `aagtgcggt` and `accgcactt` in random sequences having on average the same composition as the *H. influenzae* complete genome.

Markov model	fitted composition	expected count of `aagtgcggt`	expected count of `accgcactt`
M0	letters	4.694	3.779
M1	2-letter words	6.279	4.847
M2	3-letter words	8.603	6.208
M3	4-letter words	18.601	15.080
M4	5-letter words	55.704	48.658
M5	6-letter words	219.081	220.284
M6	7-letter words	549.815	574.734
M7	8-letter words	719.440	722.366

most over-represented 9-letter words in the complete genome of the bacteria *Haemophilus influenzae* (1830140 letters long), we find the two reverse complementary oligonucleotides `aagtgcggt` and `accgcactt` which occur respectively 740 and 731 times. As an illustration, Table 15.1 gives the expected count of these two words when fitting the sequence composition of smaller words. These two 9-letter words are very well known from the biologists: they are the two DNA *uptake* sequences involved in discriminating self from foreign entering DNA during competence in the bacteria.

Another example is the word `gctggtgg` which is the "crossover hotspot instigator" (*Chi*) motif in the bacteria *Escherichia coli* and is involved in chromosome maintenance. Chi is among the five most over-represented 8-letter words in the *E. coli* genome (4638858 letters long). This example will be detailed in Section 15.2.5.

In contrast, many restriction sites (generally 6-letter words) are strongly under-represented along bacterial genomes, which is not surprising because they induce a double-strand break of the bacterial DNA. The aim of this section is precisely to show how to assess the significance of over- and under-representations.

When we want to analyze the distribution of a word along a sequence or when we want to know if a word occurs significantly more often in one sequence compared to another one (Section 15.3), it is relevant to model the occurrences themselves in order to fit the observed frequencies of this word. However, if the problem is precisely to know if a given word occurs in a DNA sequence with a frequency that seems either too low or too high, one needs to compare it to an expected frequency. Usually, we compare the observation with what one would expect in random sequences sharing common properties with the DNA sequence. Under classical sequence models (Section 15.2.1), we can analytically calculate the moments of the count (Section 15.2.2) and sometimes obtain its

distribution or some approximations (Section 15.2.3), leading to p-values (Section 15.2.4). We will end this section by presenting how the Chi motif of *Staphylococcus aureus* was predicted, because of its exceptional frequency, before being experimentally validated [Halpern *et al.* (2007)].

15.2.1 Sequence models

The commonly used sequence models have the property to fit the letter composition of the observed sequence and more generally its composition in small words of a given length. For instance, it is common to fit the 3-letter word composition of coding DNA sequences because the letters of these sequences are read 3 by 3 by the ribosome, which translates each disjoint triplet into amino acids to form a protein. The most intuitive model is therefore the permutation model (or shuffling model), consisting in shuffling the letters of the observed sequence so that the composition remains exactly the same. Preserving exactly the letter composition is an easy task, but it is more difficult for 2-letter words or longer words, from both algorithmic and probabilistic points of view. In that respect, stationary Markov chains are particularly interesting if one accepts fitting the composition on average rather than exactly. Moreover, if one wants to take some periodicity or a heterogeneous composition along the sequence into account, permutation models become very complicated to manipulate.

In our discussion, we will consider a random sequence $\mathbf{S} = X_1 X_2 \cdots X_n$ on the 4-letter DNA alphabet, i.e. $X_i \in \mathcal{A} := \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

Permutation models These models assume that random sequences are uniformly drawn from the set $\mathcal{S}_m$ of sequences having exactly the same counts of words of length 1 up to m as the observed DNA sequence, for a given integer $m \geq 1$. The probability of a sequence $\mathbf{S}$ is then $1/|\mathcal{S}_m|$. For $m = 1$ or $m = 2$, for instance, we have

$$|\mathcal{S}_1| = \frac{n!}{N_{\text{obs}}(\mathtt{a})! \times N_{\text{obs}}(\mathtt{c})! \times N_{\text{obs}}(\mathtt{g})! \times N_{\text{obs}}(\mathtt{t})!}$$

$$|\mathcal{S}_2| = \prod_{a \in \mathcal{A}} \frac{N_{\text{obs}}(a+)!}{\prod_{b \in \mathcal{A}} N_{\text{obs}}(ab)!} \times H_{X_n, X_1}(\mathcal{S}),$$

where $N_{\text{obs}}(\cdot)$ denotes the count in the observed sequence $\mathbf{S}_{\text{obs}}$, $N_{\text{obs}}(a+) := \sum_b N_{\text{obs}}(ab)$ and $H_{X_n,X_1}(\mathcal{S})$ is the cofactor corresponding to row X_n and column X_1 of the matrix $\big(\mathbb{1}\{a = b\} - N(ab)/N(a+)\big)_{a,b \in \mathcal{A}}$ [Whittle (1955)]. Note that the constraint for $\mathbf{S} \in \mathcal{S}_2$ to have the same letter composition as $\mathbf{S}_{\text{obs}}$ is equivalent to starting (resp. ending) with the first (resp. last) letter of $\mathbf{S}_{\text{obs}}$. Indeed, we have $N_{\text{obs}}(a+) = N_{\text{obs}}(a)$ for all $a \in \mathcal{A}$ except for the last nucleotide of $\mathbf{S}_{\text{obs}}$ for which the counts differ from 1. Knowing the letter composition in addition to the dinucleotide composition determines the last letter X_n of the

sequences $\mathbf{S} \in \mathcal{S}$. We use the same procedure for the first letter X_1 by using the numbers $N_{\text{obs}}(+b)$ of dinucleotides that end with b.

Working with these permutation models requires a lot of combinatorics.

Stationary Markov chains Let us consider the first order stationary Markov model, denoted by M1. This means that the random letters X_i are not independent and satisfy the following Markov property:

$$\mathbb{P}(X_i = b \mid X_1, X_2, \ldots, X_{i-1}) = \mathbb{P}(X_i = b \mid X_{i-1}), \quad \forall b \in \mathcal{A}.$$

The transition probabilities will be denoted as follows:

$$\pi(a, b) = \mathbb{P}(X_i = a \mid X_{i-1} = b), \forall a, b \in \mathcal{A};$$

$\Pi = (\pi(a, b))_{a,b}$ will denote the transition matrix. Moreover, all X_i's have the same distribution, namely the stationary distribution μ which satisfies the relation $\mu = \mu\Pi$.

The transition probabilities are estimated by their maximum likelihood estimators (MLEs), i.e.

$$\widehat{\pi}(a, b) = \frac{N(ab)}{N(a+)}, \quad a, b, \in \mathcal{A}, \tag{15.1}$$

where $N(\cdot)$ denotes the number of occurrences in the sequence $\mathbf{S} = X_1 X_2 \cdots X_n$. Moreover, the letter probability $\mu(a)$ is usually estimated by $\widehat{\mu}(a) = \frac{N(a)}{n}$.

An important consequence of this estimation is that the plug-in estimator of the expected number of ab in model M1 is approximately equal to the observed count of ab in the DNA sequence. Indeed, we will see in Section 15.2.2 that $\mathbb{E}[N(ab)] = (n-1)\mu(a)\pi(a, b)$, which leads to

$$\widehat{\mathbb{E}}[N(ab)] := (n-1)\widehat{\mu}(a)\widehat{\pi}(a, b) \simeq N(ab).$$

In other words, model M1 fits on average the 2-letter word composition of the observed sequence.

Similarly, the stationary m-th order Markov chain model (Mm) fits on average the $(m+1)$-letter word composition of the observed sequence. In practice, the choice of the order m of the model Mm is important because it defines the set of reference sequences and, as we will see in Section 15.2.5, this choice often has a strong influence on the statistical results. This influence can already be observed in Table 15.1: the expected counts vary with respect to the chosen model.

Since model Mm on the $\mathcal{A}$ alphabet can be considered as a model M1 on the larger alphabet $\mathcal{A}^m$, we will focus on first order Markov chains in this chapter.

Phased Markov chains for coding sequences The interest in considering phased Markov chains comes from the analysis of coding DNA sequences. Such sequences are split into adjacent 3-letter words called codons, each of which is translated into an amino acid to form a protein. The succession of codons ensures the reading frame for the translation. The nucleotides of a coding DNA sequence are then alternatively the first letter of a codon, the second letter of a codon, the third letter of a codon, and so on. The phase of a nucleotide is its position with respect to the codons; a letter can then be in three different phases in a coding sequence. The three positions of a codon do not have the same importance. First of all, an amino acid is often determined by the two first letters of a codon according to the genetic code. Moreover, the 3D structure of the protein usually implies constraints on the succession of amino acids. It is therefore important to take the phase of the nucleotides into account when modeling coding DNA sequences.

In a phased Markov chain of order 1, the transition probability from letter a to letter b depends on the phase $\phi \in \{1,2,3\}$ of the nucleotide b. We then have the three following transition probabilities:

$$\pi_\phi(a,b) = \mathbb{P}(X_{3i+\phi} = b \mid X_{3i+\phi-1} = a), a, b \in \mathcal{A}.$$

We can also define the distributions $\boldsymbol{\mu}_\phi$ of letters on each phase $\phi \in \{1,2,3\}$. They satisfy $\mu_1 = \mu_3\Pi_1$, $\mu_2 = \mu_1\Pi_2$ and $\mu_3 = \mu_2\Pi_3$.

When estimating these parameters by the maximum likelihood method, we can fit on average the composition of the coding DNA sequence in ab's on phase 1, in ab's on phase 2 and ab's on phase 3, for all $a, b \in \mathcal{A}$.

With an appropriate change of alphabet, the phased Markov model on the $\mathcal{A}$ alphabet can be considered like a model M1 on $\mathcal{A} \times \{1,2,3\}$. It suffices to rewrite the sequence $\mathbf{S}$ over the alphabet $\mathcal{A} \times \{1,2,3\}$ by defining $X_i^\star = (X_i, i \text{ modulo } 3)$. The transition probability from (a,ϕ') to (b,ϕ) is then equal to $\pi_\phi(a,b)$ if $\phi = \phi' + 1$ modulo 3, and 0 otherwise.

Heterogeneous Markov models Some entire chromosomes have been completely sequenced for several years, and it was quickly noticed that their composition is more or less heterogeneous. There may be many reasons for this heterogeneity: genes are more constrained than intergenic regions because they have to code for functional proteins, bacteria can exchange genomic regions (horizontal transfers) but they all have their own signature in terms of composition, etc. It is thus natural to use heterogeneous Markov models. Usually the heterogeneity is considered like a piecewise homogeneity, i.e. homogeneous regions alternate along the genome. If the heterogeneity is known in advance (for instance genes/intergenic regions), one may then use piecewise homogeneous Markov models. When the aim is precisely to recover the heterogeneous structure, then the most popular models in genome analysis are hidden Markov

models. Note that a hidden Markov chain with a hidden state space $\mathcal{Q}$ and an observation space $\mathcal{A}$ can be considered as a Markov chain on $\mathcal{A} \times \mathcal{Q}$.

15.2.2 Mean and variance for the count

The derivation of the expectation and the variance of a word count under the permutation model based on $\mathcal{S}_2$ can be found in Cowan (1991) and Prum *et al.* (1995) [see Schbath (1995b) and Robin *et al.* (2005) for the letter permutation model].

In this section, we assume that the sequence $\mathbf{S} = X_1 X_2 \cdots X_n$ is a first order stationary Markov chain (model M1) with nonzero transition probabilities.

The number of occurrences $N(\mathbf{w})$ of an h-letter word $\mathbf{w} = w_1 w_2 \cdots w_h$ in the sequence $\mathbf{S} = X_1 X_2 \cdots X_n$ can be simply defined by

$$N(\mathbf{w}) = \sum_{i=1}^{n-h+1} Y_i(\mathbf{w}), \tag{15.2}$$

where $Y_i(\mathbf{w})$ equals 1 if and only if an occurrence of $\mathbf{w}$ starts at position i in the sequence and 0 otherwise. Therefore, to get the mean and variance of the count, we need to study the distribution of the random indicators $Y_i(\mathbf{w})$'s, namely their expectation, variance and covariances.

Random indicator of an occurrence The position of an occurrence of $\mathbf{w}$ is defined by the position of its first letter w_1. We define the random indicator $Y_i(\mathbf{w})$ of an occurrence of $\mathbf{w}$ at position i, $1 \leq i \leq n-h+1$, in $\mathbf{S}$ by

$$Y_i(\mathbf{w}) = \begin{cases} 1 \text{ if } (X_i, X_{i+1}, \ldots, X_{i+h-1}) = (w_1, w_2, \ldots w_h), \\ 0 \text{ otherwise.} \end{cases}$$

It is a random Bernoulli variable with parameter $\mathbb{P}(Y_i(\mathbf{w}) = 1)$ given by

$$\begin{aligned} \mathbb{P}(Y_i(\mathbf{w}) = 1) &= \mathbb{P}(X_i = w_1, \ldots, X_{i+h-1} = w_h) \\ &= \mu(w_1) \times \pi(w_1, w_2) \times \cdots \times \pi(w_{h-1}, w_h). \end{aligned}$$

For convenience, $\mu(\mathbf{w})$ will denote the probability for the word $\mathbf{w}$ to appear at a given position in the sequence. The $Y_i(\mathbf{w})$'s are then Bernoulli variables with expectation $\mu(\mathbf{w})$ and variance $\mu(\mathbf{w})[1 - \mu(\mathbf{w})]$, with

$$\mu(\mathbf{w}) = \mu(w_1) \times \prod_{j=2}^{h} \pi(w_{j-1}, w_j). \tag{15.3}$$

However, these random indicators $Y_i(\mathbf{w})$ are not independent, not only because the sequence is Markovian, but most importantly because occurrences of a given word may overlap in a sequence. Consequently, their sum over the positions $i = \{1, \ldots, n-h+1\}$ (namely the number of occurrences—or count—of the word) is not distributed according to a binomial distribution.

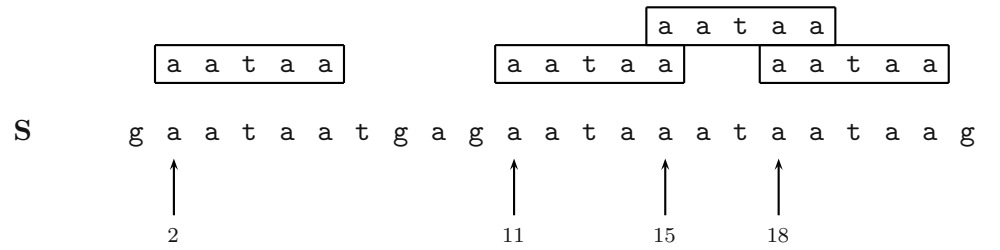


Figure 15.1. Four occurrences of aataa in sequence **S** leading to two clumps of aataa, the first one of size 1 and the second one of size 3.

Overlaps Occurrences of a given word may overlap in a sequence. For instance, $\mathbf{w} = \texttt{aataa}$ occurs four times in the sequence given in Figure 15.1, at positions $i = 2$, 11, 15 and 18. The third occurrence overlaps both the second and the fourth occurrences, leading to a clump of three overlapping occurrences of aataa starting at position 11.

The overlapping structure of a word can be described by two equivalent quantities: the overlapping indicators or the periods.

Overlapping indicators The overlapping indicator $\varepsilon_u(\mathbf{w})$, for $1 \leq u \leq h$, is equal to 1 if two occurrences of $\mathbf{w}$ can overlap on u letters, meaning that the last u letters of $\mathbf{w}$ are identical to its first u letters, and 0 otherwise:

$$\varepsilon_u(\mathbf{w}) = \begin{cases} 1 \text{ if } (w_{h-u+1}, w_{h-u+2}, \ldots, w_h) = (w_1, w_2, \ldots, w_u), \\ 0 \text{ otherwise.} \end{cases}$$

By definition, $\varepsilon_h(\mathbf{w}) = 1$. A *non-overlapping* word $\mathbf{w}$ is such that $\varepsilon_u(\mathbf{w}) = 0$ for all $1 \leq u \leq h-1$.

Periods of a word An integer $p \in \{1, \ldots, h-1\}$ is said to be a period of $\mathbf{w}$ if and only if two occurrences of $\mathbf{w}$ can start at a distance p apart ($\varepsilon_{h-p}(\mathbf{w}) = 1$). It implies the following periodicity: $w_j = w_{j+p}$ for all $j \in \{1, \ldots, h-p\}$.

We denote by $\mathcal{P}(\mathbf{w})$ the set of periods of the word $\mathbf{w}$. For instance, $\mathcal{P}(\texttt{aataataa}) = \{3, 6, 7\}$. Periods that are not a strict multiple of the smallest period are said to be *principal* since they will be more important, as we will see later. $\mathcal{P}'(\mathbf{w})$ denotes the set of the principal periods of $\mathbf{w}$; for instance, $\mathcal{P}'(\texttt{aataataa}) = \{3, 7\}$.

In the rest of our discussion, we will use the periods rather than the overlapping indicators because this simplifies formulas. We will denote by $\mathbf{w}^p\mathbf{w}$ the word composed of two overlapping occurrences of $\mathbf{w}$ starting at a distance p apart:

$$\mathbf{w}^p\mathbf{w} = w_1 \cdots w_p w_1 \cdots w_h.$$

Dependence between occurrences The variables $Y_i(\mathbf{w})$ and $Y_{i+d}(\mathbf{w})$, $d > 0$, are not independent. Their covariance is defined by

$$\begin{aligned}\mathbb{C}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w})] &= \mathbb{E}[Y_i(\mathbf{w}) \times Y_{i+d}(\mathbf{w})] - \mathbb{E}[Y_i(\mathbf{w})] \times \mathbb{E}[Y_{i+d}(\mathbf{w})] \\ &= \mathbb{P}(Y_i(\mathbf{w}) = 1,\ Y_{i+d}(\mathbf{w}) = 1) - [\mu(\mathbf{w})]^2. \end{aligned} \tag{15.4}$$

To calculate the probability $\mathbb{P}(Y_i(\mathbf{w}) = 1,\ Y_{i+d}(\mathbf{w}) = 1)$, we distinguish two cases: $1 \leq d < h$ (two overlapping occurrences) and $d \geq h$ (two disjoint occurrences).

- The probability that $\mathbf{w}$ occurs both at positions i and $i + d$, $1 \leq d < h$, is different from 0 only if d is a period of $\mathbf{w}$. In this case, it is equal to $\mu(\mathbf{w}^d\mathbf{w})$.
- The probability that two disjoint occurrences of $\mathbf{w}$ are separated by $d-h$ letters ($d \geq h$) is given by $\mu(\mathbf{w})\pi^{d-h+1}(w_h, w_1)\mu(\mathbf{w})/\mu(w_1)$, where $\pi^\ell(\cdot,\cdot)$ denotes ℓ-step transition probabilities in $\mathbf{S}$.

The covariance between two random indicators of occurrence is thus

$$\mathbb{C}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w})] = \begin{cases} -[\mu(\mathbf{w})]^2 & \text{if } 0 < d < h,\ d \notin \mathcal{P}(\mathbf{w}), \\ \mu(\mathbf{w}^d\mathbf{w}) - [\mu(\mathbf{w})]^2 & \text{if } d \in \mathcal{P}(\mathbf{w}), \\ [\mu(\mathbf{w})]^2 \left[\dfrac{\pi^{d-h+1}(w_h, w_1)}{\mu(w_1)} - 1\right] & \text{if } d \geq h. \end{cases} \tag{15.5}$$

Mean and variance of the count Finally, we get the following expression for the expectation and the variance of $N(\mathbf{w})$:

$$\mathbb{E}[N(\mathbf{w})] = \sum_{i=1}^{n-h+1} \mathbb{E}[Y_i(\mathbf{w})] = (n-h+1)\mu(\mathbf{w}) \tag{15.6}$$

$$\begin{aligned}\mathbb{V}[N(\mathbf{w})] &= \sum_{i=1}^{n-h+1} \mathbb{V}[Y_i(\mathbf{w})] + 2 \sum_{i=1}^{n-h+1} \sum_{j=i+1}^{n-h+1} \mathbb{C}[Y_i(\mathbf{w}), Y_j(\mathbf{w})] \\ &= (n-h+1)\mu(\mathbf{w})\big(1-\mu(\mathbf{w})\big) + 2 \sum_{i=1}^{n-h+1} \sum_{d=1}^{n-h-i+1} \mathbb{C}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w})], \end{aligned} \tag{15.7}$$

where $\mu(\mathbf{w})$ is given by Equation (15.3), and the covariance term is given by Equation (15.5).

15.2.3 Word count distribution

We will now focus on the statistical distribution of the count $N(\mathbf{w})$. Several methods have been proposed to derive the exact distribution of $N(\mathbf{w})$ in a sequence of independent letters (model M0) or in model M1. Most of them use

pattern matching principles or language theory (see for instance Chapter 7 from Lothaire (2005)). The most probabilistic approach is probably the one that uses the following duality principle: $\mathbb{P}(N(\mathbf{w}) \geq j) = \mathbb{P}(T_j \leq n)$, where T_j denotes the position of the j-th occurrence of the word $\mathbf{w}$ along a random sequence $\mathbf{S}$ of length n. The distribution of T_j can be obtained via the distribution of the distance between two successive occurrences of $\mathbf{w}$ [see Robin and Daudin (1999)]. However, all these methods are fastidious to implement, with many technical limitations as soon as the sequence is long, or if the order of the Markov model is greater than 1, or if the motif is complex. In practice, approximate distributions are used. In this section, we will present two approximations of the word count distribution that have been theoretically proved under some asymptotic framework: the Gaussian approximation, which is valid if the expected count is far enough from zero, and a compound Poisson approximation, which is adapted for the count of rare and clumping events. The quality of these approximations has been studied in Robin and Schbath (2001) and in Nuel (2006). No theoretical result exists so far on the binomial approximation that would result from neglecting the dependence between the occurrences.

Gaussian approximation

Recall that $N(\mathbf{w})$ is a sum of $(n-h+1)$ random Bernoulli variables $Y_i(\mathbf{w})$ with mean $\mu(\mathbf{w})$ and variance $\mu(\mathbf{w})[1-\mu(\mathbf{w})]$.

Asymptotic normality If the Bernoulli variables $Y_i(\mathbf{w})$ were independent, then the classical central limit theorem would ensure that the count the convergence in distribution is a special probabilistic convergence for random variables to a Gaussian variable. But the $Y_i(\mathbf{w})$'s are not independent for two reasons: the occurrences of $\mathbf{w}$ can overlap, and the letters of the sequence are not independent. Nonetheless, by using a central limit theorem for Markov chains, the asymptotic normality of the count can be established:

$$\frac{N(\mathbf{w}) - \mathbb{E}[N(\mathbf{w})]}{\sqrt{\mathbb{V}[N(\mathbf{w})]}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } n \to +\infty. \tag{15.8}$$

Estimating the parameters In the previous convergence, both the expectation and variance of the count depend on the model parameters, which are not known in practice. Let us estimate the expected count by its plug-in estimator, i.e. by replacing the transition probabilities $\pi(a,b)$ by their MLEs $\widehat{\pi}(a,b) = N(ab)/N(a+)$ and the probability $\mu(w_1)$ by $\widehat{\mu}(w_1) = N(w_1)/n$ in Equation (15.6). We then consider the following estimator:

$$\widehat{\mathbb{E}}[N(\mathbf{w})] = \frac{N(w_1w_2) \times \cdots \times N(w_{h-1}w_h)}{N(w_2) \times \cdots \times N(w_{h-1})}. \tag{15.9}$$

Because the estimator $\widehat{\mathbb{E}}_1[N(\mathbf{w})]$ is expressed like a function of several asymptotically Gaussian counts, the δ-method ensures that there exists a constant $v^2(\mathbf{w})$ such that

$$\frac{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]}{\sqrt{(n-h+1)v^2(\mathbf{w})}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } n \to +\infty. \tag{15.10}$$

However, since $\widehat{\mathbb{E}}[N(\mathbf{w})]$ is random, the variance of $\{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]\}$ is different from $\mathbb{V}[N(\mathbf{w})]$ and $(n-h+1)v^2(\mathbf{w})$ is therefore not related to $\mathbb{V}[N(\mathbf{w})]$.

Asymptotic variance Several approaches have been used to derive the asymptotic variance $(n-h+1)v^2(\mathbf{w})$. The first one is the δ-method in Lundstrom (1990): it uses the fact that $n^{-1/2}\{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]\}$ is a function of the asymptotically Gaussian vector $\big(N(\mathbf{w}), N(w_1w_2), \ldots, N(w_{h-1}w_h),\ N(w_2), \ldots, N(w_{h-1})\big)$ from (15.8). However, the function and the size of this vector depend both on the length and on the 2-letter composition of $\mathbf{w}$, so it does not give a unified formula for the asymptotic variance.

Prum *et al.* (1995) proposed a second method: they showed that the estimator $\widehat{\mathbb{E}}[N(\mathbf{w})]$ is asymptotically equivalent to $\mathbb{E}[N(\mathbf{w}) \,|\, \mathcal{S}_2]$, the expected count of $N(\mathbf{w})$ under the 2-letter word permutation model, and that $v^2(\mathbf{w})$ is the limit of $n^{-1}\mathbb{V}[N(\mathbf{w}) \,|\, \mathcal{S}_2]$. They obtained

$$\begin{aligned} v^2(\mathbf{w}) = \mu(\mathbf{w}) + 2 \sum_{p \in \mathcal{P}(\mathbf{w}),\, p<h-1} \mu(\mathbf{w}^p\mathbf{w}) \\ + [\mu(\mathbf{w})]^2 \left[\sum_a \frac{[N_{\mathbf{w}}(a+)]^2}{\mu(a)} - \sum_{a,b} \frac{[N_{\mathbf{w}}(ab)]^2}{\mu(ab)} + \frac{1 - 2N_{\mathbf{w}}(w_1+)}{\mu(w_1)} \right], \end{aligned} \tag{15.11}$$

where $N_{\mathbf{w}}(\cdot)$ stands for the count inside the word $\mathbf{w}$. The overlaps of $\mathbf{w}$ on two or more letters explicitly appear in this formula ($p < h-1$). The overlap on a unique letter is taken into account in the $[\mu(\mathbf{w})]^2$ term.

Since model M1 allows more variability than the corresponding permutation model, one expects the variance $(n-h+1)v^2(\mathbf{w})$ to be smaller than the variance $\mathbb{V}[N(\mathbf{w})]$. This is not difficult to show in the Bernoulli model ($m = 0$); for higher models, it has been numerically verified.

Generalizations to $m > 1$ and to phased models can be found in Schbath *et al.* (1995) and Schbath (1995b). When $m = h-2$, i.e. in the Markov chain model fitting the counts of all the $(h-1)$-letter words (we call this model the maximal model regarding the analysis of h-letter words), a third approach can be used to derive the asymptotic variance. This approach is based on martingale theory and provides a simpler expression for the asymptotic variance [see Prum *et al.* (1995) or Reinert *et al.* (2000)].

Compound Poisson approximation

Poisson approximations can also be used for the count of rare events, i.e. when $\mathbb{E}[N(\mathbf{w})] = \mathrm{O}(1)$. Note that this condition implies that $\log n = \mathrm{O}(h)$ (long enough words). In this section, we will assume the rare event condition but also assume that $h = \mathrm{o}(n)$.

A nice method to establish Poisson approximations of counts is the Chen–Stein method [see Arratia *et al.* (1990) for an introduction and Barbour *et al.* (1992b) for a more general presentation]. This method gives a bound on the total variation distance between the distribution of a sum of dependent Bernoulli variables and the Poisson distribution with the same expectation. The lower the dependence, the better the Poisson approximation quality. Unfortunately, the local dependence between occurrences of an overlapping word $\mathbf{w}$ is too important, and a Poisson approximation of the distribution of $N(\mathbf{w})$ generally does not hold. One can clearly show that the bound provided by the Chen–Stein method does not converge to zero [it is of order $\mu(\mathbf{w}^{p_0}\mathbf{w})$ with p_0 the minimal period of $\mathbf{w}$, see Schbath (1995a)]. But one can also show that a geometric distribution (discrete version of the exponential distribution) does not fit the distribution of the distance between two successive occurrences of an overlapping word [Robin and Daudin (1999)].

The solution is to take advantage of the clump structure (clumps do not overlap) and to use the following relations between the number of occurrences $N(\mathbf{w})$ and the clumps (size and count). Indeed we have

$$N(\mathbf{w}) = \sum_{i=1}^{\widetilde{N}(\mathbf{w})} K_i(\mathbf{w}), \tag{15.12}$$

where $\widetilde{N}(\mathbf{w})$ is the number of clumps of $\mathbf{w}$ and $K_i(\mathbf{w})$ is the size of the i-th clump, but we also have

$$N(\mathbf{w}) = \sum_{k>0} k\widetilde{N}_k(\mathbf{w}), \tag{15.13}$$

where $\widetilde{N}_k(\mathbf{w})$ is the number of clumps of $\mathbf{w}$ of size k in $\mathbf{S}$. Since a compound Poisson variable is defined like $\sum_{k>0} k\, Z_k$ where the Z_k's are independent Poisson variables, or like $\sum_{i=1}^{Z} C_i$ with Z a Poisson variable and the C_i's independent and identically distributed (i.i.d.) variables, the Poisson approximation of the number of clumps (of any size or of size k) is the core of the compound Poisson approximation of the word count. In the remainder of this section, we will explicitly define the clumps and give some of their probabilistic properties.

Random indicator of a clump occurrence A clump of a word $\mathbf{w}$ in a sequence $\mathbf{S}$ is a maximal succession of overlapping occurrences of $\mathbf{w}$. The size of

a clump is the number of occurrences of $\mathbf{w}$ of which the clump is composed. For instance, in Figure 15.1, there are two clumps of `aataa`: one of size 1 starting at position 2, the other one of size 3 starting at position 11. The position of a clump of $\mathbf{w}$ in the sequence is defined by the position (start) of the first occurrence of $\mathbf{w}$ in the clump. Let us define $\widetilde{Y}_i(\mathbf{w})$ as the random indicator that an occurrence of a clump of $\mathbf{w}$ starts at position i in $\mathbf{S}$. A clump of $\mathbf{w}$ occurs at position i if and only if an occurrence of $\mathbf{w}$ occurs at position i without overlapping a previous occurrence of $\mathbf{w}$. Therefore, if we neglect end effects (i.e. when $i < h$), we can write

$$\widetilde{Y}_i(\mathbf{w}) = Y_i(w)[1 - Y_{i-1}(w)] \times \cdots \times [1 - Y_{i-h+1}(w)]. \tag{15.14}$$

(End effects are corrected by considering an infinite sequence.) Now an occurrence of $\mathbf{w}$ which overlaps a previous occurrence of $\mathbf{w}$ is necessarily preceded by a prefix $w_1 \cdots w_p$ of $\mathbf{w}$, where p is a period of $\mathbf{w}$. If we restrict ourselves to principal periods, this is a necessary and sufficient condition [Schbath (1995a)]. For instance, an occurrence of `aataataa` overlaps a previous occurrence of `aataataa` if and only if it is preceded either by `aat` (prefix of size 3) or by `aataata` (prefix of size 7). If it was preceded by `aataat` (prefix of size 6), it would also be preceded by `aat`.

Therefore, we have

$$\widetilde{Y}_i(\mathbf{w}) = \sum_{p \in \mathcal{P}'(\mathbf{w})} [1 - Y_{i-p}(w_1 \cdots w_p)] \times Y_i(w).$$

Clump probability Let us denote by $\widetilde{\mu}(\mathbf{w})$ the probability that a clump of $\mathbf{w}$ occurs at a given position, i.e. $\widetilde{\mu}(\mathbf{w}) = \mathbb{E}[\widetilde{Y}_i(\mathbf{w})]$. The previous equation gives

$$\widetilde{\mu}(\mathbf{w}) = [1 - a(\mathbf{w})] \times \mu(\mathbf{w}), \tag{15.15}$$

where $a(\mathbf{w})$ is the probability that an occurrence of $\mathbf{w}$ overlaps a previous occurrence of $\mathbf{w}$ and is given by

$$a(\mathbf{w}) = \sum_{p \in \mathcal{P}'(\mathbf{w})} \prod_{j=1}^{p} \pi(w_j, w_{j+1}). \tag{15.16}$$

Symmetrically, the probability that an occurrence of $\mathbf{w}$ overlaps a next occurrence of $\mathbf{w}$ is also equal to $a(\mathbf{w})$. Therefore, $a(\mathbf{w})$ will be simply called the probability of self-overlap of $\mathbf{w}$. Note that $a(\mathbf{w}) = 0$ if and only if $\mathbf{w}$ is a non-overlapping word (we assumed that all transition probabilities were nonzero). In that case we also have $\widetilde{Y}_i(\mathbf{w}) = Y_i(\mathbf{w})$ and $\widetilde{\mu}(\mathbf{w}) = \mu(\mathbf{w})$.

Poisson approximation for the number of clumps Let us define the number of clumps of $\mathbf{w}$ by $\widetilde{N}(\mathbf{w}) := \sum_{i=1}^{n-h+1} \widetilde{Y}_i(\mathbf{w})$. The mean number of clumps is then equal to $(n-h+1)\widetilde{\mu}(\mathbf{w}) = [1-a(\mathbf{w})]\mathbb{E}[N(\mathbf{w})]$ from (15.15). The Poisson approximation of $\widetilde{N}(\mathbf{w})$ follows from a direct application of the Chen–Stein method to the Bernoulli variables $\widetilde{Y}_i(\mathbf{w})$ [Schbath (1995a)]. The error bound is indeed of order $(\rho^h + h\mu(\mathbf{w}))$ where $0 < \rho < 1$ is the second largest eigenvalue (in modulus) of the transition matrix Π. Recall that $n\mu(\mathbf{w}) = \mathrm{O}(1)$ from the rare event condition and that $h = \mathrm{o}(n)$.

The exact distribution of the number of clumps of $\mathbf{w}$ in model M1 has been recently derived through its generating function [Stefanov *et al.* (2007)] and compared to the Poisson distribution; The conclusion was that the smaller the expected count of the word, the better the Poisson approximation.

Size of a clump A clump is of size k if and only if the first occurrence of $\mathbf{w}$ in the clump overlaps from the right a second occurrence (probability $a(\mathbf{w})$), the second occurrence of $\mathbf{w}$ in the clump overlaps a third occurrence (probability $a(\mathbf{w})$), ..., the $(k-1)$-th occurrence overlaps a k-th occurrence of $\mathbf{w}$ (probability $a(\mathbf{w})$), and this k-th occurrence of $\mathbf{w}$ does not overlap a next occurrence (probability $1-a(\mathbf{w})$). Thus, if we denote by $K_i(\mathbf{w})$ the size of the i-th clump of $\mathbf{w}$ in the sequence, the random variable $K_i(\mathbf{w})$ is geometrically distributed:

$$\mathbb{P}(K_i(\mathbf{w}) = k) = [1 - a(\mathbf{w})] \times [a(\mathbf{w})]^{(k-1)}. \tag{15.17}$$

Compound Poisson approximation for rare word counts As previously stated, the Poisson approximations of the number of clumps of any size and more particularly of size k for $k \geq 1$ are the key ingredients for the compound Poisson approximation of $N(\mathbf{w})$. Indeed, let us denote by $\mathcal{CP}(\lambda_k, k \geq 1)$ the compound Poisson distribution of $\sum_{k>0} kZ_k$ with $Z_k \sim \mathcal{P}(\lambda_k)$. Since $N(\mathbf{w}) = \sum_{k>0} k\widetilde{N}_k(\mathbf{w})$, the total variation distance properties give

$$d_{\mathrm{TV}}(\mathcal{L}(N(\mathbf{w})), \mathcal{CP}(\mathbb{E}[\widetilde{N}_k(\mathbf{w})], k \geq 1)) \leq d_{\mathrm{TV}}(\mathcal{L}(\widetilde{N}_k(\mathbf{w}), k \geq 1), \otimes\mathcal{P}(\mathbb{E}[\widetilde{N}_k(\mathbf{w})])).$$

The joint Poisson approximation of $(\widetilde{N}_k(\mathbf{w}), k \geq 1)$ is more involved to obtain than the one for $\widetilde{N}(\mathbf{w})$ [Schbath (1995a)], but the error bound is of the same order and

$$\mathbb{E}[\widetilde{N}_k(\mathbf{w})] = [1 - a(\mathbf{w})]^2[a(\mathbf{w})]^{(k-1)}\mathbb{E}[N(\mathbf{w})].$$

The above formula means that the limiting compound Poisson distribution $\mathcal{CP}(\mathbb{E}[\widetilde{N}_k(\mathbf{w})], k \geq 1)$ is in fact a Pólya–Aeppli distribution (also called a geometric-Poisson distribution) with parameter $(\mathbb{E}[\widetilde{N}(\mathbf{w}), a(\mathbf{w})])$ [Johnson *et al.* (1992)].

Direct compound Poisson approximation methods exist and can be alternatively applied to the word count [Erhardsson (1999), Erhardsson (2000)]. Their

advantage is that they provide better error bounds, but they give the same limiting compound Poisson distribution as above [see Lothaire (2005), Chapter 6].

Generalization to Mm and phased models As in the Gaussian approximation, the generalization to the phased Markov model of order 1 is done by rewriting the sequence with the new alphabet $\mathcal{A} \times \{1, 2, 3\}$ (see Section 15.2.1). However, note that the occurrence of a single word $\mathbf{w}$ in sequence $\mathbf{S}$ corresponds to the occurrence of a word family composed of three phased words in the new sequence. Therefore, one has to use the compound Poisson approximation for the count of a set of words in M1 (see Section 15.4.1).

When one changes the alphabet (see Section 15.2.1) to generalize the compound Poisson approximation in model M1 to model Mm, $m > 1$, one must be very careful with the word overlaps. Indeed, there is no one-to-one transformation between clumps of $\mathbf{w}$ in $\mathbf{S}$ and clumps of $\mathbf{w}^\star$ (word $\mathbf{w}$ written on $\mathcal{A}^m$) in the new sequence $\mathbf{S}^\star$. Let us take an example with $m = 2$. Set $\mathbf{w} = \texttt{aataa}$ and let $\mathbf{S}$ be the following sequence on the $\mathcal{A}$ alphabet:

$$\mathbf{S} = \underline{\texttt{gaataa}}\texttt{tgag}\underline{\texttt{aataaataataag}}.$$

$\mathbf{S}$ contains four occurrences of $\mathbf{w}$ and two clumps of $\mathbf{w}$ (one of size 1, the other one of size 3). Now, we write the word and the sequence in the new alphabet $\mathcal{A}^2$. For this, we set $\texttt{ga} = \gamma$, $\texttt{aa} = \alpha$, $\texttt{at} = \beta$, $\texttt{ta} = \tau$, $\texttt{tg} = \delta$, $\texttt{ag} = \kappa$. We have

$$\mathbf{w}^\star = \alpha\beta\tau\alpha \quad \text{and} \quad \mathbf{S}^\star = \gamma\underline{\alpha\beta\tau\alpha}\beta\delta\gamma\kappa\gamma\underline{\alpha\beta\tau\alpha}\,\underline{\alpha\beta\tau\alpha\beta\tau\alpha}\kappa.$$

We can see that the word $\mathbf{w}^\star$ still appears four times in the sequence $\mathbf{S}^\star$ ($N(\mathbf{w})$ is equal to the count of $\mathbf{w}^\star$ in $\mathbf{S}^\star$) but there are now three clumps of $\mathbf{w}^\star$ in $\mathbf{S}^\star$ (two of size 1 and one of size 2). This is due to the fact that $\mathbf{w}^\star$ has just one unique period ($\mathcal{P}(\alpha\beta\tau\alpha) = \{3\}$), whereas $\mathbf{w}$ has two periods ($\mathcal{P}(\texttt{aataa}) = \{3, 4\}$). Therefore, when the results for the word $\mathbf{w}^\star$ in M1 are "translated" into the alphabet $\mathcal{A}$, some overlaps will not appear explicitly in the formulas. In Mm, only the overlaps on m letters or more will be taken into account since the principal periods of $\mathbf{w}^\star$ are the periods of $\mathbf{w}$ that are less than or equal to $(h - m)$. The word $\mathbf{w}^\star$ is non-overlapping as soon as $\mathbf{w}$ is not sufficiently self-overlapping.

15.2.4 *p*-values and scores of exceptionality

The significance of the over-representation of a word $\mathbf{w}$ in a given DNA sequence is measured by the p-value $p(\mathbf{w})$:

$$p(\mathbf{w}) = \mathbb{P}\{N(\mathbf{w}) \geq N_{\text{obs}}(\mathbf{w})\},$$

where $N_{\text{obs}}(\mathbf{w})$ is the observed count of $\mathbf{w}$ in the DNA sequence. If $p(\mathbf{w})$ is close to 0, then the word is exceptionally frequent: there is no chance to observe it

so many times in random sequences. On the other hand, the significance of an under-representation is measured by the p-value $p'(\mathbf{w}) = \mathbb{P}\{N(\mathbf{w}) \leq N_{\text{obs}}(\mathbf{w})\}$. If $p'(\mathbf{w})$ is close to 0, then $\mathbf{w}$ is exceptionally rare under the model: there is no chance that $\mathbf{w}$ occurs so rarely in random sequences. Since the exact distribution of the count $N(\mathbf{w})$ is rarely available in practice, approximate p-values are calculated to detect exceptional words and are usually converted into scores of exceptionality.

Approximate *p*-values A natural way of approximating p-values is to use an approximate distribution of $N(\mathbf{w})$; for instance, a Gaussian distribution for highly expected words or a compound Poisson distribution for rarely expected words, as we have seen in Section 15.2.3. Calculating approximate p-values only requires us to compute the tail of the Gaussian or compound Poisson distribution. An efficient algorithm to compute tails of geometric-Poisson distributions has been proposed by Nuel (2008).

For exceptional words, i.e. words whose count strongly deviates from what is expected, large deviation theory is probably the most accurate way to approximate p-values. This approach has been studied in Nuel (2004). Since it requires sophisticated numerical analysis and longer computation times, this method should be restricted to the most exceptional words (filtered from Gaussian or compound Poisson approximations for instance).

Score of exceptionality In practice, it is often more convenient to manipulate scores from $\mathbb{R}$ than probabilities of the form $p(\mathbf{w}) = \mathbb{P}\{N(\mathbf{w}) \geq N_{\text{obs}}(\mathbf{w})\}$, especially when the ones we are interested in are very close to 0 or very close to 1. For symmetrical reasons we prefer to use the probit transformation rather than the $-\log$ transformation. Therefore, to each probability $p(\mathbf{w})$ we associate the score $u(\mathbf{w})$ such that

$$\mathbb{P}\{\mathcal{N}(0,1) \geq u(\mathbf{w})\} = p(\mathbf{w}).$$

Therefore, words with a high positive score are exceptionally frequent, whereas words with a negative but high absolute value score are exceptionally rare in the observed sequence.

The Gaussian approximation of $N(\mathbf{w})$ has a great practical advantage: it allows us to directly calculate the score of exceptionality $u(\mathbf{w})$ without calculating the associated p-value. Indeed, if we set

$$u(\mathbf{w}) = \frac{N(\mathbf{w}) - \widehat{\mathbb{E}}[N(\mathbf{w})]}{\sqrt{\widehat{\sigma}^2(\mathbf{w})}}, \tag{15.18}$$

where $\widehat{\mathbb{E}}[N(\mathbf{w})]$ is the estimator of the expected count given by Equation (15.9), and $\widehat{\sigma}^2(\mathbf{w})$ is a plug-in estimator of $(n-h+1)v^2(\mathbf{w})$ (cf. Equation (15.11)), namely

$$\widehat{\sigma}^2(\mathbf{w}) = \widehat{\mathbb{E}}[N(\mathbf{w})] + 2 \sum_{p\in\mathcal{P}(\mathbf{w}),p<h-1} \widehat{\mathbb{E}}[N(\mathbf{w}^p\mathbf{w})] \tag{15.19}$$
$$+\{\widehat{\mathbb{E}}[N(\mathbf{w})]\}^2 \left[\sum_a \frac{[N_{\mathbf{w}}(a+)]^2}{N(a)} - \sum_{a,b} \frac{[N_{\mathbf{w}}(ab)]^2}{N(ab)} + \frac{1-2N_{\mathbf{w}}(w_1+)}{N(w_1)}\right],$$

then we have

$$\mathbb{P}\{N(\mathbf{w}) \geq N_{\text{obs}}(\mathbf{w})\} \simeq \mathbb{P}\{\mathcal{N}(0,1) \geq u(\mathbf{w})\}.$$

15.2.5 Example of DNA motif discovery

Chi motifs in bacterial genomes Chi motifs have been identified in several bacterial genomes, and they are not conserved through species. Their identification in a new species is still a challenge. They are involved in the repair of double-strand DNA breaks by homologous recombination. More precisely, they interact specifically with an enzyme that processes along the DNA and degrades it (exonuclease activity). When the enzyme encounters a Chi site, its exonuclease activity is strongly reduced and altered, but it still continues to separate the two DNA strands, forming a substrate for homologous pairing and repair of the deleted DNA parts. Since Chi motifs protect the bacterial genome from degradation and stimulate its repair, it seems important that these motifs appear as frequently as possible along the bacterial genome. Biologists expect them to be significantly over-represented.

Moreover, Chi activity is strongly orientation dependent. The Chi motif is only recognized when the enzyme enters a double-strand DNA molecule from the right side of the motif. In many bacteria for which the Chi motif has been identified, the Chi orientation is correlated with the direction of DNA replication, meaning that it occurs preferentially on the leading strand [El Karoui *et al.* (1999), Halpern *et al.* (2007)]. The over-representation of Chi should then be important on the leading strands. Biologists classically measure the asymmetry strand of a motif by calculating its skew. The skew of a motif $\mathbf{w}$ is simply the ratio $N(\mathbf{w})/N(\overline{\mathbf{w}})$, where $\overline{\mathbf{w}}$ is the reverse complement of the word $\mathbf{w}$; in other words $N(\overline{\mathbf{w}})$ is simply the count of $\mathbf{w}$ in the complementary strand. Therefore, biologists expect Chi to be relatively skewed, i.e. with a skew much greater than one.

***E. coli* as a learning case** The Chi motif of *E. coli* has been known for a long time: it is the 8-letter word `gctggtgg`. If we study the statistical properties of the Chi frequency along the *E. coli* genome, we note some significant characteristics. First of all, its 762 occurrences in the complete genome (concatenation of both leading stands, $n = 4.6\,10^6$) are significantly high whatever model we

Table 15.2. Statistics of `gctggtgg` in the complete genome (left) and in the backbone genome (right) of *E. coli* K12 under various models Mm. The rank is obtained while sorting the 65,536 scores by decreasing order.

	complete genome 762 occurrences				backbone 675 occurrences			
m	$\widehat{\mathbb{E}}_m[N]$	$\widehat{\sigma}^2_m$	u_m	rank	$\widehat{\mathbb{E}}_m[N]$	$\widehat{\sigma}^2_m$	u_m	rank
0	85.9	85.8	72.96	3	73.10	73.02	70.44	3
1	84.9	84.8	73.54	1	71.47	71.32	71.46	1
2	206.8	203.9	38.88	1	186.68	183.82	36.02	1
3	355.5	338.9	22.08	5	315.26	299.68	20.78	1
4	355.3	314.4	22.94	2	309.79	272.90	22.11	2
5	420.9	298.0	19.76	1	376.68	262.42	18.42	1
6	610.1	203.3	10.65	3	539.09	176.02	10.24	1

choose. In other words, its high frequency cannot be explained by the genome composition. As we can see in Table 15.2, Chi has very high over-representation scores and is always among the five most exceptionally frequent 8-letter words. Second, if we restrict the analysis to the *E. coli* backbone[1] ($n = 3.7\,10^6$), Chi becomes the most exceptionally frequent 8-letter word in five models, especially in the maximal model M6 (see Table 15.2). Analyzing only the backbone seems therefore to reduce the noise produced by the regions which are either highly variable or specific to one or few strains (mobile elements). Indeed, there is a priori no biological reason for Chi to occur in such regions.

The choice of the model does not seem to affect the significance of the Chi frequency (it is always exceptional), but this is not a general picture. Note that, when the order of the Markov model increases, the model better fits the sequence composition and fewer exceptional words are found. This is illustrated by the boxplots of Figure 15.2. Moreover, in a high order model we have a more accurate knowledge of the sequence composition than in a low order model: the significance of a word frequency then has no reason to be the same. This point is illustrated by the plot of Figure 15.2 which compares scores in models M1 and M6. We recognize the Chi motif, which is clearly outside the cloud, but let us take the case of the word `ggcgctgg`. It occurs 761 times in the *E.coli* backbone, and it has a significantly high score of 62.4 in model M1 (it is the second most exceptional word) but has a score of 0.8 in model M6 (rank 17100). It simply means that its high frequency can be explained by the composition of 7-letter words; indeed, it is expected about 749 times in M6.

[1]The backbone of a bacterial genome is composed of the genomic regions conserved in several strains of the bacteria. Here, we used the backbone obtained from the alignment of the three strains K12, O157:H7 and CFT and available at `http://genome.jouy.inra.fr/mosaic/`

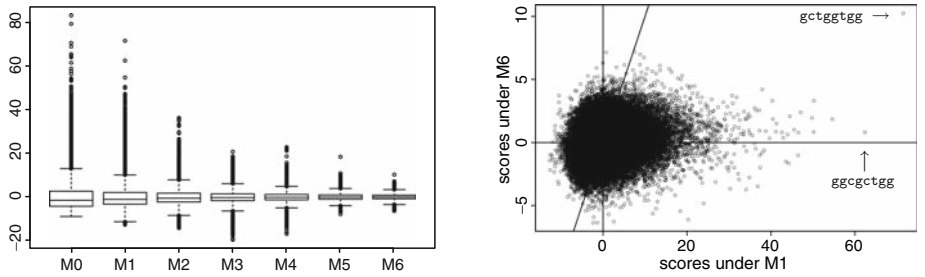


Figure 15.2. Exceptionality scores for the 65,536 8-letter words in the *E. coli* backbone. Left: Boxplots of the scores under models M0 to m6. Right: Scores under models M1 (x-axis) and M6 (y-axis).

The third characteristic of Chi in the *E.coli* backbone is that it is significantly skewed. Its skew is equal to 3.20, and the method described in Section 15.4.1 to assess skew significance gives a score of 6.53 in M6 (p-value of $3.3\,10^{-11}$).

Identification of Chi motif in *S. aureus* We will describe here the strategy used in Halpern *et al.* (2007) to identify the Chi motif in the bacteria *S. aureus*. The first step was to extract the backbone of the *S. aureus* genome by comparing the genome of six strains of the bacteria. The obtained backbone contains about $2.44\,10^6$ letters.

The second step was to search for motifs which are frequent enough, exceptionally frequent and relatively skewed. They started by analyzing 8-letter words (as for *E. coli*) but none of the most over-represented and skewed motifs were frequent enough to be retained as potential Chi candidates. They thus focused on 7-letter words. Scores of exceptionality were calculated with the Gaussian approximation and in the maximal model, namely model M5. Six motifs have an exceptionality score greater than 11 (see Table 15.3 or Figure 15.3 for a global view). Two of them have a negative skew score, so they were not retained. A biological experiment was then performed to test for *S. aureus* Chi activity of the four candidates: `gaaaatg`, `ggattag`, `gaagcgg` and `gaattag`. The conclusion was that `gaagcgg` is necessary and sufficient to confer Chi activity in *S. aureus*. This strategy has also been successfully used to predict and validate the Chi motif of three species of the *Streptococcus* genus [Halpern *et al.* (2007)].

Table 15.3. The 10 most exceptionally frequent 7-letter words under model M5 in the *S. aureus* complete genome. Columns correspond respectively to the word, its observed count, its estimated expected count, its normalizing factor, its score of over-representation under model M5, its observed skew and its skew score under model M0.

w	$N_{\text{obs}}(\mathbf{w})$	$\widehat{\mathbb{E}}_5[N(\mathbf{w})]$	$\widehat{\sigma}^2_5(\mathbf{w})$	$u_5(\mathbf{w})$	Skew	Score
taaaaaa	1542	1214.3	603.4	13.34	1.61	−1.28
gaaaatg	1067	789.9	454.2	13.00	2.48	1.13
taaaatt	1356	1062.6	552.8	12.48	1.04	−1.53
ggattag	266	143.2	97.5	12.43	2.53	1.52
gaagcgg	272	162.4	88.1	11.67	7.56	2.91
gaattag	614	420.7	274.4	11.67	3.89	7.23
gaaaaag	1177	942.1	518.0	10.32	3.52	2.53
taagatt	316	201.3	130.9	10.03	1.07	−2.98
ttaaaag	1059	856.5	431.6	9.75	2.00	3.85
gatttag	657	488.1	305.9	9.66	2.16	4.25

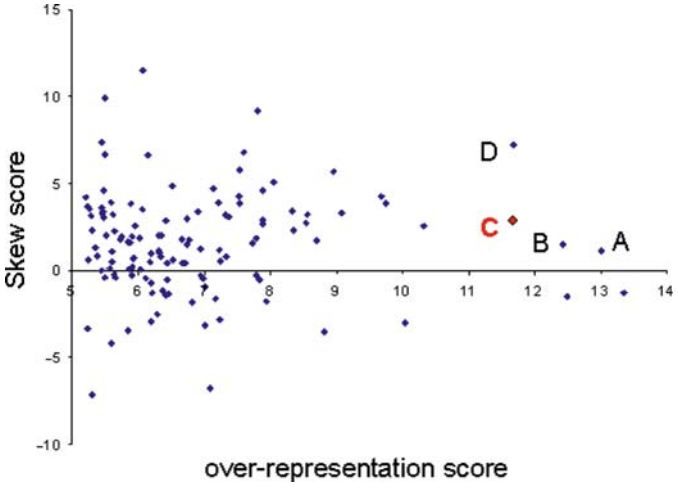


Figure 15.3. Over-representation scores under M5 and skew scores under M0 for the most over-represented 7-letter words (over-representation scores greater than 5) in the complete genome of *S. aureus*. The four best candidates (motifs A to D) are indicated. Motif C (gaagcgg) is the functional Chi site of *S. aureus*.

15.3 Words with Exceptional Distribution

The way the occurrences of a given motif **w** are spread along a sequence or among different sequences or subsequences may provide functional information. When the motif (and its functional properties) is known, this gives us hints about the function of the regions where it occurs (or where it is avoided). Conversely, new interesting motifs may be discovered by comparing their relative frequencies in different well-defined sequences or subsequences (e.g. regions of a genome).

15.3.1 Compound Poisson process

For both problems, we need a probabilistic model describing the motif occurrences process to assess the significance of the observed results. In this section, we will focus on the (compound) Poisson process, which is simple and provides a surprisingly good approximation of the distribution of the word count [Robin and Schbath (2001)].

In this model, the sequence is viewed as a continuous line. To account for possible overlaps between occurrences, the word is assumed to occur in clumps along the sequence. We assume that the counting process of the clumps $\{C(x)\}_{x\geq 0}$ is a homogeneous Poisson process with intensity λ (in all of Section 15.3, we will avoid indexing the quantities by $(\mathbf{w})$ because there will be no ambiguity). Each clump contains a random number of occurrences, referred to as the clump size. The clump sizes $\{K_1, K_2, \dots\}$ are supposed to be i.i.d. with distribution $p(k)$. The counting process $\{N(x)\}_{x\geq 0}$ is hence the compound Poisson process defined as

$$N(x) = \sum_{c=1\dots C(x)} K_c.$$

In the case of a single fixed word, the clump size has a geometric distribution: $p(k) = (1-a)a^{k-1}$, where a stands for the overlapping probability of the word (see Section 15.2.3). In the case of more complex motifs, $p(k)$ may have a more complicated form [Robin (2002)]. The estimates of parameters λ and a depend on the biological question: empirical estimates will fit the observed word frequency (and clumping), while estimates based on a Markov chain model will account for the sequence composition.

15.3.2 Words significantly unbalanced between two sequences

We first consider the detection of motifs having different frequencies between two sequences $\mathbf{S}_1$ and $\mathbf{S}_2$. To avoid artifacts and spurious detections, the testing procedure must account for the different lengths and composition of the

sequences, and for the fact that the word may have an unexpected frequency in one or both of them.

We only consider the non-overlapping case (i.e. $a = 0$). In sequence $\mathbf{S}_i$ $(i = 1, 2)$, the count N_i of $\mathbf{w}$ is supposed to have a Poisson distribution

$$N_i \sim \mathcal{P}(\lambda_i), \qquad \lambda_i = k_i \ell_i \mu_i,$$

where ℓ_i is the length of $\mathbf{S}_i$, $\mu_i = \mu_i(\mathbf{w})$ is the occurrence probability of $\mathbf{w}$ under a Markov model fitted to the composition of $\mathbf{S}_i$ (see Section 15.2.2) and k_i is the exceptionality coefficient of $\mathbf{w}$ in $\mathbf{S}_i$. This framework is described in Robin *et al.* (2007).

Our purpose is to test if the counts of $\mathbf{w}$ in both sequences deviate from their expected values in the same way. We hence want to test the hypothesis $\mathbf{H}_0 : \{k_1 = k_2\}$ versus $\{k_1 \neq k_2\}$. A test procedure can be derived from the following property: for two independent Poisson variables N_1 and N_2 with respective means λ_1 and λ_2, the conditional distribution of N_1 given the sum $N_1 + N_2$ is binomial $\mathcal{B}(N_1 + N_2, \lambda_1/(\lambda_1 + \lambda_2))$. Hence, we have under $\mathbf{H}_0$:

$$N_1 | (N_1 + N_2) \sim \mathcal{B}\left(N_1 + N_2, \ell_1\mu_1/[\ell_1\mu_1 + \ell_2\mu_2]\right).$$

The distribution of the counts of overlapping words is characterized by two parameters (λ and a). For such words, the frequency comparison must be stated in both terms. Assuming that the overlapping probability is the same in the two sequences leads us to define the same binomial test procedure as above on the number of clumps (rather than the number of occurrences itself), which is supposed to have a Poisson distribution (see Section 15.2.3).

To illustrate this procedure, we consider the occurrences of the Chi motif $\mathbf{w}$ = `gctggtgg` in the genome of *E. coli*. This genome can be split into a very conserved part (called the backbone) that is common to various strains of *E. coli* and a remaining part (called variable segments) that is specific to the strain under study: K12. The occurrences of Chi actually never overlap in the whole genome; the number of clumps is the number of occurrences. Chi occurs 691 times in the backbone[2] and 66 times in the variable segments, while the expected numbers of clumps $\ell_i\tilde{\mu}_i$ under model M1 are 73.6 and 11.3, respectively, so $\ell_1\mu_1/(\ell_1\mu_1 + \ell_2\mu_2) = 86.7\%$. It seems therefore more frequent in the backbone than in the loops. To assess the significance of this difference, we calculate the p-value $\Pr\{\mathcal{B}(757, 86.7\%) \geq 691\} = 5.12\,10^{-5}$, which shows that Chi is significantly more frequent in the most conserved regions of the genome, which is consistent with its favorable function.

Testing the equality of the two overlapping probabilities ($\mathbf{H}_0 : \{a_1 = a_2\}$) leads to a hypergeometric test [see Robin *et al.* (2007)].

[2]In contrast to Section 15.2.5, the backbone here is the one obtained from the alignment of two strains: K12 and 0157:H7.

15.3.3 Detecting regions significantly enriched with or devoid of a word

We now want to detect genome regions where the occurrences of a given word **w** are unexpectedly frequent (or rare). The standard strategy in such a situation is to use scan statistics, i.e. distances between successive occurrences. This strategy was first proposed in a genomic context by Karlin and Macken (1991). In this setting, the occurrences are supposed to occur according to a homogeneous Poisson process, which actually corresponds to a non-overlapping word.

Overlapping words can be studied in the compound Poisson model. Since the clump size has a geometric distribution, the distance D between two successive occurrences is either (i) 0 (if the two occurrences belong to the same clump) or (ii) exponential (if they belong to two successive clumps). (i) occurs with probability a and (ii) with probability $(1-a)$. The cumulative distribution function (cdf) of D is hence $F(y) = 1-(1-a)e^{-\lambda y}$. The analogous exact distribution is derived in Robin and Daudin (2001) in the Markov chain model. Because the occurrence process is a renewal process, the cdf F_r of the r-scan, i.e. the cumulated distance D^r between the i-th occurrence and the $(i+r)$-th is simply the r times self-convolution of F: $F_r = F^{\otimes r}$.

Let $D_1^r, D_2^r, \ldots$ denote the successive r-scans. The richest region in terms of occurrences is characterized by the smallest $D_{\min}^r = \min_i D_i^r$. To check if the observed minimum distance $d_{\min}^r$ is significantly small, we need to evaluate $\Pr\{D_{\min}^r \leq d_{\min}^r\}$. A Poisson approximation strategy is proposed by Dembo and Karlin (1992):

$$\Pr\{D_{\min}^r \leq d_{\min}^r\} \approx 1 - \exp[-(N-r)F_r(d_{\min})],$$

where N is the total number of occurrences. Chen–Stein bounds for this approximation are provided. These results can be applied for both the compound Poisson process [Robin (2002)] and Markov chain [Robin and Daudin (2001)] frameworks.

As an illustration, we consider the occurrences of the Chi motif in the genome of *Haemophilus influenzae*, and study their distribution using 3-scans (see Section 15.2.5 to get the description of the Chi motif). The x-axis of Figure 15.4 gives the positions in Mbps, and the y-axis gives the intensity $3/D^3$ multiplied by 10^3 (in log scale); peaks correspond to rich regions. We observe several peaks, the highest one being near the center, i.e. near the terminus of replication. Chi motifs are expected to be frequent here because this region is crucial in the replication mechanism of the cell. The four horizontal lines give, in ascending order, the theoretical mean intensity, the lower bound of the Chen–Stein approximation, the Chen–Stein threshold and the upper bound. We see that several peaks are significant under the M1 model, but the mean intensity of the occurrence process is highly underestimated by this model. Using MLEs,

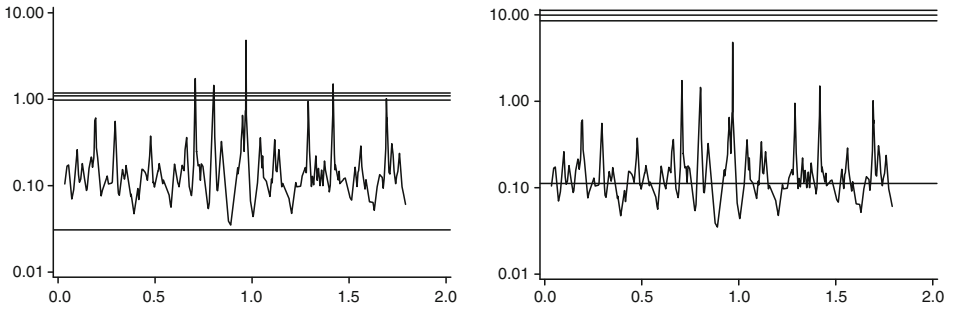


Figure 15.4. Significance of the intensity peaks for the occurrences of the Chi site of *H. influenzae*.

the compound Poisson model fits the observed mean intensity. In this model, even the highest peak is no longer significant.

15.4 More Sophisticated Patterns

Biological motifs are not always exact and simple words. They often contain some uncertainties (degenerated motifs) like the Chi motif `gntggtgg` of *H. influenzae* (the `n` stands for any of the four DNA letters). In this case, we have to consider the occurrences of a set of words rather than a single word. In the case of transcription factor binding sites, we have to deal with several (exact or not) words that should occur at a constrained distance apart (structured motifs). In Section 15.4.1, we give major extensions required to generalize the results on simple words presented in the previous sections to a set of words. Then, we present some results for structured motifs in Section 15.4.2.

15.4.1 Family of words

Let $\mathcal{W}$ be a set (family) of r words: $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$. To simplify the exposition, we will assume that all of the r words have the same length h. In the general case, one just makes the assumption that no word from the family, is part of another word of the family, and the results can be easily generalized.

Distribution of the count of a word family (model M1) The number of occurrences of the word family, denoted by $N(\mathcal{W})$, is simply the sum of the counts of each word taken from $\mathcal{W}$:

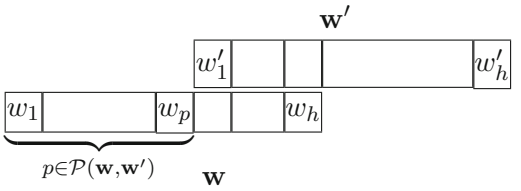
$$N(\mathcal{W}) = \sum_{j=1}^{r} N(\mathbf{w}_j).$$

The expected count $\mathbb{E}[N(\mathcal{W})]$ is then simply the sum of the r expected counts $\mathbb{E}[N(\mathbf{w}_j)]$, $j = 1, \ldots, r$. For the variance, we have $\mathbb{V}[N(\mathcal{W})] = \sum_{j=1}^{r} \mathbb{V}[N(\mathbf{w}_j)] + 2\sum_{j<j'} \mathbb{C}[N(\mathbf{w}_j), N(\mathbf{w}_{j'})]$, so we just need to derive the covariance between two word counts (see below). The Gaussian approximation of $N(\mathcal{W})$ is immediate, and it is easy to derive a score of exceptionality for any family of words. For the compound Poisson approximation, it is much more involved. A first strategy could be to approximate separately the clumps of each word, and then to combine the associated Poisson variables [Reinert and Schbath (1998)]. Unfortunately, words from $\mathcal{W}$ can overlap each other, and this will lead to a bad approximation for overlapping families. The alternative is to consider clumps of the word family itself, i.e. clumps composed of overlapping occurrences of $\mathcal{W}$ [Roquain and Schbath (2007)]. This leads to a compound Poisson distribution, whose parameters are derived from an overlapping probability matrix $(A(w_j, w_{j'}))_{1\leq j,j'\leq r}$, but which is not a geometric Poisson distribution. Tails of general compound Poisson distributions can be calculated by using the algorithm from Barbour *et al.* (1992a).

Covariance between two word counts in M1 Let there be two different words $\mathbf{w}$ and $\mathbf{w}'$ of length h. The covariance $\mathbb{C}[N(\mathbf{w}), N(\mathbf{w}')]$ is given by

$$\mathbb{C}[N(\mathbf{w}), N(\mathbf{w}')] = -\mathbb{E}[N(\mathbf{w})]\,\mathbb{E}[N(\mathbf{w}')] + \sum_{i\neq j} \mathbb{E}[Y_i(\mathbf{w})Y_j(\mathbf{w}')].$$

Because of symmetry, let us restrict ourselves to the calculation of $\mathbb{E}[Y_i(\mathbf{w})Y_{i+d}(\mathbf{w}')]$ for $d > 0$. If $0 < d < h$, an occurrence of $\mathbf{w}'$ at position $i + d$ would overlap an occurrence of $\mathbf{w}$ at position i. We then need to introduce the possible lags between an occurrence of $\mathbf{w}$ and a following overlapping occurrence of $\mathbf{w}'$.



Let $\mathcal{P}(\mathbf{w}, \mathbf{w}')$ be the set of these possible lags, namely

$$p \in \mathcal{P}(\mathbf{w}, \mathbf{w}') \Longleftrightarrow w'_j = w_{j+p}, \quad \forall j \in \{1, \ldots, h - p\}.$$

Overlaps are not necessarily symmetric so $\mathcal{P}(\mathbf{w}, \mathbf{w}') \neq \mathcal{P}(\mathbf{w}', \mathbf{w})$. For instance, `atcg` can be overlapped from the right by `cgct` after a lag of 2 ($\mathcal{P}(\texttt{atcg}, \texttt{cgct}) = \{2\}$), whereas `cgct` cannot be overlapped from the right by `atcg` ($\mathcal{P}(\texttt{cgct}, \texttt{atcg}) = \emptyset$).

If $p \in \mathcal{P}(\mathbf{w}, \mathbf{w}')$, let $\mathbf{w}^p\mathbf{w}'$ be the word composed of two overlapping occurrences of $\mathbf{w}$ and $\mathbf{w}'$: $\mathbf{w}^p\mathbf{w}' = w_1 \cdots w_p w'_1 \cdots w'_h$.

By analogy with Equation (15.5), one can show that

$$\mathbb{E}[Y_i(\mathbf{w}), Y_{i+d}(\mathbf{w}')] = \begin{cases} 0 & \text{if } 0 \leq d < h, d \notin \mathcal{P}(\mathbf{w}, \mathbf{w}'), \\ \mu(\mathbf{w}^d \mathbf{w}') & \text{if } d \in \mathcal{P}(\mathbf{w}, \mathbf{w}'), \\ \mu(\mathbf{w})\mu(\mathbf{w}') \frac{\pi^{d-h+1}(w_h, w_1')}{\mu(w_1')} & \text{if } d \geq h, \end{cases}$$

which finally leads to the following expression for the covariance:

$$\begin{aligned} \mathbb{C}[N(\mathbf{w}), N(\mathbf{w}')] = & -\mathbb{E}[N(\mathbf{w})]\,\mathbb{E}[N(\mathbf{w}')] + \sum_{p \in \mathcal{P}(\mathbf{w}, \mathbf{w}')} (n-h-p+1)\mu(\mathbf{w}^p \mathbf{w}') \\ & + \sum_{p \in \mathcal{P}(\mathbf{w}', \mathbf{w})} (n-h-p+1)\mu(\mathbf{w}'^p \mathbf{w}) \\ & + \mu(\mathbf{w})\mu(\mathbf{w}') \\ & \times \sum_{t=1}^{n-2h+1} (n-2h-t+2) \left[\frac{\pi^t(w_h, w_1')}{\mu(w_1')} + \frac{\pi^t(w_h', w_1)}{\mu(w_1)} \right]. \end{aligned}$$

Note that it is also possible to calculate the asymptotic variance of $N(\mathcal{W}) - \sum_j \widehat{\mathbb{E}}[N(\mathbf{w}_j)]$ by using the conditional covariances of $(N(\mathbf{w}_j), N(\mathbf{w}_\ell))$ in the permutation model (see Schbath *et al.* (1995)).

Skew distribution As we have seen in Section 15.2.5, biologists may be interested in the statistical significance of the skew of a word $\mathbf{w}$. The skew is defined like the ratio $N(\mathbf{w})/N(\overline{\mathbf{w}})$ where $\overline{\mathbf{w}}$ is the reverse complementary[3] word of $\mathbf{w}$ (for instance, if $\mathbf{w} = \texttt{gctggtgg}$ then $\overline{\mathbf{w}} = \texttt{ccaccagc}$). To calculate the significance of the skew, one then has to get (or to approximate) the following p-value:

$$\mathbb{P}\left(\frac{N(\mathbf{w})}{N(\overline{\mathbf{w}})} \geq b \right),$$

where b is the observed skew. This requires at least the joint distribution of $(N(\mathbf{w}), N(\overline{\mathbf{w}}))$.

If we assume that $(N(\mathbf{w}), N(\overline{\mathbf{w}}))$ can be approximated by a Gaussian vector with mean $(\widehat{\mathbb{E}}[N(\mathbf{w})], \widehat{\mathbb{E}}[N(\overline{\mathbf{w}})])$ and covariance matrix Σ, the above p-value can be approximated by

$$\mathbb{P}\left(\mathcal{N}(0,1) \geq \frac{b\widehat{\mathbb{E}}[N(\overline{\mathbf{w}})] - \widehat{\mathbb{E}}[N(\mathbf{w})]}{\sqrt{\Sigma_{11} - 2b\Sigma_{12} + b^2\Sigma_{22}}} \right).$$

The right term of the preceding inequality will then be considered like a score to measure the significance of the skew. Typically, Σ_{11} and Σ_{22} are given by Equation (15.19), and Σ_{12} can be obtained similarly because of the conditional covariances between counts.

[3] `a` is the complement of `t` whereas `c` is the complement of `g`.

If $N(\mathbf{w})$ and $N(\overline{\mathbf{w}})$ are more likely to be (compound) Poisson distributed, no solution exists for now. If $\mathbf{w}$ and $\overline{\mathbf{w}}$ do not overlap each other, their counts can be approximated by two independent geometric Poisson variables [Reinert and Schbath (1998)], but it does not help to derive an asymptotic distribution for the skew.

Distances between multiple words Because of the possible overlaps between words of the family, the distribution of the intersite distances between two word family occurrences depends on which word actually occurs first and which word occurs next [Robin (2002)]. Therefore, in the general case, the occurrences of a set of words do not constitute a renewal process, and the methodology described in Section 15.3.3 cannot be used to get the r-scan distribution. In the Markov chain framework, the occurrences of a set of words turns out to be a semi-Markov process.

15.4.2 Structured motifs

A structured motif is composed of several words which should occur in a given order and at some distance apart from each other. Let consider the simple case of two fixed words $\mathbf{u}$ and $\mathbf{v}$. We define a structured motif $\mathbf{m}$ like a pattern whose $\mathbf{u}$ is a prefix, $\mathbf{v}$ is a suffix and whose length is $|\mathbf{u}| + d + |\mathbf{v}|$, $d \geq 0$. Moreover, we impose $d_1 \leq d \leq d_2$. Since d_1 can be large (typically 12 to 20 for transcription factor binding sites), it is not reasonable to view a structured motif like a set of words (i.e. a very degenerated word). Dedicated methods should then be provided. The two main questions related to structured motif occurrences are: (i) what is the probability that a random sequence contains at least one occurrence of a given structured motif? (ii) Is this structured motif more over-represented in front of genes than along the whole chromosome? For the first question, an approximate probability has been derived by assuming that the random indicator of occurrence $Y_i(\mathbf{m})$ only depends on $Y_{i-1}(\mathbf{m})$ [Robin *et al.* (2002)]. More recently, the generating function of the waiting time for the first occurrence of a structured motif was proposed [Stefanov *et al.* (2007); see also Stefanov (2009)]. For the second question, one can use the test described in Section 15.3.2 which just requires us to compute $\mu(\mathbf{m}) = \mathbb{E}[Y_i(\mathbf{m})]$, the occurrence probability of $\mathbf{m}$. An example of the transcription factor binding site discovery method can be found in Touzain *et al.* (2008).

Occurrence probability The probability for $\mathbf{m}$ to occur at a given position in a random sequence $X_1, X_2, \ldots, X_n$ (model M1) is given by

$$\mu(\mathbf{m}) = \mu(\mathbf{u}) \sum_{d=d_1}^{d_2} \mathbb{P}(D_{\mathbf{u},\mathbf{v}} = d)\mu(\mathbf{v})/\mu(v_1),$$

where $D_{\mathbf{u},\mathbf{v}}$ is the random distance between an occurrence of $\mathbf{u}$ and the next occurrence of $\mathbf{v}$, and v_1 is the first letter of $\mathbf{v}$. The distribution of $D_{\mathbf{u},\mathbf{v}}$ is given in Robin and Daudin (2001) [see also Stefanov (2009)].

15.5 Ongoing Research and Open Problems

Multiple testing problem Multiple testing problems immediately arise in motif detection studies: looking for exceptional 8-letter words leads to performing thousands of tests at the same time. The control of the false discovery rate [Benjamini and Hochberg (1995)] has received huge attention in the last few years in the gene expression context, but it is still neglected in most motif statistic studies. The main difficulty comes from the dependency between the counts—and hence between the tests—of all words under study. Under the null (Markov) model, all word counts are correlated, since they are observed on the same sequence. The covariance between any pair of counts is actually known (see Section 15.4.1), but is difficult to account for in multiple testing procedures, partly because of high dimensionality problems.

Sequence classification Many genomes, e.g. bacterial ones, can be characterized in terms of oligonucleotides composition. This phenomenon is often referred to as the "genome signature." Several new genomic approaches aim at classifying sequences with similar origins: comparative genomics aims at finding similarities between complete genomes, typically in an evolutionary perspective; meta-genome analysis considers sets of hundreds of species living in the same environment (soil, human intestine) and deals with mixtures of subsequences coming from these different species.

As seen before, the Mm Markov chain model accounts for the composition of a sequence in $(m+1)$-letter works. Mixture models [McLachlan and Peel (2000)] provide a natural framework to classify objects into unknown groups. Such a model assumes that the sequences actually come from Q groups, each characterized by one transition matrix; sequence i coming from group number q is a random path with transition matrix $\mathbf{\Pi}_q$. The expectation-maximization algorithm is the standard way to estimate both group proportions and matrices $\mathbf{\Pi}_q$, which make $(Q-1)+3Q4^m$ independent parameters. However, mixture models generally lead to model selection problems, typically to choose the unknown number of groups Q. In the case of sequences, this problem turns out to be very complex because of different sequence lengths: long sequences tend to discriminate very easily from each other, while small sequences have almost no influence on the global model. Combinatorial arguments are needed to evaluate the number of "efficient" parameters, i.e. the number of transition probabilities for which some information can actually be derived from the data.

High throughput sequencing This new technology is likely to be used in many biological experiments in the next decade, typically in the place of microarrays. It consists in sequencing a huge number (40 millions) of small DNA fragments (25 nucleotides) in one run. It can be used to count the number of copies of the transcripts of a given gene, to evaluate its expression level, or to explore the meta-genome of a given ecosystem. Dealing with such large datasets is an open problem. Markov models and motif statistics can probably help to organize all this information, but we admit that we still do not really know how.

References

1. Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method, *Statistical Science*, **5**, 403–434.

2. Barbour, A. D., Chen, L. H. Y. and Loh, W.-L. (1992a). Compound Poisson approximation for nonnegative random variables via Stein's method, *Annals of Probability*, **20**, 1843–1866.

3. Barbour, A. D., Holst, L. and Janson, S. (1992b). *Poisson Approximation*, Oxford University Press, London.

4. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, B*, **57**, 289–300.

5. Cowan, R. (1991). Expected frequencies of DNA patterns using Whittle's formula, *Journal of Applied Probability*, **28**, 886–892.

6. Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes, *Annals of Applied Probability*, **2**, 329–357.

7. El Karoui, M., Biaudet, V., Schbath, S. and Gruss, A. (1999). Characteristics of Chi distribution on several bacterial genomes, *Research in Microbiology*, **150**, 579–587.

8. Erhardsson, T. (1999). Compound Poisson approximation for Markov chains using Stein's method, *Annals of Probability*, **27**, 565–596.

9. Erhardsson, T. (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains, *Annals of Applied Probability*, **10**, 573–591.

10. Halpern, D., Chiapello, H., Schbath, S., Robin, S., Hennequet-Antier, C., Gruss, A. and El Karoui, M. (2007). Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modelling, *PLoS Genetics*, **3**, e153.

11. Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Wiley, New York.

12. Karlin, S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data, *Journal of the American Statistical Association*, **86**, 27–35.

13. Lothaire, M. (2005). *Applied Combinatorics on Words*, volume 105 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, London.

14. Lundstrom, R. (1990). *Stochastic models and statistical methods for DNA sequence data*, Ph.D. thesis, University of Utah, Salt Lake City.

15. McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

16. Nuel, G. (2004). LD-SPatt: Large deviations statistics for patterns on Markov chains, *Journal of Computational Biology*, **11**, 1023–1033.

17. Nuel, G. (2006). Numerical solutions for patterns statistics on Markov chains, *Statistical Applications in Genetics and Molecular Biology*, **5**, Article 26.

18. Nuel, G. (2008). Cumulative distribution function of a geometric Poisson distribution, *Journal of Statistical Computation and Simulation*, **78**, 385–394.

19. Prum, B., Rodolphe, F. and de Turckheim, E. (1995). Finding words with unexpected frequencies in DNA sequences, *Journal of the Royal Statistical Society, B*, **57**, 205–220.

20. Reinert, G. and Schbath, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains, *Journal of Computational Biology*, **5**, 223–254.

21. Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words, *Journal of Computational Biology*, **7**, 1–46.

22. Robin, S. (2002). A compound Poisson model for words occurrences in DNA sequences, *Journal of the Royal Statistical Society, C*, **51**, 437–451.

23. Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters, *Journal of Applied Probability*, **36**, 179–193.

24. Robin, S. and Daudin, J.-J. (2001). Exact distribution of the distances between any occurences of a set of words, *Annals of the Institute of Statistical Mathematics*, **53**, 895–905.

25. Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Occurrence probability of structured motifs in random sequences, *Journal of Computational Biology*, **9**, 761–773.

26. Robin, S., Rodolphe, F. and Schbath, S. (2005). *DNA, Words and Models*, Cambridge University Press, English version of *ADN, mots et modèles*, BELIN 2003.

27. Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences, *Journal of Computational Biology*, **8**, 349–359.

28. Robin, S., Schbath, S. and Vandewalle, V. (2007). Statistical tests to compare motif count exceptionalities, *BMC Bioinformatics*, **8**, 1–20.

29. Roquain, E. and Schbath, S. (2007). Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain, *Advances in Applied Probability*, **39**, 128–140.

30. Schbath, S. (1995a). Compound Poisson approximation of word counts in DNA sequences, *ESAIM: Probability and Statistics*, **1**, 1–16.

31. Schbath, S. (1995b). *Etude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN*, Ph.D. thesis, Université René Descartes, Paris V.

32. Schbath, S., Prum, B. and de Turckheim, E. (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences, *Journal of Computational Biology*, **2**, 417–437.

33. Stefanov, V. (2009). Occurrence of Patterns and Motifs in Random Strings, *Scan Statistics: Methods and Applications*, Glaz, J., Pozdnyakov, V. and Wallenstein, S., eds., Birkhäuser, Boston, MA, 351–367.

34. Stefanov, V., Robin, S. and Schbath, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences, *Discrete Applied Mathematics*, **155**, 868–880.

35. Touzain, F., Schbath, S., Debled-Rennesson, I., Aigle, B., Leblond, P. and Kucherov, G. (2008). SIGffRid: a tool to search for σ factor binding sites in bacterial genomes using comparative approach and biologically driven statistics, *BMC Bioinformatics*, **9**, 1–23.

36. Whittle, P. (1955). Some distribution and moment formulae for the Markov chain, *Journal of the Royal Statistical Society, B*, **17**, 235–242.

16

Occurrence of Patterns and Motifs in Random Strings

Valeri T. Stefanov

School of Mathematics and Statistics, University of Western Australia, Crawley, Australia

Abstract: Patterns and motifs on finite alphabets are of interest in many applied areas, such as computational molecular biology, computer science, communication theory, and reliability theory. The exact distribution theory associated with occurrences of patterns (single or compound) and motifs, in random strings of letters, is treated in this chapter. The strings are generated by a Markov source, and for the case of single patterns, they are generated by general discrete-time or continuous-time models. Here, the interest is in finding closed-form expressions for the distributions of the following quantities: (i) the waiting time until the first occurrence of a pattern (motif), (ii) the intersite distances between consecutive occurrences of such, and (iii) the count of occurrences of a pattern, or more generally, the weighted count of occurrences of a compound pattern, both within a finite time horizon. General exact distribution results are discussed. Also, a brief guide on various methodological tools used in the area is provided in the Introduction.

Keywords and phrases: Pattern, motif, waiting time, Markov chain, semi-Markov process

16.1 Introduction

Patterns and motifs on finite alphabets are of interest in many applied areas, such as computational molecular biology, computer science, communication theory, and reliability theory. A word on an alphabet is called a single pattern, and a set of distinct single patterns (words) is called a compound pattern. The strings (texts) of letters can be generated either by independent and identically distributed multinomial trials, or by general discrete-time or continuous-time models (Markov chains or semi-Markov processes). The main interest, from a

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_16,

probabilistic/statistical point of view, is in finding practicable closed-form expressions for the distributions of the following quantities: the waiting time until the first occurrence of a pattern (single or compound) or motif, the intersite distances between consecutive occurrences of such, and the count(s) of occurrences of a pattern(s) or motif within a finite time horizon. Motifs are special cases of compound patterns which usually contain a huge number of distinct single patterns.

The theory on pattern occurrence attracted a variety of methodological tools. For example, the following methodologies have been widely used in the literature: combinatorial methods and classical probabilistic methods based on conditioning arguments, Markov chain embeddings, Markov renewal embeddings, exponential families, martingale techniques, and automata theory. The usefulness of these methodologies to the area is well illustrated in the sources which follow.

Runs are the simplest patterns. Feller (1950) showed how recurrent event theory can be used to solve problems about success runs. For a comprehensive account of the literature on runs see Balakrishnan and Koutras (2002). The key to handling complex patterns was provided by Conway's leading numbers, which account for the overlapping structure of a pattern. Guibas and Odlyzko (1981) derived results applying elementary methods, and Chryssaphinou and Papastavridis (1990) extended them to more general models [see also Robin and Daudin (1999, 2001), Rukhin (2002, 2006), Han and Hirano (2003), and Inoue and Aki (2007)]. Li (1980) introduced martingale techniques to the area, and Gerber and Li (1981) combined the latter with a relevant Markov chain embedding. Martingale tools have also been used in Pozdnyakov *et al.* (2005), Glaz *et al.* (2006), and Pozdnyakov (2008).

Markov chain embeddings have been widely used in the area for treating problems on pattern occurrence; a few relevant sources are Fu (1996), Chadjiconstantinidis, Antzoulakos, and Koutras (2000), Antzoulakos (2001), Fu and Chang (2002), and Fu and Lou (2003). Blom and Thorburn (1982) made connections with Markov renewal theory, and this was systematically exploited by Biggins and Cannings (1987) and Biggins (1987). Stefanov and Pakes (1997) introduced exponential family methodology, combined with a minimal Markov chain embedding, and Stefanov (2000) extended it in combination with suitable Markov renewal embeddings to handle some special compound patterns (sets of runs).

Nicodème, Salvy, and Flajolet (2002) used automata theory comprehensively. Nuel (2008) combined automata theory with Markov chain embeddings and elaborated on a route which leads, for any given pattern(s), to a minimal embedding Markov chain. Reinert, Schbath, and Waterman (2000) provided a survey on some probabilistic tools used in the theory of patterns, and Szpankowski (2001) treated problems on pattern occurrence associated with

average case analysis of string searching algorithms. The first exact distributional results on structured motifs are found in Stefanov, Robin, and Schbath (2007) [cf. also Robin *et al.* (2002), Nuel (2008), and Pozdnyakov (2008)].

In this chapter, results are discussed which provide explicit, closed-form solutions for the distributions of the aforementioned random quantities associated with the occurrence of patterns and structured motifs. These results are derived using predominantly simple probabilistic tools. Also, for a given alphabet, they require a preliminary (easy) evaluation of a few basic characteristics, and then each pattern case is covered in an automated way.

In Sections 16.2 and 16.3 we discuss single patterns. The strings are generated by discrete- or continuous-time semi-Markov processes. The exact distribution of the waiting time until the first occurrence of a pattern, given any (fixed) portion of it has been reached, is found. Also joint distributional results are discussed. The method relies on the knowledge of basic characteristics associated with the underlying model used to generate the strings. These basic characteristics are the probability generating functions (pgf's) of the waiting times until another letter of the alphabet is reached. In other words, we need to know only the pgf's of the waiting times until the simplest special patterns consisting of a single letter from the alphabet are first reached. These pgf's can be evaluated using well-known analytical results if the underlying model is a discrete- or continuous-time finite-state semi-Markov process. In terms of these basic characteristics, simple recurrence relations are provided; these lead to exact evaluation of the relevant pgf's for any pattern. The results on single patterns, as provided in Sections 16.2 and 16.3, lead to an easy solution for compound patterns, which consist of a small to moderate number of distinct single patterns. This is discussed in Subsection 16.4.1. The distribution of the count, and more generally the weighted count, of a compound pattern within a finite time horizon is discussed in Subsection 16.4.2. A neat explicit expression is derived for this distribution in terms of the aforementioned waiting time distributions. The result in Subsection 16.4.2 has not appeared in the literature before. Structured motifs are covered in Subsection 16.4.3. It is shown that results on compound patterns, consisting of only two single patterns, are enough to derive exact distribution results on structured motifs.

16.2 Patterns: Discrete-Time Models

In this section we explain how to derive a closed-form expression for the pgf of the waiting time to reach a pattern (word) starting from either a given letter or an already-achieved portion of the pattern. The strings of letters are generated by a finite-state discrete-time Markov chain whose state space and states are also called alphabet and letters, respectively.

Let $\{X(n)\}_{n\geq 0}$ be an ergodic finite-state Markov chain with discrete-time parameter, state space $\{1,2,\ldots,N\}$, and one-step transition probabilities $p_{i,j}$, $i,j=1,2,\ldots,N$. Denote by $g_{i,j}(t)$ the pgf of the waiting time, $\tau_{i,j}$, to reach state j from state i, that is $g_{i,j}(t)=E(t^{\tau_{i,j}})$, and

$$\tau_{i,j}=\inf\{n: X(n)=j|X(0)=i\}.$$

We assume $\tau_{i,i}=0$, and therefore $g_{i,i}(t)=1$, for each i. The first return time to state i is denoted by $\tilde{\tau}_{i,i}$, that is,

$$\tilde{\tau}_{i,i}=\inf\{n>0: X(n)=i|X(0)=i\},$$

and its pgf is denoted by $\tilde{g}_{i,i}(t)$.

The pattern of interest is $\mathbf{w}_k=w_1w_2\ldots w_k$, where $1\leq w_i\leq N$, $i=1,2,\ldots,k$. For $j<k$, the subpattern $\mathbf{w}_j$ is also called a prefix of $\mathbf{w}_k$. For each j, $j=2,3,\ldots,k-1$, and $r<j$, and each n, $n=1,2,\ldots,N$, denote by $I_{r,j,n}$ the indicator function which is equal to one if and only if none of the strings $w_iw_{i+1}\ldots w_jn$ for $i=2,3,\ldots,r$ is a prefix of $\mathbf{w}_k$ but $w_{r+1}w_{r+2}\ldots w_jn$ is. Also, the indicator function $I_{j,j,n}$ is equal to one if and only if none of the strings $w_iw_{i+1}\ldots w_jn$ for $i=2,3,\ldots,j$ is a prefix of $\mathbf{w}_k$.

Denote by $G_j^{(s)}(t)$ ($\tilde{G}_j^{(s)}(t)$), $j=1,2,\ldots,k$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$ from state s, allowing (not allowing) the initial state s to contribute to the pattern. Also, denote by $G_j^{(\mathbf{w}_r)}(t), 1\leq r\leq j$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$, given that the pattern $\mathbf{w}_r$ has already been reached (note that $G_j^{(\mathbf{w}_j)}(t)=1$). The following theorem provides a simple route for evaluating these pgf's knowing the pgf's, $g_{i,j}(t)$, of the transition times between the states of the original Markov chain $X(n)$. The expressions for the pgf's $g_{i,j}(t)$ are easily recoverable from well-known analytical results [see Theorem 2.19 on page 81 of Kijima (1997)], for any given finite-state Markov chain with not too large a state space.

Theorem 16.2.1 *Let the pattern of interest be* $\mathbf{w}_k$. *The following recurrence relations hold for each* j, $j=1,2,\ldots,k-1$, *and each* r, $r=1,2,\ldots,j$ *(with the convention* $\sum_{i=1}^{0}=0$*):*

$$\tilde{G}_{j+1}^{(s)}(t)=\frac{p_{w_j,w_{j+1}}t\tilde{G}_j^{(s)}(t)}{1-\sum\limits_{n=1,\,n\neq w_{j+1}}^{N}p_{w_j,n}t\left(\sum\limits_{i=1}^{j-1}I_{i,j,n}G_j^{(\mathbf{w}_{j-i+1})}(t)+I_{j,j,n}G_j^{(n)}(t)\right)},$$

$$G_{j+1}^{(\mathbf{w}_r)}(t)=\frac{p_{w_j,w_{j+1}}tG_j^{(\mathbf{w}_r)}(t)}{1-\sum\limits_{n=1,\,n\neq w_{j+1}}^{N}p_{w_j,n}t\left(\sum\limits_{i=1}^{j-1}I_{i,j,n}G_j^{(\mathbf{w}_{j-i+1})}(t)+I_{j,j,n}G_j^{(n)}(t)\right)},$$

where

$$\begin{aligned}
\tilde{G}_{j+1}^{(s)}(t) &= G_{j+1}^{(s)}(t), \qquad \text{if } s \neq w_1, \\
\tilde{G}_{j+1}^{(w_1)}(t) &= \tilde{g}_{w_1,w_1}(t) G_{j+1}^{(w_1)}(t), \\
G_1^{(s)}(t) &= g_{s,w_1}(t), \\
\tilde{G}_1^{(s)}(t) &= \tilde{g}_{w_1,w_1}(t) = \sum_{n=1}^{N} p_{w_1,n} t g_{n,w_1}(t),
\end{aligned}$$

and the $g_{i,j}(t)$ and the indicator functions $I_{i,j,n}$ are as above.

The pgf of the intersite distance between consecutive occurrences of the pattern $\mathbf{w}_k$ is given by $G_k^{(\mathbf{w}_j)}(t)$, where j is the largest integer such that $\mathbf{w}_j$ is a proper prefix as well as a suffix of the pattern $\mathbf{w}_k$. Also, the pgf of the waiting time until the r-th occurrence of the pattern $\mathbf{w}_k$, given the initial state i, is equal to $G_k^{(i)}(t)\left(G_k^{(\mathbf{w}_j)}(t)\right)^{r-1}$, where j has the same property as above.

The proof of Theorem 16.2.1 is based on the following simple idea. Let $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$ be the waiting time for the first return (strictly positive) from pattern $\mathbf{w}_j$ to itself given that the pattern $\mathbf{w}_{j+1}$ is not achieved. Of course, the pattern $\mathbf{w}_{j+1}$ is not achieved if the first state visited is not state w_{j+1}. Therefore, the pgf of $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$ is equal to

$$g_{\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}}(t) = \sum_{n=1,\, n\neq w_{j+1}}^{N} \frac{p_{w_j,n} t}{1 - p_{w_j,w_{j+1}}} \left(\sum_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n} G_j^{(n)}(t) \right).$$

Then, the waiting time to reach pattern $\mathbf{w}_{j+1}$ starting from state s is equal to one plus a geometric sum of independent random variables, $Y_1, Y_2, \ldots$, say, such that Y_1 has the distribution of the waiting time to reach subpattern $\mathbf{w}_j$ from state s and the remaining Y_n have the distribution of $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$. This implies that

$$\tilde{G}_{j+1}^{(s)}(t) = \frac{p_{w_j,w_{j+1}} t \tilde{G}_j^{(s)}(t)}{1 - \sum_{n=1,\, n\neq w_{j+1}}^{N} p_{w_j,n} t \left(\sum_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n} G_j^{(n)}(t) \right)}.$$

A detailed proof of Theorem 16.2.1 is found in Stefanov (2003).

16.3 Patterns: General Discrete-Time and Continuous-Time Models

In this section, extensions of the result from the preceding section are presented. Finite-state semi-Markov processes, with either discrete- or continuous-time parameters, are the underlying models for generating the strings. Also, joint distributions of the waiting time to reach a pattern, together with the associated counts of occurrences of each letter, are of interest.

16.3.1 Waiting times

The notation from the preceding section is further used here for identifying the counterparts of similar quantities. For example, $g_{i,j}(t)$ will again denote the pgf of the waiting time to reach state j from state i in the more general discrete- or continuous-time model considered here.

Let $\{X(u)\}_{u\geq 0}$ (the time parameter u may be either discrete or continuous) be a semi-Markov process whose associated embedded discrete-time Markov chain has a finite state space $\{1, 2, \ldots, N\}$ and one-step transition probabilities $p_{i,j}$, $i, j = 1, 2, \ldots, N$. For a formal definition of a semi-Markov process see Çinlar (1975). Denote by $\phi_{i,j}(t)$ the pgf of the holding (sojourn) time in state i, given that the next state to be visited is state j (if the holding time distributions are discrete, then the time parameter is discrete). We denote by $g_{i,j}(t)$ the pgf of the waiting time, $\tau_{i,j}$, to reach state j from state i; that is, $g_{i,j}(t) = E(t^{\tau_{i,j}})$, where

$$\tau_{i,j} = \inf\{u : X(u) = j | X(0) = i\}.$$

We assume $\tau_{i,i} = 0$, and therefore $g_{i,i}(t) = 1$, for each i. The first return time to state i is denoted by $\tilde{\tau}_{i,i}$ and its pgf by $\tilde{g}_{i,i}(t)$. Of course, if $X(u)$ is a discrete-time Markov chain,

$$\tilde{\tau}_{i,i} = \inf\{u > 0 : X(u) = i | X(0) = i\},$$

and if $X(u)$ is a continuous-time Markov chain,

$$\tilde{\tau}_{i,i} = \inf\{u > 0 : X(u) = i, X(u-) \neq i | X(0) = i\}.$$

If $X(u)$ is a general semi-Markov process, then $\tilde{\tau}_{i,i}$ is understood to be the waiting time to reach state i from itself given that at least one transition has been made in the associated embedded discrete-time Markov chain. This clarifies the interpretation of $\tilde{\tau}_{i,i}$ in case one-step transitions are allowed from a state to itself in the embedded discrete-time Markov chain.

Again, as in the preceding section, the pattern of interest is denoted by $\mathbf{w}_k$. Denote by $G_j^{(s)}(t)$ $(\tilde{G}_j^{(s)}(t))$, $j = 1, 2, \ldots, k$, the pgf of the waiting time

to reach the pattern $\mathbf{w}_j$ from state s, allowing (not allowing) the initial state s to contribute to the pattern. Also denote by $G_j^{(\mathbf{w}_r)}(t), 1 \le r \le j$, the pgf of the waiting time to reach the pattern $\mathbf{w}_j$, given that the pattern $\mathbf{w}_r$ has already been reached (note that $G_j^{(\mathbf{w}_j)}(t) = 1$). The following theorem provides a simple route for evaluating these pgf's in terms of the following characteristics of the original semi-Markov process $X(u)$: the pgf's, $g_{i,j}(t)$, of the transition times between the states, the pgf's, $\phi_{i,j}(t)$, of the holding time distributions, and the transition probabilities, $p_{i,j}$, of the embedded discrete-time Markov chain.

Theorem 16.3.1 *Let the pattern of interest be* $\mathbf{w}_k$. *The following recurrence relations hold for each* j, $j = 1, 2, \ldots, k-1$, *and each* r, $r = 1, 2, \ldots, j$ *(with the convention* $\sum_{i=1}^{0} = 0$*):*

$$\tilde{G}_{j+1}^{(s)}(t) = \frac{p_{w_j,w_{j+1}}\phi_{w_j,w_{j+1}}(t)\tilde{G}_j^{(s)}(t)}{1 - \sum\limits_{\substack{n=1,\\ n\neq w_{j+1}}}^{N} p_{w_j,n}\phi_{w_j,n}(t)\left(\sum\limits_{i=1}^{j-1} I_{i,j,n}G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n}G_j^{(n)}(t)\right)},$$

$$G_{j+1}^{(\mathbf{w}_r)}(t) = \frac{p_{w_j,w_{j+1}}\phi_{w_j,w_{j+1}}(t)G_j^{(\mathbf{w}_r)}(t)}{1 - \sum\limits_{\substack{n=1,\\ n\neq w_{j+1}}}^{N} p_{w_j,n}\phi_{w_j,n}(t)\left(\sum\limits_{i=1}^{j-1} I_{i,j,n}G_j^{(\mathbf{w}_{j-i+1})}(t) + I_{j,j,n}G_j^{(n)}(t)\right)},$$

where

$$\begin{aligned}
\tilde{G}_{j+1}^{(s)}(t) &= G_{j+1}^{(s)}(t), \qquad \text{if } s \neq w_1,\\
\tilde{G}_{j+1}^{(w_1)}(t) &= \tilde{g}_{w_1,w_1}(t)G_{j+1}^{(w_1)}(t),\\
G_1^{(s)}(t) &= g_{s,w_1}(t),\\
\tilde{G}_1^{(w_1)}(t) &= \tilde{g}_{w_1,w_1}(t) = \sum_{n=1}^{N} p_{w_1,n}\phi_{w_1,n}(t)g_{n,w_1}(t).
\end{aligned}$$

The proof is based on the same idea as that used to prove Theorem 16.2.1. Similarly to the preceding section, denote by $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$ the waiting time to reach $\mathbf{w}_j$ from itself given that the pattern $\mathbf{w}_{j+1}$ is not achieved. Then one may notice that the waiting time to reach pattern $\mathbf{w}_{j+1}$ starting from state s is equal to the sum of two independent random variables, where the first has a pgf which equals $\phi_{w_j,w_{j+1}}(t)$ and the second one is a geometric sum of independent random variables, $Y_1, Y_2, \ldots$, say, such that Y_1 has the distribution of the waiting time to reach subpattern $\mathbf{w}_j$ from state s and the remaining Y_n have the distribution of $\tau_{\mathbf{w}_j|\bar{\mathbf{w}}_{j+1}}$.

16.3.2 Joint generating functions associated with waiting times

In this subsection we consider the same general semi-Markov model $X(u)$ that has been introduced in the preceding subsection. Recall that its embedded discrete-time Markov chain has N states. Throughout this subsection these states will be called 'symbols'. Again the notation from the preceding subsections is further used in this subsection for identifying the counterparts of similar quantities (such as $G_j^{(s)}(\cdot)$, etc.). Note that basic quantities of the underlying model, such as $\tau_{i,j}$ and $\phi_{i,j}$, have the same meaning as that in the preceding subsection.

Let $C_i(u)$ be the count of occurrences of symbol i up to time u, and let $g_{i,j}(\underline{\mathbf{t}})$, where $\underline{\mathbf{t}} = (t_0, t_1, \ldots, t_N)$, be the joint pgf of $(\tau_{i,j}, C_1(\tau_{i,j}), \ldots, C_N(\tau_{i,j}))$, where the $\tau_{i,j}$ have been introduced in the preceding subsection. Likewise, let $\tilde{g}_{i,i}(\underline{\mathbf{t}})$ be the joint pgf of $(\tilde{\tau}_{i,i}, C_1(\tilde{\tau}_{i,i}), \ldots, C_N(\tilde{\tau}_{i,i}))$, where again the $\tilde{\tau}_{i,i}$ have been introduced in the preceding subsection. Note that $g_{i,i}(\underline{\mathbf{t}}) = 1$. Denote by $\nu_j^{(s)}$ the waiting time to reach the pattern $\mathbf{w}_j$ from state s. Let $G_j^{(s)}(\underline{\mathbf{t}})$, $(\tilde{G}_j^{(s)}(\underline{\mathbf{t}}))$, be the joint pgf of $\nu_j^{(s)}, C_1(\nu_j^{(s)}), \ldots, C_N(\nu_j^{(s)})$, allowing (not allowing) the first symbol to contribute to the pattern. Further, let $\nu_j^{(\mathbf{w}_r)}$ be the waiting time to reach the pattern $\mathbf{w}_j$ from the already-reached prefix $\mathbf{w}_r$, and let $G_j^{(\mathbf{w}_r)}(\underline{\mathbf{t}})$ be the joint pgf of $\nu_j^{(\mathbf{w}_r)}, C_1(\nu_j^{(\mathbf{w}_r)}), \ldots, C_N(\nu_j^{(\mathbf{w}_r)})$. Note that the methodology introduced in Stefanov (2000; see Section 3) yields explicit expressions for the pgf's $g_{i,j}(\underline{\mathbf{t}})$ associated with any given semi-Markov process, whose embedded discrete-time Markov chain has a relatively small number of states. Therefore, the recurrence relations in the following theorem provide a simple route for explicit evaluation of the joint pgf's of the waiting time to reach, or the intersite distance between two consecutive occurrences of, a pattern and the associated counts of occurrences of the corresponding symbols (letters).

Theorem 16.3.2 *Let the pattern of interest be* $\mathbf{w}_k$. *The following recurrence relations hold for each* j, $j = 1, 2, \ldots, k-1$, *and each* r, $r = 1, 2, \ldots, j$:

$$\tilde{G}_{j+1}^{(s)}(\underline{\mathbf{t}}) = \frac{p_{w_j,w_{j+1}} t_{w_{j+1}} \phi_{w_j,w_{j+1}}(t_0) \tilde{G}_j^{(s)}(\underline{\mathbf{t}})}{1 - \sum\limits_{\substack{n=1,\\ n\neq w_{j+1}}}^{N} p_{w_j,n} t_n \phi_{w_j,n}(t_0) \left(\sum\limits_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(\underline{\mathbf{t}}) + I_{j,j,n} G_j^{(n)}(\underline{\mathbf{t}}) \right)},$$

$$G_{j+1}^{(\mathbf{w}_r)}(\underline{\mathbf{t}}) = \frac{p_{w_j,w_{j+1}} t_{w_{j+1}} \phi_{w_j,w_{j+1}}(t_0) G_j^{(\mathbf{w}_r)}(\underline{\mathbf{t}})}{1 - \sum\limits_{\substack{n=1,\\ n\neq w_{j+1}}}^{N} p_{w_j,n} t_n \phi_{w_j,n}(t_0) \left(\sum\limits_{i=1}^{j-1} I_{i,j,n} G_j^{(\mathbf{w}_{j-i+1})}(\underline{\mathbf{t}}) + I_{j,j,n} G_j^{(n)}(\underline{\mathbf{t}}) \right)},$$

where

$$\begin{aligned}
\tilde{G}_{j+1}^{(s)}(\underline{\mathbf{t}}) &= G_{j+1}^{(s)}(\underline{\mathbf{t}}), \qquad \text{if } s \neq w_1, \\
\tilde{G}_{j+1}^{(w_1)}(\underline{\mathbf{t}}) &= \tilde{g}_{w_1,w_1}(\underline{\mathbf{t}}) G_{j+1}^{(w_1)}(\underline{\mathbf{t}}), \\
G_1^{(s)}(\underline{\mathbf{t}}) &= g_{s,w_1}(\underline{\mathbf{t}}), \\
\tilde{G}_1^{(w_1)}(\underline{\mathbf{t}}) &= \tilde{g}_{w_1,w_1}(\underline{\mathbf{t}}) = \sum_{n=1}^{N} p_{w_1,n} t_n \phi_{w_1,n}(t_0) g_{n,w_1}(\underline{\mathbf{t}}).
\end{aligned}$$

The proof of this theorem is found in Stefanov (2003).

16.4 Compound Patterns

Throughout this section we assume that the strings are generated by discrete-time Markov chains.

16.4.1 Compound patterns containing a small number of single patterns

Denote by $\mathbf{W}$ a compound pattern which consists of k distinct single patterns, $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots, \mathbf{w}^{(k)}$. The latter may have different lengths, and it is assumed that none of them is a proper substring of any of the others. Let $\mathbf{a}$ be an arbitrary pattern; in particular, if $\mathbf{a}$ has length 1, that is, it is equal to a particular letter, s say, then we will denote $\mathbf{a}$ by s. Introduce the following quantities.

$T_{\mathbf{a},\mathbf{W}}$ — the waiting time, starting from pattern $\mathbf{a}$, to reach for the first time the compound pattern $\mathbf{W}$; if $\mathbf{a}$ equals one of the $\mathbf{w}^{(i)}$, then this waiting time is assumed to be greater than 0;

$T_{\mathbf{a},\mathbf{W}|\mathbf{w}^{(j)}}$ — the waiting time, starting from pattern $\mathbf{a}$, to reach for the first time the compound pattern $\mathbf{W}$, given that $\mathbf{W}$ is reached via $\mathbf{w}^{(j)}$;

$T_{\mathbf{a},\mathbf{b}}$ — the waiting time to reach pattern $\mathbf{b}$ starting from pattern $\mathbf{a}$;

$X_{i,j}$ — the interarrival time between two consecutive occurrences of pattern $\mathbf{W}$, given that the starting pattern is $\mathbf{w}^{(i)}$ and the reached pattern is $\mathbf{w}^{(j)}$;

$r_{i,j}$ — the probability that the first reached pattern from $\mathbf{W}$ is $\mathbf{w}^{(j)}$, given that the starting pattern is $\mathbf{w}^{(i)}$.

Of course, $X_{i,j} = T_{\mathbf{w}^{(i)},\mathbf{W}|\mathbf{w}^{(j)}}$. Introduce the following pgf's:

$$G_{\mathbf{a},\mathbf{W},j}(t) = \sum_{n=1}^{\infty} P\left(T_{\mathbf{a},\mathbf{W}} = T_{\mathbf{a},\mathbf{W}|\mathbf{w}^{(j)}} = n\right) t^n, \qquad j = 1, 2, \ldots, k,$$

and recall that by $G_Y(t)$ we denote the pgf of a random variable Y. Clearly,

$$r_{i,j} = P\left(T_{\mathbf{w}^{(i)},\mathbf{W}} = T_{\mathbf{w}^{(i)},\mathbf{W}|\mathbf{w}^{(j)}}\right) = G_{\mathbf{w}^{(i)},\mathbf{W},j}(1).$$

Also, it is easy to see that

$$G_{X_{i,j}}(t) = \frac{G_{\mathbf{w}^{(i)},\mathbf{W},j}(t)}{r_{i,j}}\,.$$

Therefore, both the $r_{i,j}$ and the pgf's $G_{X_{i,j}}(t)$ can be recovered from the pgf's $G_{\mathbf{w}^{(i)},\mathbf{W},j}(t)$. The following theorem [see Chryssaphinou and Papastavridis (1990) and Gerber and Li (1981)] provides, for each pattern $\mathbf{a}$, a system of linear equations from which one can recover the pgf's $G_{\mathbf{a},\mathbf{W},j}(t)$ and $G_{T_{\mathbf{a},\mathbf{W}}}(t)$ in terms of the pgf's $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$. The $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$ are derived from the results in Section 16.2.

Theorem 16.4.1 *The following identities hold:*

$$\begin{aligned} G_{T_{\mathbf{a},\mathbf{W}}}(t) &= \sum_{j=1}^{k} G_{\mathbf{a},\mathbf{W},j}(t), \\ G_{T_{\mathbf{a},\mathbf{w}^{(i)}}}(t) &= \sum_{j=1}^{k} G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t) G_{\mathbf{a},\mathbf{W},j}(t), \qquad i = 1, 2, \ldots, k. \end{aligned}$$

In particular, we get the following explicit expressions for the $G_{\mathbf{w}^{(i)},\mathbf{W},j}(t)$ in terms of the $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$ if the compound pattern $\mathbf{W} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$ consists of two patterns. For brevity, $G_{T_{i,j}}$ below stands for $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$.

$$\begin{aligned} G_{\mathbf{w}^{(1)},\mathbf{W},1}(t) &= \frac{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}}^2}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\,, \\ G_{\mathbf{w}^{(1)},\mathbf{W},2}(t) &= \frac{G_{T_{1,1}} G_{T_{1,2}} - G_{T_{1,1}} G_{T_{2,1}}}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\,, \\ G_{\mathbf{w}^{(2)},\mathbf{W},1}(t) &= \frac{G_{T_{2,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,2}}}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\,, \\ G_{\mathbf{w}^{(2)},\mathbf{W},2}(t) &= \frac{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{2,1}}^2}{G_{T_{1,1}} G_{T_{2,2}} - G_{T_{1,2}} G_{T_{2,1}}}\,. \end{aligned}$$

16.4.2 Weighted counts of compound patterns

A quantity of interest is the count of occurrences of a compound pattern, $\mathbf{W}$ say (as introduced in Subsection 16.4.1), within a finite time horizon. A more general quantity is the weighted count of pattern occurrences which attaches a weight, h_i say, to each occurrence of a single pattern, $\mathbf{w}^{(i)}$, from $\mathbf{W}$. More specifically, introduce

$$H_{\mathbf{W}}(t) = \sum_{i=1}^{k} h_i N_{\mathbf{w}^{(i)}}(t),$$

where $N_{\mathbf{w}^{(i)}}(t)$ is the count of occurrences of pattern $\mathbf{w}^{(i)}$ within a time interval of length t. Recall the meaning of the $r_{i,j}$, $X_{i,j}$, and $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$ which are introduced in Subsection 16.4.1. Of course, the occurrence of $\mathbf{W}$ can be modelled by a k-state semi-Markov process, where an entry to state i identifies an occurrence of pattern $\mathbf{w}^{(i)}$. The one-step transition probabilities of the embedded discrete-time Markov chain of this semi-Markov process are the $r_{i,j}$. The holding time at state i, given that the next state to be visited is state j, is identified by the random variable $X_{i,j}$. For each initial letter, s say, we augment this semi-Markov process with one initial state, 0 say, and relevant one-step transition probabilities and holding times as follows (we denote the probability to move from state 0 to state j by $r_{0,j}$):

$$r_{0,0} = 0, \quad r_{0,j} = G_{s,\mathbf{W},j}(1), \quad j = 1, 2, \ldots, k,$$

and the holding time at state 0, given that the next state to be visited is state j, is identified by $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$, where the latter and $G_{s,\mathbf{W},j}(t)$ are introduced in Subsection 16.4.1. Now consider the semi-Markov processes, Y_t say, derived from that above as follows. The state space has $(k+1)^2$ states, identified by the pairs (i, j), $i, j = 0, 1, \ldots, k$. The process Y_t enters state (i, j) if pattern $\mathbf{w}^{(i)}$ is reached, given that the next occurrence of $\mathbf{W}$ is via pattern $\mathbf{w}^{(j)}$. The initial states are the states $(0, j)$ for $j = 1, 2, \ldots, k$, and the initial probabilities are the $r_{0,j}$. Clearly, the holding time distributions for this new semi-Markov process do not depend on the next state visited. Also, the holding time in state (i, j) is identified by the random variable $X_{i,j}$, and that in state $(0, j)$ by $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$. Then the weighted count $H_{\mathbf{W}}(t)$, introduced above, is equal to

$$H_{\mathbf{W}}(t) = \sum_{i=0}^{k} \sum_{j=0}^{k} h_i N_{(i,j)}(t),$$

where $N_{(i,j)}(t)$ counts the number of visits of Y_t to state (i, j) within a time interval of length t. Denote by $\nu_{(i_1,j_1),(i_2,j_2)}$ the first passage time of Y_t from state (i_1, j_1) to state (i_2, j_2) and by $L_{H_{\mathbf{W}}}^{\nu_{(i_1,j_1),(i_2,j_2)}}(s_1, s_2)$ the joint Laplace transform of the random variables $\nu_{(i_1,j_1),(i_2,j_2)}$ and $H_{\mathbf{W}}\left(\nu_{(i_1,j_1),(i_2,j_2)}\right)$, that is,

$$L_{H_{\mathbf{W}}}^{\nu_{(i_1,j_1),(i_2,j_2)}}(s_1, s_2) = E\left(\exp\left(-s_1\nu_{(i_1,j_1),(i_2,j_2)} - s_2 H_{\mathbf{W}}(\nu_{(i_1,j_1),(i_2,j_2)})\right)\right).$$

Closed-form expressions for the $L_{H_{\mathbf{W}}}^{\nu_{(i_1,j_1),(i_2,j_2)}}(s_1,s_2)$ are derivable in terms of the $r_{i,j}$ and the Laplace transforms of the $X_{i,j}$ and the $T_{s,\mathbf{W}|\mathbf{w}^{(j)}}$, as explained in Stefanov (2006) for general reward functions on semi-Markov processes. Let

$$L_{t,H_{\mathbf{W}}}^{(s)}(s_1,s_2)=\int_0^\infty\int_0^\infty e^{-s_1t-s_2x}P\left(H_{\mathbf{W}}(t)\le x|\text{ the initial letter is }s\right)dxdt$$

The following theorem follows from a general result on reward functions for semi-Markov processes [see Theorem 2.1 in Stefanov (2006)]. It provides an explicit, closed-form expression for the Laplace transform, $L_{t,H_{\mathbf{W}}}^{(s)}(s_1,s_2)$, of the weighted count of $\mathbf{W}$ occurrences within a time interval of length t, in terms of the $r_{i,j}$, the Laplace transforms, $\mathcal{L}[X_{i,j}](\cdot)$, of the interarrival times $X_{i,j}$ of the compound pattern $\mathbf{W}$, and the Laplace transforms, $\mathcal{L}[T_{s,\mathbf{W}|\mathbf{w}^{(j)}}](\cdot)$, of the waiting time to reach $\mathbf{W}$ from an initial letter s, for $s=1,2,\ldots,N$.

Theorem 16.4.2 *The following identity holds for the Laplace transform* $L_{t,H_{\mathbf{W}}}^{(s)}$:

$$L_{t,H_{\mathbf{W}}}^{(s)}(s_1,s_2)=\sum_{m=1}^{k}r_{0,m}\sum_{i,j=1}^{k}\frac{\left(1-\mathcal{L}[X_{i,j}](s_1+s_2h_i)\right)L_{H_{\mathbf{W}}}^{\nu_{(0,m),(i,j)}}(s_1,s_2)}{s_2(s_1+s_2h_i)\left(1-L_{H_{\mathbf{W}}}^{\nu_{(i,j),(i,j)}}(s_1,s_2)\right)},$$

where the joint Laplace transforms $L_{H_{\mathbf{W}}}^{\nu_{(i_1,j_1),(i_2,j_2)}}(s_1,s_2)$ *have been introduced above.*

16.4.3 Structured motifs

Structured motifs are special compound patterns, usually containing a huge number of single patterns. In this subsection we consider both the waiting time until the first occurrence, and the intersite distance between consecutive occurrences, of a structured motif. The interest in these waiting times is due to the biological challenge of identifying promoter motifs along genomes. A structured motif is composed of several patterns separated by a variable distance. If the number of patterns is n, then the structured motif is said to have n boxes. The formal definition of a structured motif with 2 boxes follows. Let $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ be two patterns of length k_1 and k_2, respectively. The alphabet size equals N, and the strings are generated by the Markov chain introduced in Section 16.2. A structured motif $\mathbf{m}$ formed by the patterns $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, and denoted by $\mathbf{m}=\mathbf{w}^{(1)}(d_1:d_2)\mathbf{w}^{(2)}$, is a string with the following property. Pattern $\mathbf{w}^{(1)}$ is a prefix and pattern $\mathbf{w}^{(2)}$ is a suffix of the string, and the number of letters between the two patterns is not smaller than d_1 and not greater than d_2. Also, it is assumed that patterns $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ appear only once in the string. The pgf's of both the waiting time, $\tau_{\mathbf{m}}^{(s)}$, to reach for the first time the structured motif

$\mathbf{m}$ from state s, and the intersite distance, $\tau_{\mathbf{m}}^{(intersite)}$, between two consecutive occurrences of $\mathbf{m}$, are of interest.

Let $\mathbf{W} = \{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$ be a compound pattern consisting of two patterns. For brevity, denote by $T_{i,j}$, $i, j \in \{1, 2\}$, the waiting time to reach pattern $\mathbf{w}^{(j)}$ from pattern $\mathbf{w}^{(i)}$, and by $T_j^{(s)}$ the waiting time to reach pattern $\mathbf{w}^{(j)}$ from state s. The quantities $r_{i,j}$ and $X_{i,j}$, $i, j \in \{1, 2\}$, are introduced in Subsection 16.4.1. Let

$$a_{i,j}(x) = P(X_{i,j} = x).$$

In order to reach the structured motif $\mathbf{m}$, we need to reach first the pattern $\mathbf{w}^{(1)}$ and, from this occurrence of $\mathbf{w}^{(1)}$, to reach the pattern $\mathbf{w}^{(2)}$ such that $d_1 + k_2 \leq X_{1,2} \leq d_2 + k_2$. Introduce the following random variables:

$$\begin{aligned} F_{12} &= (X_{1,2} \mid X_{1,2} < d_1 + k_2 \ \text{ or } \ X_{1,2} > d_2 + k_2), \\ S_{12} &= (X_{1,2} \mid d_1 + k_2 \leq X_{1,2} \leq d_2 + k_2). \end{aligned}$$

F_{12} corresponds to an occurrence of $\mathbf{w}^{(2)}$ that fails to achieve the structured motif, whereas for S_{12}, $\mathbf{w}^{(2)}$ achieves the structured motif. One may notice that the pgf's of F_{12} and S_{12} are given by

$$\begin{aligned} G_{F_{12}}(t) &= \left(G_{X_{12}}(t) - \sum_{x=d_1+k_2}^{d_2+k_2} a_{1,2}(x) t^x \right) (1 - q_S)^{-1} \\ G_{S_{12}}(t) &= \left(\sum_{x=d_1+k_2}^{d_2+k_2} a_{1,2}(x) t^x \right) q_S^{-1}, \end{aligned}$$

where q_S is the probability of 'success' ($\mathbf{w}^{(2)}$ achieves the structured motif), i.e., the probability that $d_1 + k_2 \leq X_{1,2} \leq d_2 + k_2$. Namely, we have

$$q_S = \sum_{x=d_1+k_2}^{d_2+k_2} a_{1,2}(x).$$

The following theorem provides explicit and calculable expressions for the pgf's of both the waiting time to reach for the first time the structured motif $\mathbf{m} = \mathbf{w}^{(1)}(d_1 : d_2)\mathbf{w}^{(2)}$ from state s, and the intersite distance between two consecutive occurrences of $\mathbf{m}$.

Theorem 16.4.3 *The pgf, $G_{\mathbf{m}}^{(s)}(t)$, of the waiting time to reach for the first time a structured motif $\mathbf{m}$ starting from state s, and the pgf, $G_{\mathbf{m}}^{(intersite)}(t)$, of the intersite distance between two consecutive occurrences of $\mathbf{m}$, admit the following explicit expressions:*

$$G_{\mathbf{m}}^{(s)}(t) = \frac{r_{1,2}\, q_S\, G_{T_1^{(s)}}(t)\, G_{S_{12}}(t)}{(1 - (1 - r_{1,2}) G_{X_{1,1}}(t)) \left(1 - (1 - q_S) \left(\frac{r_{1,2}\, G_{T_{2,1}}(t)\, G_{F_{12}}(t)}{1 - (1 - r_{1,2}) G_{X_{1,1}}(t)} \right) \right)},$$

$$G_{\mathbf{m}}^{(intersite)}(t) = \frac{r_{1,2}\, q_S\, G_{T_{2,1}}(t)\, G_{S_{12}}(t)}{\big(1-(1-r_{1,2})G_{X_{1,1}}(t)\big)\left(1-(1-q_S)\left(\frac{r_{1,2}\, G_{T_{2,1}}(t)\, G_{F_{12}}(t)}{1-(1-r_{1,2})G_{X_{1,1}}(t)}\right)\right)},$$

where $G_{F_{12}}(t), G_{S_{12}}(t)$, *and* q_S *are given above.*

The proof of this theorem is found in Stefanov, Robin, and Schbath (2007). Note that, in view of this theorem, the availability of the pgf's $G_{X_{i,j}}(t)$, $i,j = 1,2$, is enough to calculate explicit, closed-form expressions for $G_{\mathbf{m}}^{(s)}(t)$ and $G_{\mathbf{m}}^{(intersite)}(t)$. Explicit expressions for the $G_{X_{i,j}}(t)$, in terms of the $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$, are derived from the identities at the end of Subsection 16.4.1. Also, recall that the $G_{T_{\mathbf{w}^{(i)},\mathbf{w}^{(j)}}}(t)$ are calculated from Theorem 16.2.1 in Section 16.2.

Neat closed-form expressions for the relevant pgf's associated with structured motifs with n boxes are found in Stefanov, Robin, and Schbath (2009).

References

1. Antzoulakos, D. L. (2001). Waiting times for patterns in a sequence of multistate trials. *Journal of Applied Probability*, **38,** 508–518.

2. Balakrishnan, N. and Koutras, M. (2002). *Runs and Scans with Applications.* Wiley, New York.

3. Biggins, J. D. (1987). A note on repeated sequences in Markov chains. *Advances in Applied Probability*, **19,** 739–742.

4. Biggins, J. D. and Cannings, C. (1987). Markov renewal processes, counters and repeated sequences in Markov chains. *Advances in Applied Probability*, **19,** 521–545.

5. Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained? *Journal of Applied Probability*, **19,** 518–531.

6. Chadjiconstantinidis, S., Antzoulakos, D. L. and Koutras, M. V. (2000). Joint distributions of successes, failures and patterns in enumeration problems. *Advances in Applied Probability*, **32,** 866–884.

7. Chryssaphinou, O. and Papastavridis, S. (1990). The occurrence of a sequence of patterns in repeated dependent experiments. *Theory of Probability and Its Applications*, **35,** 167–173.

8. Çinlar, E. (1975). *Introduction to Stochastic Processes.* Prentice-Hall, Englewood Cliffs, NJ.

9. Feller, W. (1950). *An Introduction to Probability Theory and Its Applications*, Vol. 1. Wiley, New York.

10. Fu, J. C. (1996). Distribution theory of runs and patterns associated with a sequence of multistate trials. *Statistica Sinica*, **6,** 957–974.

11. Fu, J. C. and Chang, Y. M. (2002). On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *Journal of Applied Probability*, **39,** 70–80.

12. Fu, J. C. and Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and its Applications*, World Scientific, Hackensack, NJ.

13. Gerber, H. and Li, S-Y. R. (1981). The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stochastic Processes and Their Applications*, **11,** 101–108.

14. Glaz, J., Kulldorff, M., Pozdnyakov, V. and Steele, J. M. (2006). Gambling teams and waiting times for patterns in two-state Markov chains. *Journal of Applied Probability*, **43,** 127–140.

15. Guibas, L. J. and Odlyzko, A. M. (1981). String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, **30,** 183–208.

16. Han, Q. and Hirano, K. (2003). Sooner and later waiting time problems for patterns in Markov dependent trials. *Journal of Applied Probability*, **40,** 73–86.

17. Inoue, K. and Aki, S. (2007). On generating functions of waiting times and numbers of occurrences of compound patterns in a sequence of multistate trials. *Journal of Applied Probability* **44,** 71–81.

18. Kijima, M. (1997). *Markov Processes for Stochastic Modeling.* Chapman & Hall, London.

19. Li, S-Y. R. (1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Annals of Probability*, **8,** 1171–1176.

20. Nicodème, P., Salvy, B. and Flajolet, P. (2002). Motif statistics. *Theoretical Computer Science*, **287,** 593–617.

21. Nuel, G. (2008). Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *Journal of Applied Probability*, **45,** 226–243.

22. Pozdnyakov, V. (2008). A note on occurrence of gapped patterns in i.i.d. sequences. *Discrete Applied Mathematics* **156,** 93–102.

23. Pozdnyakov, V., Glaz, J., Kulldorff, M. and Steele, J. M. (2005). A martingale approach to scan statistics. *Annals of the Institute of Statistical Mathematics*, **57,** 21–37.

24. Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, **7,** 1–46.

25. Robin, S. and Daudin, J. (1999). Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, **36,** 179–193.

26. Robin, S. and Daudin, J. (2001). Exact distribution of the distances between any occurrences of a set of words. *Annals of the Institute of Statistical Mathematics*, **36**, 895–905.

27. Robin, S., Daudin, J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Occurrence probability of structured motifs in random sequences. *Journal of Computational Biology*, **9,** 761–773.

28. Rukhin, A. (2002). Distribution of the number of words with a prescribed frequency and tests of randomness. *Advances in Applied Probability*, **34,** 775–797.

29. Rukhin, A. (2006). Correlation matrices of chains for Markov sequences, and testing for randomness. (Russian) *Teoriya Veroyatnostei i ee Primeneniya*, **51,** 712–731.

30. Stefanov, V. T. (2000). On some waiting time problems. *Journal of Applied Probability*, **37,** 756–764.

31. Stefanov, V. T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach. *Journal of Applied Probability*, **40,** 881–892.

32. Stefanov, V. T. (2006). Exact distributions for reward functions on semi-Markov and Markov additive processes. *Journal of Applied Probability*, **43,** 1053–1065.

33. Stefanov, V. T. and Pakes, A. G (1997). Explicit distributional results in pattern formation. *Annals of Applied Probability*, **7,** 666–678.

34. Stefanov, V. T., Robin, S. and Schbath, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Applied Mathematics*, **155**, 868–880.

35. Stefanov, V. T., Robin, S. and Schbath, S. (2009). Occurrence of structured motifs in random sequences: arbitrary number of boxes. (in preparation).

36. Szpankowski, W. (2001). *Average Case Analysis of Algorithms on Sequences.* John Wiley & Sons, New York.

17

Detection of Disease Clustering

Toshiro Tango

Department of Technology Assessment and Biostatistics, National Institute of Public Health, Wako-shi, Japan

Abstract: In epidemiological studies, it is often of interest to evaluate whether a disease is randomly distributed over time and/or space after being adjusted for a known heterogeneity, which may provide clues to the etiology of disease. To do this, we can apply tests for spatial randomness, or disease clustering. In this paper, I review the existing tests for disease clustering and discuss the advantages and disadvantages of these test statistics. These tests are illustrated and compared with several real temporal and spatial data sets.

Keywords and phrases: Cluster detection test, epidemiology, global clustering test, likelihood ratio, relative risk, spatial statistics

17.1 Introduction

There has been great public concern about the clustering of health events such as the occurrence of childhood leukemia, birth defects, and cancer. To investigate whether clustering is real and significant, many different tests have been proposed for different purposes. Besag and Newell (1991) classified these tests into two families: *focused tests* and *general tests*. The former family of tests assesses the clustering around a pre-fixed point like a nuclear installation. The latter is aimed at investigating the question of whether clustering occurs over the study region. *General tests* were further classified by Kulldorff (1998) into two groups: the first group, *global clustering tests* (GCTs), is designed for evaluating whether cases tend to come in groups or whether cases are located close to each other no matter when and where they occur, and the second group, *cluster detection tests* (CDTs), is designed to both detect local clusters and evaluate their significance. Recently, Kulldorff (2006) discussed the general framework into which most of the many different proposed test statistics for spatial randomness can be placed.

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_17,

This paper is concerned with *general tests* and is organized as follows. Section 17.2 reviews tests for detecting temporal clustering, and Section 17.3 reviews tests for spatial clustering. This paper concludes with a discussion in Section 17.4.

17.2 Temporal Clustering

17.2.1 Disjoint tests

Ederer, Myers, and Mantel (1964) developed a GCT for temporal clustering using a cell-occupancy approach. They divided the time period into m disjoint subintervals. Under the null hypothesis of no clustering, the n cases are randomly distributed among the subintervals (i.e., are multinomially distributed). The test statistic M is the maximum number of cases occurring in a subinterval, i.e., $M = \max(n_1, ..., n_m)$. If the health event is rare and of unknown etiology, M is summed over several locations and time periods. The sum is tested by using a single degree of freedom chi-square test. Ederer, Myers, and Mantel (1964) and Mantel, Kryscio, and Myers (1976) provide tables of the exact null distribution of M for selected values of m and n.

17.2.2 Scan statistics for individual time points data

Naus (1965) proposed a CDT for temporal clustering that is known as the *scan statistic* and is applicable when individual time points data $(t_1, \ldots, t_n)$ are available during the study period. The test statistic S_d, the maximum number of cases observed in an interval of length d, is found by "scanning" all intervals of length d, known as the scanning window of fixed size d, in the time period. In certain cases, this approach is intuitively more appealing than the disjoint interval approach of Ederer, Myers, and Mantel (1964), but it is more complicated mathematically. A major challenge with the scan statistic has been to find analytical results concerning its statistical significance. Unfortunately, the computations necessary to obtain exact p-values for the scan statistic are complex and often not feasible. For selected interval lengths, time lengths, and sample sizes, the tables of p-values provided by Naus (1966) and Wallenstein (1980) can be used. Knox and Lancashire (1982) found a pragmatic approximation to the p-value but it was not so good. In 1987, Wallenstein and Neff proposed a simple but excellent approximation for small p-values such as $p < 0.10$. Let T denote the length of the entire study period and $w = d/T$. Then we have

$$\Pr\{S_d \geq k \mid n, T\} \approx (\frac{k}{w} - n - 1)b(k \mid n, w) + 2 \sum_{i=k+1}^{n} b(i \mid n, w), \qquad (17.1)$$

where

$$b(i \mid n, w) = \binom{n}{i} w^i (1-w)^{n-i}.$$

Although this formula often gives a poor approximation for larger p-values, it does not matter in terms of statistical significance. For example, when $n = 62$, $k = 7,\ d = 1,\ T = 24$ in examples of trisomy data, we have $p \approx 1.09 > 1$, indicating that the test result is not significant anyway.

Naus (1996) compared the power of the scan test with that of the Ederer, Myers, and Mantel (1964) test and concluded that if the scanning interval is small and the data are continuous over the interval, the scan test is the more powerful of the two. Weinstock (1981) proposed a generalization of the scan test that adjusts for changes in the population at risk. Later, Nagarwalla (1996) extended the scan statistic to one with a variable window, whose size does not need to be chosen *a priori*. Let $(t_1, ..., t_n)$ denote a random sample of n points from the density $f(t)$ in an interval $[0,\ T]$. For the hypothesis testing problem $H_0 : f(x) = 1/T$, $H_1 : f(x) = 1/T + \delta$ for $a \le x \le a + d$, the test is the maximized likelihood ratio test statistic λ, which allows for clusters of variable width d:

$$\lambda = \sup_{d,\ k \ge n_0} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \left(\frac{T}{d}\right)^k \left(\frac{T}{T-d}\right)^{n-k}, \qquad (17.2)$$

where $k = k(a, d)$ is the number of points in the window $(a,\ a + d]$. Nagarwalla gave a simple algorithm for the implementation of the method, but Monte Carlo hypothesis testing is used to obtain the p-value since it is not possible to obtain the null distribution of λ analytically.

17.2.3 Clustering index

Tango (1984) developed a GCT for temporal clustering based on the distribution of counts in m disjoint subintervals. However, it can provide a statistic to estimate the clustering periods which made large contributions to significant clustering. The test is useful when the data are grouped. The test statistic, known as a clustering index, is a quadratic form involving the relative frequencies in each subinterval and a measure of closeness between subintervals,

$$C = \boldsymbol{r}^t \mathbf{A} \boldsymbol{r} = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{n_i n_j}{n^2} a_{ij}, \quad (0 < C \le 1), \qquad (17.3)$$

where $\boldsymbol{r}^t = (n_1, ..., n_m)/n$ and the entries a_{ij} of the $m \times m$ symmetric matrix $\boldsymbol{A}$ are arbitrary known measures of closeness between the ith and jth subintervals with the property $a_{ii} = 1$ and where a_{ij} is a monotonically nonincreasing

function of d_{ij}, the time between the ith and jth subintervals. Tango used the following form as a natural choice:

$$a_{ij} = \exp(-d_{ij}) = \exp(- \mid i - j \mid).$$

The clustering index obtains a maximum value of 1 when all cases occur in the same subinterval. Although the statistic is easy to calculate, the proposed asymptotic null distribution was rather complex for simple use. Whittemore and Keller (1986) showed that the distribution of Tango's index is asymptotically normal with mean and variance that are simple to compute. However, later on, Tango (1990) showed that their normal approximation was very poor for moderately large sample sizes and suggested a central chi-square distribution with degrees of freedom ν adjusted by the skewness as a better approximation, i.e.,

$$\Pr\{C > c \mid H_0\} \approx \Pr\left\{\chi^2_\nu > \nu + \sqrt{2\nu}\left(\frac{c - E(C)}{\sqrt{Var(C)}}\right)\right\}, \tag{17.4}$$

where

$$\begin{aligned}
E(C) &= m^{-2}\{\mathbf{1}^t \boldsymbol{A}\mathbf{1} + n^{-1}\mathrm{tr}[\boldsymbol{AV}]\} \\
\mathrm{Var}(C) &= m^{-4}n^{-1}\{4\mathbf{1}^t \boldsymbol{AVA}\mathbf{1} + 2n^{-2}\mathrm{tr}[(\boldsymbol{AV})^2]\} \\
\nu &= 8/(\sqrt{\beta_1(C)})^2 \\
\sqrt{\beta_1(C)} &= \frac{8\{3\mathbf{1}^t(\boldsymbol{AV})^2\boldsymbol{A}\mathbf{1} + n^{-1}\mathrm{tr}[(\boldsymbol{AV})^3]\}}{\sqrt{n}\{4\mathbf{1}^t \boldsymbol{AVA}\mathbf{1} + 2n^{-1}\mathrm{tr}[(\boldsymbol{AV})^2]\}^{3/2}} \\
\mathbf{1} &= (1, ..., 1)^t \text{ (length } m) \\
\boldsymbol{V} &= \mathrm{diag}(m\mathbf{1}) - \mathbf{1}\mathbf{1}^t,
\end{aligned}$$

where $\mathrm{diag}(\boldsymbol{x})$ is the $m \times m$ diagonal matrix with the vector $\boldsymbol{x}$. If the null hypothesis of no clustering is rejected, we can apply the same idea adopted in the spatial clustering index [Tango (2000)], i.e., the most likely center of clustering period may be identified by the subinterval i with maximum of

$$U_i = \frac{1}{C}\sum_{j=1}^{m} \frac{n_i n_j}{n^2} a_{ij}, \quad (\sum_{i=1}^{m} U_i = 1), \tag{17.5}$$

which denotes the percentage of the ith subinterval's contribution to the significant clustering. Empirically, the subintervals with high outlying percentages will be likely periods of clusters.

17.2.4 Other methods

Bailar, Eisenberg, and Mantel (1970) suggested a GCT for detecting temporal clustering based on the number of pairs of cases in a given area that occur

within a specified length of time d of each other. The numbers of close pairs occurring in q areas are summed. The test statistic is assumed to be approximately normally distributed. Larsen, Holmes, and Heath (1973) developed a rank order GCT for detecting temporal clustering. The time period is divided into disjoint subintervals that are numbered sequentially (i.e., ranked). The test statistic K is the sum of absolute differences between the rank of the subinterval in which a case occurred and the median subinterval rank. Small values of K indicate unimodal clustering. Generally, the K statistics for multiple geographic areas are summed. The resulting statistic is asymptotically normal with simple mean and variance. This test is sensitive only to unimodal clustering; it cannot distinguish multiple clustering from randomness. Molinari, Bonaldi, and Daures (2001) proposed a CDT by applying a piecewise-constant regression model which allows for multiple cluster detection. They used the Akaike information criterion and the Bayesian information criterion to determine the optimal model including the number of clusters.

17.2.5 Illustration with congenital oesophageal atresia data

The data we use to illustrate several tests here consists of individual dates of birth of $n = 35$ cases of the birth defects oesophageal atresia and tracheo-oesophageal fistula observed in a hospital in Birmingham, U.K., from 1950 through 1955. The study was first published by Knox (1959) and subsequently analyzed by Weinstock (1981) using a scan statistic with a fixed window and by Nagarwalla (1996) using a scan statistic with a variable window. The data is shown in Table 17.1. The second column is the number of days past 1 January 1950 on which each case was observed. The third, fourth, and fifth columns of the table denote the frequency of cases per 100 days, 200 days, and 365 days (one year), respectively. Visual inspection of the data suggests that there occurs a clustering during three close subintervals [1200, 1299], [1300, 1399], [1400, 1499] and another less striking concentration occurs in the last three subintervals [1900, 1999], [2000, 2099], [2100, 2199]. We shall show the results of application of the scan statistic with a fixed window, the scan statistics with a variable window, and the clustering index.

1. Scan statistic with a fixed window $d = 100$

   $S_d = 7$ for the cluster of 7 cases from the day 1233 (17 May 1953) to the day 1305 (28 July 1953). Using the approximation (17.1) we obtain $p = 0.088$.

2. Scan statistic with a fixed window $d = 200$

   $S_d = 10$ for the cluster of 10 cases from the day 1233 (17 May 1953) to the day 1390 (21 October 1953). Using (17.1) we obtain $p = 0.0499$.

Table 17.1. $n = 35$ cases of oesophageal atresia and tracheo-oesophageal fistula over 2191 days from 1950 to 1955. Day 1 was set as *1 January 1950*. (Data from Knox, 1959)

Interval	Day number	Frequency per d days		
		$d = 100$	200	365
0–99		0		
100–199	170	1	1	
200–299		0		
300–399	316	1	1	2
400–499	445, 468	2		
500–599		0	2	
600–699		0		
700–799		0	0	2
800–899		0		
900–999	938	1	1	
1000–1099	1034	1		2
1100–1199	1128	1	2	
1200–1299	1233, 1248, 1249, 1252, 1259, 1267	6		
1300–1399	1305, 1385, 1388, 1390	4	10	
1400–1499	1446, 1454, 1458, 1461, 1491	5		14
1500–1599	1583	1	6	
1600–1699	1699	1		
1700–1799	1702, 1787	2	3	
1800–1899		0		6
1900–1999	1924, 1974	2	2	
2000–2099	2049, 2051, 2067, 2075	4		
2100–2199	2108, 2151, 2174	3	7	9
Total		35		

3. Scan statistic with a fixed window $d = 300$

   $S_d = 15$ for the cluster of 15 cases from the day 1233 (17 May 1953) to the day 1491 (30 January 1954). Using (17.1) we obtain $p = 0.0014$.

4. Scan statistic with a fixed window $d = 365$ [Weinstock (1981)]

   $S_d = 16$ for the cluster of 16 cases from the day 1233 (17 May 1953) to the day 1583 (2 May 1954). Using (17.1) we obtain $p = 0.0027$.

5. Scan statistic with a variable window [Nagarwalla (1996)]

   Results of four different scan statistics with fixed windows $d = 100, 200, 300$, and 365 suggest the optimal window could exist between 200

and 365. With $n_0 = 5$, the maximum likelihood ratio (17.2) is $\lambda^* = 43,968$, and the most likely cluster is the set of 15 cases from the day 1233 (17 May 1953) to the day 1491 (30 January 1954), which is the same as that of the scan statistic with fixed window $d = 300$. The optimal and minimum window is $1491 - 1233 + 1 = 259$. Using Monte Carlo testing with 9999 replicates, the observed rank of λ^* due to Nagarwalla's computation is 58, i.e., $p = 0.0058$.

6. Clustering index for the frequency data per 100 days

   Observed standardized clustering index is $c = 5.015$ and using the approximation (17.4) we obtain $p = 0.00027$. By examining the percent contribution U_i to C, we can see that three successive subintervals [1200, 1299], [1300, 1399], [1400, 1499] (15 cases from the day 1233 to the day 1491) have quite large values compared with those of other subintervals, and their contribution is 61.7%, indicating strong clustering period in these three successive subintervals. Furthermore, we can indicate another possible clustering period in two successive subintervals [2000, 2099], [2100, 2199] (7 cases from the day 2049 to the last day 2174) which contributed about 18.7%.

7. Clustering index for the frequency data per 200 days

   Observed standardized clustering index is $c = 5.222$ and using (17.4) we obtain $p = 0.0004$. By examining the percent contribution U_i to C, we can see a cluster in the two successive subintervals [1200, 1399], [1400, 1599] (16 cases from the day 1233 to the day 1583) which has 61.8% contribution. Furthermore, we can indicate another possible clustering period in the last subinterval [2000, 2199] (7 cases from the day 2049 to the last day 2174) which contributed about 18.0%.

8. Clustering index for the frequency data per one year (365 days)

   Observed standardized clustering index is $c = 4.745$ and using (17.4) we obtain $p = 0.0014$. By examining the percent contribution U_i to C, we can see a cluster in the subinterval [1095, 1459] (14 cases from the day 1128 to the day 1458) which has about 51.3% contribution. Furthermore, we can indicate another possible clustering period in the last subinterval [1825, 2190] (9 cases from the day 1924 to the last day 2174) which contributed about 23.6%.

17.2.6 Illustration with trisomy data

In this section, we shall consider a grouped data of $N = 62$ cases of trisomy among karyotyped spontaneous abortions of pregnancies, by calendar month of

the last menstrual period, July 1975 to June 1977, in three New York hospitals. This study was first analyzed by Wallenstein (1980) and subsequently by Tango (1984, 1990). The data is shown in Table 17.2. The trisomy data was tabulated in two ways: (1) monthly data over 24 months, (ii) bimonthly data over 24 months. Visual inspection of the data suggests that a cluster seems to occur during the period November 1976 to January 1977. The results are as follows.

1. Scan statistic [Wallenstein (1980)]

   Wallenstein (1980) applied the scan statistic with a fixed window to individual trisomy data (not shown in his paper). In his illustration, he set $d = 60$ days and found $S_d = 14$, $p = 0.038$ based on his unpub-

Table 17.2. Frequency of trisomy among karyotyped spontaneous abortions of pregnancies, by calendar month of the last menstrual period, July 1975 to June 1977, in three New York hospitals. (Data from Wallenstein, 1980; Tango, 1984)

Year	Month	Frequency	
		per month	per two months
1975	7	0	
	8	4	4
	9	1	
	10	2	3
	11	1	
	12	3	4
1976	1	1	
	2	3	4
	3	2	
	4	2	4
	5	3	
	6	4	7
	7	1	
	8	1	2
	9	1	
	10	2	3
	11	4	
	12	7	11
1977	1	7	
	2	2	9
	3	2	
	4	6	8
	5	1	
	6	2	3
Total		62	

lished extensive table. Linear interpolation based on his Table 17.1 yields $p = 0.040$. Using the approximation (17.1) we obtain $p = 0.037$. In this example, the maximum number of trisomies in two consecutive months was also 14. In general, inspection of *all* 60-day intervals may yield a higher value than the maximum number of two consecutive months.

2. Clustering index [Tango (1984, 1990)]

   All the following three results are significant at the 5% level: (i) for monthly data over 24 months, $C = 0.1139,\ p = 0.023$, (ii) for bimonthly data over 24 months, $C = 0.1975,\ p = 0.035$, and (iii) for monthly data over the last 12 months, $C = 0.2354,\ p = 0.0046$. Using U_i, we can find a likely cluster in the period from November 1976 to January 1977 which has 18 cases and 45.5% contribution.

3. Use of SaTScan

   SaTScan is a free software developed by Kulldorff *et al.* (2007) implementing several types of spatial, temporal, and space-time scan statistics. Purely temporal, analysis is essentially the same idea as Nagarwalla's scan statistic with a variable window for individual data. The details will be described in the next section. We shall show the results only for monthly data over 24 months. The most likely cluster is the set of 28 cases from November 1976 to April 1977. Using Monte Carlo testing with 999 replicates, the observed rank of the log-likelihood ratio statistic is 22, i.e., $p = 0.022$.

17.3 Spatial Clustering

For spatial analysis, it was/is sometimes practically impossible to obtain individual point location data in space due to confidentiality restrictions on individual privacy. Therefore, most tests for spatial clustering developed so far have been designed for regional count data. Although there are some important tests using individual point data or a sample of case-control location data, e.g., Cuzick and Edwards's test (1990) based on k-nearest neighbors and its generalized version by Tango (2007), in what follows, I shall confine myself to considering the situation where an entire study area is divided into m administrative regions (for example, county, census tract, block group) and the region $i(= 1, ..., m)$ has the observed number of cases n_i and the expected number of cases e_i under the null hypothesis of no clustering such that

$$n = \sum_{i=1}^{m} n_i = \sum_{i}^{m} e_i. \tag{17.6}$$

17.3.1 Tests based on adjacencies

Geary (1954) developed a test of spatial clustering that assesses whether rates for adjacent areas are more similar than would be expected if they were randomly distributed among the geographic areas. The test statistic is the ratio of the sum of mean squared differences between rates for pairs of adjacent areas to the weighted sum of mean-squared differences between rates for all pairs of areas. If the rates are geographically distributed at random, the test statistic is close to one; otherwise, it is less than one. Geary derived an expression for the approximate variance of the ratio. If the number of areas is not too small, the ratio is asymptotically normally distributed. Ohno, Aoki, and Aoki (1979) and Ohno and Aoki (1981) developed a simple test for spatial clustering that uses rates for geographic areas (e.g., census tracts, counties, or states) rather than data for individual cases. The test assesses whether the rates in adjacent areas are more similar than would be expected under the null hypothesis of no clustering. For this test, the rate for each area is classified into one of several categories, and each pair of adjacent areas is identified. The test statistic is the number of adjacent concordant pairs; i.e., the number of pairs of areas that are adjacent and have rates in the same category. An overall clustering measure summed across all categories can be obtained as well as category-specific clustering measures. The observed number of adjacent concordant pairs is compared with the expected number by using a chi-square test. Ohno, Aoki, and Aoki (1979) provide a simple formula for calculating the expected number of pairs. Grimson, Wang, and Johnson (1981) proposed a test of spatial clustering for use in detecting clusters of geographic areas designated as high risk. The null hypothesis is that high-risk areas are randomly distributed within a larger area and do not cluster. Given the number of high-risk areas, the test statistic is the number of pairs of high-risk areas that are adjacent to each other. This statistic is equivalent to the category-specific statistic from Ohno, Aoki, and Aoki (1979).

Note that, although these tests based on adjacencies are easy to use, they do not properly take the sampling variability of rates into account, and so they are not recommendable in the sense that they may produce spurious results in practice.

17.3.2 Tests based on scanning regions

As the first method using scanning local regional rates, Openshaw *et al.* (1988) developed a geographical analysis machine (GAM) that is an exploratory tool

for searching for potential clusters. GAM constructs overlapping circles of different radii centered at each grid point defined *a priori*, counts the number of cases and the number of people at risk within the circle, and displays those circles with local incidence proportions exceeding some predefined threshold. However, GAM has attracted much criticism since it produces large numbers of highly correlated ovelapped circles. Turnbull *et al.* (1990), on the other hand, proposed a more statistically sound cluster evaluation permutation procedure (CEPP), where, for each region, a window is constructed by absorbing the nearest neighboring regions such that each window contains just a pre-fixed population size R. These windows vary in geographic shape and size but maintain a constant population size at risk so that observed counts are identically distributed. However, these windows of cases and populations overlap, and the counts are not independently distributed. The test statistic of the CEPP is given by the maximum number of cases in the window, which is not necessarily integer due to the adjustment of each population size to R. Monte Carlo testing is needed to obtain the p-value for the test statistic.

Besag and Newell (1991) considered windows with a pre-fixed number of cases k rather than a pre-fixed population size. It was originally designed for quite rare diseases, and thus a typical value of k might be small such as $k = 2, 4, ...$ Each region with nonzero cases is considered in turn as the center of a possible cluster. When considering a particular region, we label it as region 0 and order the remaining regions by their distance to the region 0. We label these regions $j = 1, 2, ..., m - 1$ and define

$$D_i = \sum_{j=0}^{i} n_{(j)}, \ u_i = \sum_{j=0}^{i} \xi_{(j)},$$

where $n_{(j)}$ and $\xi_{(j)}$ denote the number of cases and population in the region labelled j, respectively. Then, the test statistic for detecting individual clusters is

$$S = \min\{i : D_i \geq k\}. \tag{17.7}$$

Namely, the nearest S regions contain the closest k cases. A small observed value of S indicates a cluster centered at region 0. The significance level for each potential cluster is

$$\Pr\{S \leq s\} = 1 - \sum_{t=0}^{k-1} \exp(-u_s Q)(u_s Q)^t / t!, \quad Q = n_+/\xi_+. \tag{17.8}$$

As the test statistic of overall clustering within the entire study area, Besag and Newell (1991) suggested the total number T_{BN} of significant ($p < 0.05$, say) individual clusters. The significance of the observed T_{BN} may be determined by Monte Carlo simulation.

17.3.3 Spatial scan statistics

Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed the spatial scan statistic, which is a spatial version of the scan statistic with a variable window size and is a generalization of CEPP. The spatial scan statistic imposes a circular window $\mathbf{Z}$ on each centroid of a region. For any of those centroids, the radius of the circle varies from zero to some preset upper limit. If the window contains the centroid of a region, then that whole region is included in the window. In total, a very large number of different but overlapping circular windows are created, each with a different location and size, and each being a potential cluster. Let $\mathbf{Z}_{ik},\ k = 1, \ldots, K_i,$ denote the window composed by the $(k-1)$-nearest neighbors to region i. Then, all the windows to be scanned by the spatial scan statistic are included in the set

$$\mathcal{Z}_1 = \{\mathbf{Z}_{ik} \mid 1 \leq i \leq m,\ 1 \leq k \leq K_i\}.$$

Under the alternative hypothesis, there is an elevated risk within some window $\mathbf{Z}$ as compared to outside:

$$\begin{aligned} H_0 &: E(N(\mathbf{Z})) = e(\mathbf{Z}), \quad \text{for all } \mathbf{Z}, \\ H_1 &: E(N(\mathbf{Z})) > e(\mathbf{Z}), \quad \text{for some } \mathbf{Z}, \end{aligned}$$

where $N()$ and $e()$ denote the random number of cases and the null expected number of cases within the specified window, respectively. For each window, it is possible to compute the likelihood to observe the observed number of cases within and outside the window, respectively. Under the Poisson assumption, which is a typical distribution for rare diseases, the test statistic is the likelihood ratio maximized for $\mathbf{Z}$:

$$\sup_{\mathbf{Z} \in \mathcal{Z}_1} \left(\frac{n(\mathbf{Z})}{e(\mathbf{Z})}\right)^{n(\mathbf{Z})} \left(\frac{n(\mathbf{Z}^c)}{e(\mathbf{Z}^c)}\right)^{n(\mathbf{Z}^c)} I\left(\frac{n(\mathbf{Z})}{e(\mathbf{Z})} > \frac{n(\mathbf{Z}^c)}{e(\mathbf{Z}^c)}\right), \tag{17.9}$$

where $\mathbf{Z}^c$ indicates all the regions outside the window $\mathbf{Z}$, and $n()$ denotes the observed number of cases within the specified window and $I()$ is the indicator function. The window $\mathbf{Z}^*$ that attains the maximum likelihood is defined as the *most likely cluster* (MLC). To find the distribution of the test statistic under the null hypothesis, Monte Carlo hypothesis testing is required. Kulldorff's spatial scan statistic has been applied to a wide variety of epidemiological studies and also to disease surveillance for the detection of disease clusters along with SaTScan Software (Kulldorff *et al.* 2007).

However, since it uses a circular window to scan the potential cluster areas, it has difficulty in correctly detecting actual noncircular clusters. To detect arbitrarily shaped clusters which cannot be detected by the circular spatial scan statistic, Patil and Taillie (2004), Duczmal and Assunção (2004), Tango

and Takahashi (2005), and Assunção *et al.* (2006) have proposed different spatial scan statistics. Patil and Taillie (2004) used the notion of "upper level set" to reduce the size of windows to be scanned and proposed the "upper level set scan statistic." However, they do not discuss how to select the level g which defines the upper level set. Duczmal and Assunção (2004), on the other hand, have applied a simulated annealing method, in which they try to examine only the most promising windows using a graph-based algorithm to obtain the local maxima of a certain likelihood function over a subset of the collection of all the connected regions. Their method seems to be very complicated, but they do not show any programmable procedure for it. Tango and Takahashi (2005) called their spatial scan statistic the *flexible spatial scan statistic* in contrast to Kulldorff's *circular spatial scan statistic* and provided FleXScan Software [Takahashi, Yokoyama, and Tango (2007)].

The *flexible spatial scan statistic* imposes an irregularly shaped window $\mathbf{Z}$ on each region by connecting its adjacent regions. For any given region i, we create the set of irregularly shaped windows with *length* k consisting of k connected regions including i and let k move from 1 to the preset maximum length of cluster K. To avoid detecting a cluster of *unlikely peculiar shape*, the connected regions are restricted as the subsets of the set of regions i and $(K-1)$-nearest neighbors to the region i. In total, as in the circular spatial scan statistic, a very large number of different but overlapping arbitrarily shaped windows are created. Let $\mathbf{Z}_{ik(j)},\ j = 1, \ldots, J_{ik}$ denote the jth window which is a set of k regions connected starting from the region i, where J_{ik} is the number of j satisfying $\mathbf{Z}_{ik(j)} \subseteq \mathbf{Z}_{ik}$ for $k = 1, \ldots, K_i = K$. Then, all the windows to be scanned are included in the set

$$\mathcal{Z}_2 = \{\mathbf{Z}_{ik(j)} \mid 1 \le i \le m,\ 1 \le k \le K,\ 1 \le j \le J_{ik}\}. \tag{17.10}$$

In other words, for any given region i, the circular spatial scan statistic considers K concentric circles, whereas the flexible scan statistic considers K concentric circles plus all the sets of connected regions (including the single region i) whose centroids are located within the Kth largest concentric circle. So, the size of $\mathcal{Z}_2$ is far larger than that of $\mathcal{Z}_1$, which is at most mK. Under the Poisson assumption, the test statistic is the same form as (17.9) where $\mathcal{Z}_1$ is replaced by $\mathcal{Z}_2$.

17.3.4 Clustering index

Tango (1995) proposed the following test statistic for spatial disease clustering:

$$\begin{aligned} C &= (\boldsymbol{r} - \boldsymbol{p})^t \boldsymbol{A} (\boldsymbol{r} - \boldsymbol{p}) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\frac{n_i - e_i}{n} \right) \left(\frac{n_j - e_j}{n} \right) a_{ij}, \end{aligned} \tag{17.11}$$

where $\boldsymbol{r}^t = (n_1, ..., n_m)/n$ denotes a vector of the observed relative frequencies, $\boldsymbol{p} = E_{H_0}(\boldsymbol{r})$, and $e_i = np_i,\ i = 1, ..., m$. This is a generalization of his temporal clustering index in that it allows for heterogeneous population size and confounding factors based on indirect standardization. Namely, let us partition the population into K categories and let n_{ik} and ξ_{ik} denote the observed number of cases and the population size, respectively, in the kth category of the confounding factor of the ith region. Then, we have

$$\boldsymbol{p} = \sum_{k=1}^{K} \frac{n_{+k}}{n} \boldsymbol{p}_k = \sum_{k=1}^{K} \frac{n_{+k}}{n} (p_{1k}, ..., p_{mk})^t, \tag{17.12}$$

where $p_{ik} = \xi_{ik} / \sum_{j=1}^{m} \xi_{jk}$. As a measure of closeness, $a_{ij}(\lambda)$, between the regions i and j, Tango (1995, 2000) recommended the double exponential form:

$$a_{ij} = \exp\left\{-4\left(\frac{d_{ij}}{\lambda}\right)^2\right\}, \tag{17.13}$$

where λ is a measure of *cluster size* and is essentially equal to the maximum distance between cases, such that any pair of cases far apart beyond the distance λ cannot be considered as a cluster. Large λ will give a test sensitive to a large cluster and small λ to a small cluster. In practical application, it is rare that we can predict the cluster size before examining data. Therefore, we usually repeat the procedure using different parameter settings and, consequently, face multiple testing problems. To take this problem into account, Tango (2000) propose, as an extended test statistic, *the minimum of the profile P-value of C for* λ where λ varies continuously from a small value near zero upwards until λ reaches about one-fourth the maximum distance d_{ij} in the study area. The proposed test statistic P_{min} is defined as

$$P_{min} = \min_{\lambda} \Pr\{C > c \mid H_0, \lambda\} = \Pr\{C > c \mid H_0, \lambda = \lambda^*\}, \tag{17.14}$$

where λ^* attains the minimum p-values of C. A practical implementation of this procedure is to use "line search" by discretization of λ. The null distribution of P_{min} can be obtained by using Monte Carlo simulation. This test is also called Tango's MEET (maximized excess event test) in the literature [e.g., Kulldorff *et al.* (2003, 2006); Song and Kulldorff (2003, 2005)].

Given λ and under the null hypothesis H_0, the test statistic C was shown to be asymptotically approximated by the same type of chi-square distribution as (17.4), where

$$
\begin{aligned}
E(C) &= n^{-1}tr(\boldsymbol{AV}) \\
Var(C) &= 2n^{-2}tr(\boldsymbol{AV})^2 \\
\nu &= 8/\{\sqrt{\beta_1(C)}\}^2 \\
\sqrt{\beta_1(C)} &= 2\sqrt{2}tr(\boldsymbol{AV})^3/\{tr(\boldsymbol{AV})^2\}^{3/2} \\
\boldsymbol{V} &= \sum_{k=1}^{K} \frac{n_{+k}}{n}\{\mathrm{diag}(\boldsymbol{p}_k) - \boldsymbol{p}_k\boldsymbol{p}_k^t\}.
\end{aligned}
$$

This chi-square approximation is generally quite accurate even for small n. If the null hypothesis of no clustering is rejected, we can use a statistic similar to (17.5) to indicate *the most likely center* i of clustering area with large values of

$$U_i = \frac{1}{C}\sum_{j=1}^{m}\left(\frac{n_i - e_i}{n}\right)\left(\frac{n_j - e_j}{n}\right)a_{ij}, \tag{17.15}$$

which denote the percentage of the ith region's contribution to the significant clustering. More specifically, we may use the following condition of standardized U_i to suggest the center of clustering areas:

$$(U_i - \bar{U})/\mathrm{SD}_U \geq 2.0 \text{ or } 3.0.$$

17.3.5 Other methods

Whittemore *et al.* (1987) developed a test statistic for spatial clustering,

$$W = \boldsymbol{r}^t\boldsymbol{D}\boldsymbol{r},$$

which is identical in form to Tango's clustering index C (17.3), but for which $\boldsymbol{D} = (d_{ij})$ is used as a measure of distance. They proved the asymptotic distribution of this index to be normal and insisted that the clustering index C (17.3) also has an asymptotic normal distribution. However, it does depend largely on the element $\boldsymbol{A}$ or $\boldsymbol{D}$ used. When the distance measure $\boldsymbol{D}$ is used, convergence to normality is very fast. On the contrary, when the closeness measure $\boldsymbol{A}$ is used, the speed is shown to be too slow, and thus normality is not valid even for fairly large sample sizes such as $n = 1000$ [Tango (1986, 1990)]. Furthermore, more substantially, it has been shown that (1) the quadratic form in $(\boldsymbol{r} - \boldsymbol{p})$ should be used to properly adjust for heterogeneous populations, and (2) the power of W often falls below the nominal α level depending on the clustering models due to the use of distance measure $\boldsymbol{D}$ [Tango (1995, 1999)]. Therefore, the test of Whittemore *et al.* cannot be recommended for practical use. Bonetti and Pagano (2005) proposed a test using the interpoint distance distribution for spatial clustering, but it generally does not perform quite as well as the spatial scan statistic and Tango's clustering index [Kulldorff *et al.* (2003), Song and Kulldorff (2003)].

17.3.6 Illustration with gallbladder cancer mortality data

As an illustration, we shall apply three tests, 1) circular spatial scan statistic, 2) flexible spatial scan statistic, and 3) spatial clustering index, to the mortality data from gallbladder cancer (male, 1993–1997) in the areas of three adjacent prefectures (Niigata, Fukushima, and Yamagata) in Japan. The total observed number of deaths for five years was 665 in this area with $m = 246$ regions (cities and villages). Before applying these three tests for spatial clustering, we drew a disease map based on the standardized mortality ratio (SMR) in Figure 17.1, which shows the maximum likelihood estimates for the relative risks. No clear spatial pattern emerges from this map. SMRs are commonly used in disease mapping, but they are very unstable in the sense that they can yield large changes in estimate with relatively small changes in expected number of cases. So, to overcome the drawbacks of the SMRs in disease mapping, Bayesian approaches have been used to obtain more smoothed estimates [for example, see Lawson, Browne, and Vidal Rodeiro (2003)]. In this paper we shall omit Bayes estimates of for disease mapping.

The results of Kulldorff's circular spatial scan statistic and Tango–Takahashi's flexible spatial scan statistic are shown in Figure 17.2 and Figure 17.3, respectively, where $K = 20$. The most likely cluster and the secondary cluster detected by the flexible spatial scan statistic are very similar to, but have a slightly different shape than, those of the circular spatial scan statistic. Regarding the application of Tango's clustering index, we took a sequence of

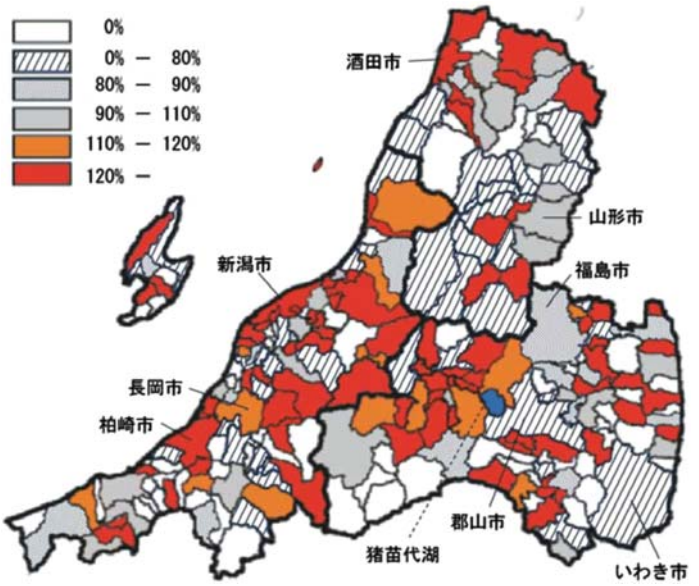


Figure 17.1. The SMRs of gallbladder cancer (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan (1996–2000).

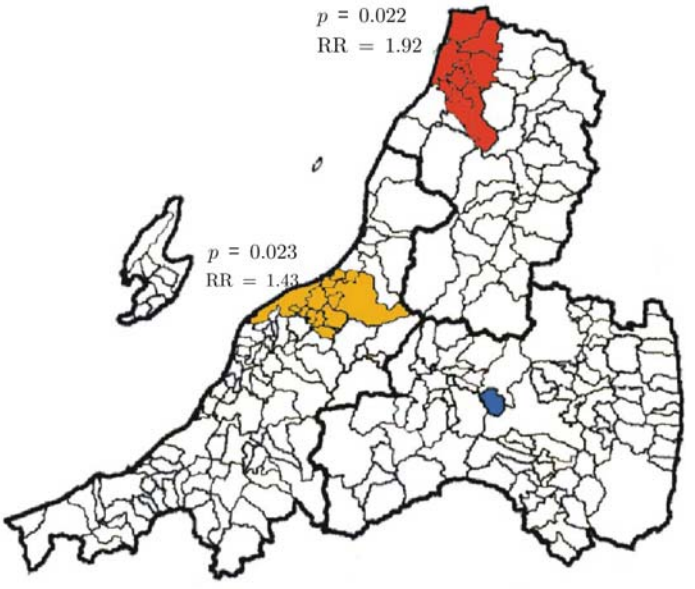


Figure 17.2. The most likely cluster (shaded area) and the secondary cluster (a lighter shaded area) detected by SaTScan for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.

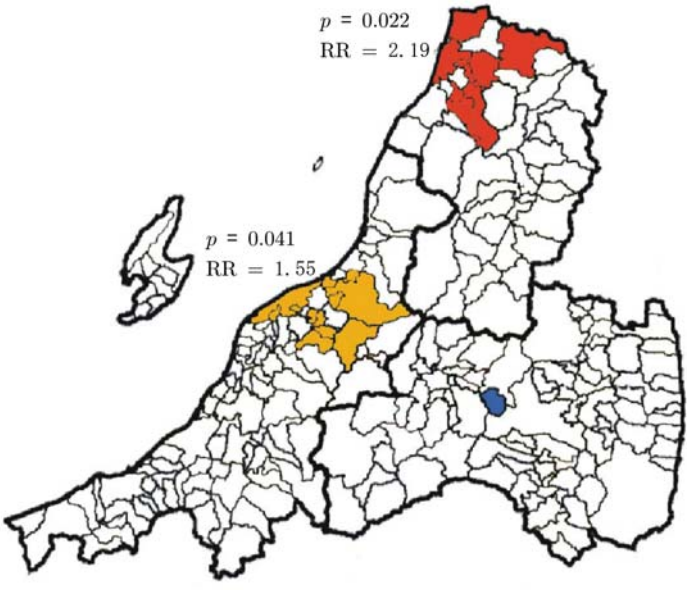


Figure 17.3. The most likely cluster (shaded area) and the secondary cluster (a lighter shaded area) detected by FleXScan for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.

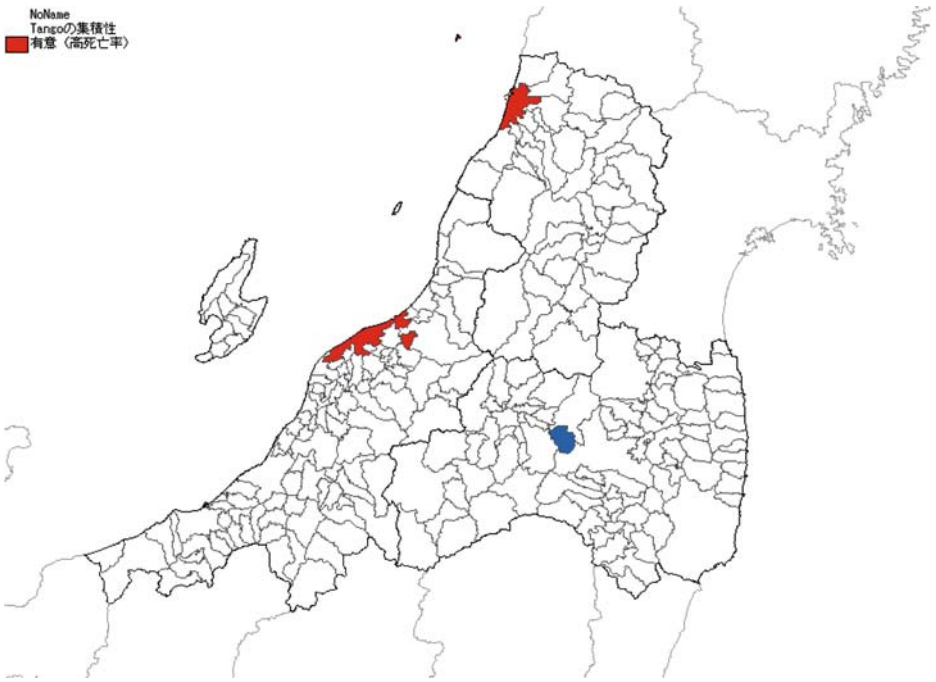


Figure 17.4. Two centers of clustering areas (shaded area) detected by Tango's spatial clustering index for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.

values of cluster size λ as $\lambda = 0.1, 5, 10, 15, ..., 100$ (km) to obtain the test statistic

$$P_{min} = \min_{\lambda \in \{0.1,5,10,...,100\}} \Pr\{C > c \mid H_0, \lambda\}$$

and obtained $P_{min} = 0.00004$ at $\lambda = 45$. This P_{min} value is the second largest among 999 Monte Carlo replicates and, therefore, the adjusted p-value of P_{min} was $2/(999+1) = 0.002$. As possible centers of clusters, regions with standardized $U_i \geq 2.0$ are indicated in Figure 17.4, and these regions are found to be included in the most likely cluster and secondary cluster detected by both the circular scan statistic and flexible scan statistic.

17.4 Discussion

Many different test statistics have been designed for detecting disease clustering in time and in space. Most tests proposed before 1995, however, suffer from multiple testing problems due to one or two unknown parameters that must

be set prior to their applications. For example, Naus's scan statistic (1965) for individual time points data has an unknown length d of the scanning window, the procedure by Turnbull *et al.* (1990) has an unknown parameter regarding the common size of the population at risk R, Cuzick and Edwards's test (1995) has an unknown number of k-nearest-neighbors, and Besag and Newell's test (1991) has an unknown number of cases k for the size of the cluster. However, tests proposed in recent years tend to take such multiple testing into account. For example, such tests include Nagarwalla's scan statistic with variable window (1996), Kulldorff's spatial scan statistic (1997), Tango and Takahashi's flexible spatial scan statistic (2005), and Tango's clustering index (2000), where we have only to specify the maximum possible cluster size.

In recent power comparisons of disease clustering tests including CDTs and GCTs by Kulldorff *et al.* (2003) and Song and Kulldorff (2003), 1) Kulldorff's circular spatial scan statistic is shown to be the most powerful for detecting localized clusters, and 2) Tango's clustering index is the most powerful for general clustering throughout the study area. Note, however, that the power estimates provided reflect only the "power to reject the null hypothesis for whatever reason" and that the probability of both rejecting the null hypothesis and detecting the true cluster correctly is a different matter. To investigate *the performance of power* of the CDT, Tango and Takahashi (2005) proposed a new bivariate power distribution $P(l, s)$, which is the probability that the significant MLC has length $l(\geq 1)$ and includes s regions within the true cluster with length s^*. The usual power is defined by $\sum_l \sum_{s=1}^{s^*} P(l, s)$. Our simulation study using $P(l, s)$ revealed that the circular spatial scan statistic shows a high level of accuracy in detecting circular clusters exactly and reasonably good power for including some true cluster regions into the MLC. However, the circular spatial scan statistic is also shown to have a tendency to detect a cluster much larger than the true cluster assumed in the simulation, even when the true cluster is circular. The flexible spatial scan statistic, on the other hand, exhibits no such high power regarding exact identification of clusters, but the support of the power distribution is shown to be concentrated in a relatively narrow range of length l on the line $s = s^*$, indicating that an observed significant MLC contains the true cluster with quite high probability.

Tango and Takahashi (2005) have also shown examples which cast a doubt on the validity of the model selection based on maximizing the likelihood ratio: Duczmal and Assunção's procedure (2004) detected a quite large and peculiar shaped MLC that had the largest likelihood ratio among the three different MLCs, identified by three different spatial scan statistics, Kulldorff's (1997), Duczmal and Assunção's (2004), and Tango and Takahashi's (2005). Such a doubt can also be seen in the above-stated simulation results of the circular spatial scan statistic that had nonnegligible probabilities of detecting much longer clusters, than the true cluster. The flexible spatial scan statistic, on the other hand, is shown not to detect such an unexpected long cluster, probably

because it has the restriction that our windows are constructed only from members of the $(K-1)$-nearest neighbors to the starting region. Nevertheless, these undesirable properties produced by the maximum likelihood ratio might suggest the use of a different criterion for model selection.

In this chapter, we did not include tests for space-time disease clustering due to the limitation of space. As far as I know, Kulldorff (2001) proposed a procedure for prospective time periodic geographical disease surveillance using a scan statistic for the first time. In the aftermath of the World Trade Center attacks on September 11, 2001 and the anthrax-laden letters that followed in October 2001, a syndromic surveillance has been poised for deployment across the USA [Lawson and Kleinman, (2005)]. Therefore, statistical methods for timely detection of an outbreak threat, which are closely related to tests for space-time clustering, will be increasingly needed.

References

1. Assunção R., Costa M., Tavares A. and Ferreira S. (2006). Fast detection of arbitrarily shaped disease clusters, *Statistics in Medicine*, **25**, 723–742.

2. Bailar III J.C., Eisenberg H. and Mantel N. (1970). Time between pairs of leukemia cases, *Cancer*, **25**, 1301–1303.

3. Besag J.E. and Newell J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A*, **154**, 143–155.

4. Bonetti M. and Pagano M. (2005). The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering, *Statistics in Medicine*, **24**, 753–773.

5. Cuzick J.C. and Edwards R. (1990). Spatial clustering for inhomogeneous populations, *Journal of the Royal Statistical Society, Series B*, **52**, 73–104.

6. Duczmal L. and Assunção R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped clusters, *Computational Statistics and Data Analysis*, **45**, 269–286.

7. Dwass M. (1957). Modified randomization test for nonparametric hypotheses, *Annals of Mathematical Statistics*, **28**, 181–187.

8. Ederer F., Myers M.H. and Mantel N. (1964). A statistical problem in space and time: do leukemia cases come in clusters? *Biometrika*, **20**, 626–638.

9. Geary R.C. (1954). The contiguity ratio and statistical mapping, *The Incorporated Statistician*, **5**, 115–145.

10. Grimson R.C., Wang K.C. and Johnson P.W.C. (1981). Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome, *Social Science & Medicine*, **15D**, 287–293.

11. Knox G. (1959). Secular pattern of congenital oesophageal atresia, *British Journal of Preventive Social Medicine*, **13**, 222–226.

12. Knox E.G. and Lancashire R. (1982). Detection of minimal epidemics, *Statistics in Medicine*, **1**, 183–189.

13. Kulldorff M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

14. Kulldorff M. (1998). Statistical methods for spatial epidemiology: tests for randomness. In *GIS and Health*, (Ed., Gatrell A. and Loytonen M.), 49–62, Taylor & Francis, London.

15. Kulldorff M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society, Series A*, **164**, 61–72.

16. Kulldorff M. (2006). Tests for spatial randomness adjusted for an inhomogeneity: a general framework, *Journal of American Statistical Association*, **101**, 1289–1305.

17. Kulldorff M. and Nagarwalla N. (1995). Spatial disease clusters: detection and inference, *Statistics in Medicine*, **14**, 799–810.

18. Kulldorff M., Tango T. and Park P.J. (2003). Power comparisons for disease clustering tests, *Computational Statistics and Data Analysis*, **42**, 665–684.

19. Kulldorff M. and Information Management Services, Inc. (2007). SaTScan v7.0: Software for the spatial and space-time scan statistics, http://www.satscan.org/

20. Larsen R.J., Holmes C.L. and Heath C.W. (1973). A statistical test for measuring unimodal clustering: a description of the test and of its application to cases of acute leukemia in metropolitan Atlanta, Georgia, *Biometrics*, **29**, 301–309.

21. Lawson A.B., Browne W.J. and Vidal Rodeiro C.L. (2003). *Disease Mapping with WinBUGS and MLwiN*, John Wiley & Sons, Chichester.

22. Lawson A.B. and Kleinman K. (eds.) (2005). *Spatial & Syndromic Surveillance for Public Health*, John Wiley & Sons, New York.

23. Mantel N., Kryscio R.J. and Myers M.H. (1976). Tables and formulas for extended use of the Ederer-Myers-Mantel disease clustering procedure, *American Journal of Epidemiology*, **104**, 576–584.

24. Molinari N., Bonaldi, C. and Daures, J.P. (2001). Multiple temporal cluster detection. *Biometrics*, **57**, 577–583.

25. Nagarwalla N. (1996). A scan statistic with a variable window. *Statistics in Medicine*, **15**, 845–850.

26. Naus J.I. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association*, **60**, 532–538.

27. Naus J.I. (1966). A power comparison of two tests of non-random clustering, *Technometrics*, **8**, 493–517.

28. Ohno Y., Aoki K. and Aoki N. (1979). A test of significance for geographic clusters of disease, *International Journal of Epidemiology*, **8**, 273–281.

29. Ohno Y. and Aoki K. (1981). Cancer deaths by city and county in Japan: a test of significance for geographic clustering of disease, *Social Science & Medicine*, **15D**, 251–258.

30. Openshaw S., Craft A.W., Charlton M. and Birth J.M. (1988). Investigation of leukemia clusters by use of a geographical analysis machine, *Lancet*, **1(8580)**, 272–273.

31. Patil G. P. and Taillie C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

32. Song C. and Kulldorff M. (2003). Power evaluation of disease clustering tests, *International Journal of Health Geographics*, **2**, 9.

33. Song C. and Kulldorff M. (2005). Tango's maximized excess events test with different weights, *International Journal of Health Geographics*, **4**, 32.

34. Takahashi K., Yokoyama T. and Tango T. (2007). FleXScan: Software for the Flexble Scan Statistic. v2.0. http://www.niph.go.jp/soshiki/gijutsu/index_e.html/.

35. Tango T. (1984). The detection of disease clustering in time, *Biometrics*, **40**, 15–26.

36. Tango T. (1990). Asymptotic distribution of an index for disease clustering, *Biometrics*, **46**, 351–357.

37. Tango T. (1995). A class of tests for detecting "general" and "focused" clustering of rare diseases, *Statistics in Medicine*, **14**, 2323–2334.

38. Tango T. (1999). Comparison of general tests for disease clustering, In *Disease Mapping and Risk Assessment for Public Health*, (Ed., A.B. Lawson *et al.*), pp. 111–117, Wiley & Sons, New York.

39. Tango T. (2000). A test for spatial disease clustering adjusted for multiple testing, *Statistics in Medicine*, **19**, 191–204.

40. Tango T. and Takahashi K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11. http://www.ij-healthgeographics.com/content/4/1/11.

41. Tango T. (2007). A class of multiplicity adjusted tests for spatial clustering based on case-control point data, *Biometrics*, **63**, 119–127.

42. Turnbull B.W., Iwano E.J., Burnnett W.S., Howe H.L. and Clark LC. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York, *American Journal of Epidemiology*, **132**, suppl. S136–143.

43. Wallenstein S. (1980). A test for detection of clustering over time, *American Journal of Epidemiology*, **111**, 367–372.

44. Wallenstein S. and Neff N. (1987). An approximation for the distribution of the scan statistic. *Statistics in Medicine*, **6**, 197–207.

45. Weinstock M.A. (1981). A generalized scan statistic test for the detection of clusters, *International Journal of Epidemiology*, **10**, 289–93.

46. Whittemore A. and Keller J.B. (1986). A letter to the editor. On Tango's index of disease clustering in time, *Biometrics*, **42**, 218.

47. Whittemore A.S., Friend N., Brown B.W. and Holly E.A. (1987). A test to detect clusters of disease, *Biometrika*, **74**, 631–635.

Index