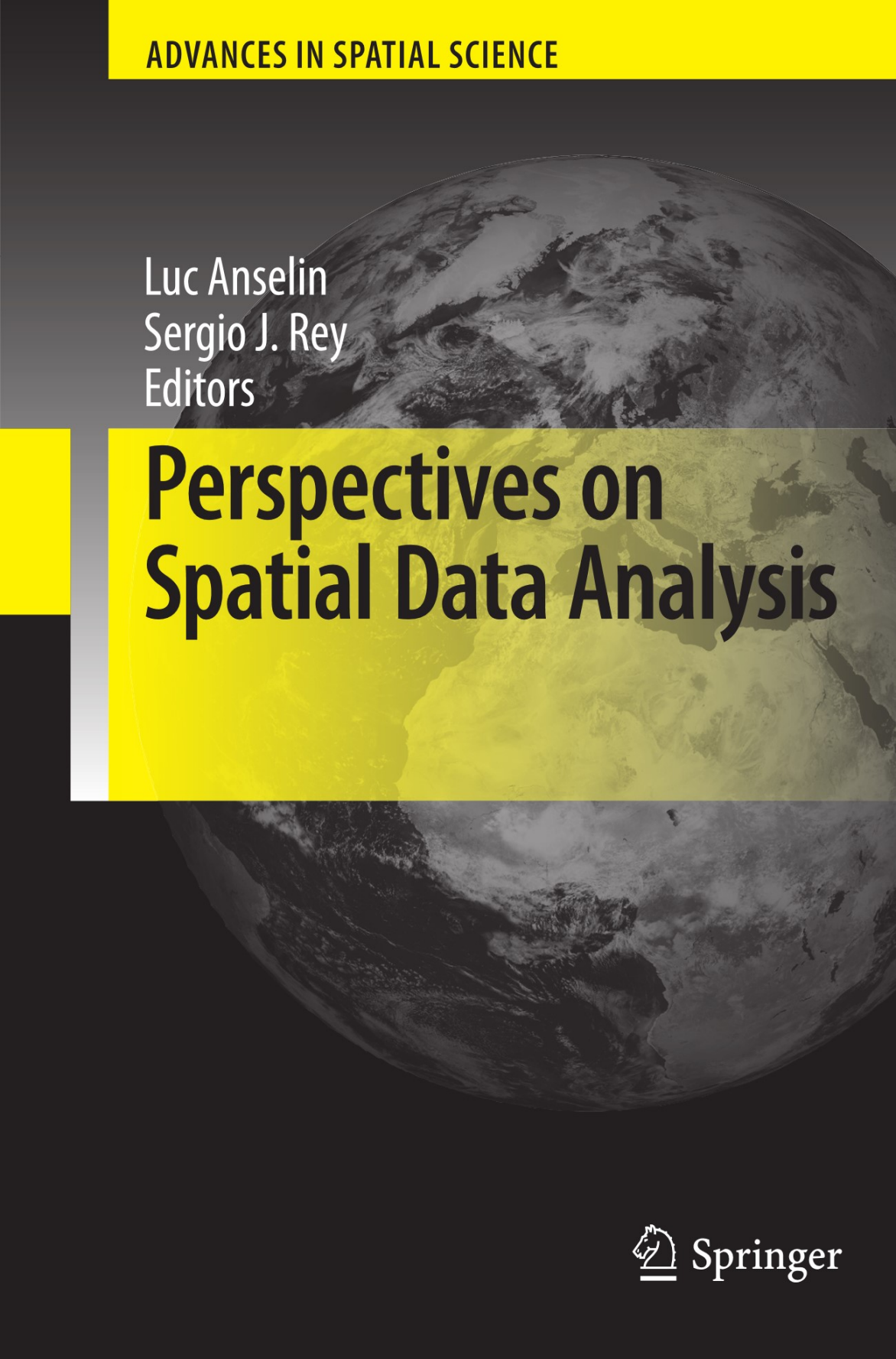
ADVANCES IN SPATIAL SCIENCE



Luc Anselin
Sergio J. Rey
Editors

Perspectives on Spatial Data Analysis

Springer

Advances in Spatial Science

For further volumes:
http://www.springer.com/series/3302

Arthur Getis

Luc Anselin • Sergio J. Rey
Editors

Perspectives on Spatial Data Analysis

Springer

Editors

Prof. Dr. Luc Anselin
Prof. Dr. Sergio J. Rey
School of Geographical Sciences
Arizona State University
Coor Hall
975 S Myrtle Avenue
Tempe, AZ 85287-5302
USA
Luc.Anselin@asu.edu
Sergio.Rey@asu.edu

ISBN 978-3-642-01975-3 e-ISBN 978-3-642-01976-0
DOI: 10.1007/978-3-642-01976-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009931324

Cover design: SPi Publisher Services

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Arthur Getis

Preface

Spatial data analysis has seen explosive growth in recent years. Both in mainstream statistics and econometrics as well as in many applied fields, the attention to space, location, and interaction has become an important feature of scholarly work. The methods developed to deal with problems of spatial pattern recognition, spatial autocorrelation, and spatial heterogeneity have seen greatly increased adoption, in part due to the availability of user friendly desktop software. Through his theoretical and applied work, Arthur Getis has been a major contributing figure in this development.

In this volume, we take both a retrospective and a prospective view of the field. We use the occasion of the retirement and move to emeritus status of Arthur Getis to highlight the contributions of his work. In addition, we aim to place it into perspective in light of the current state of the art and future directions in spatial data analysis.

To this end, we elected to combine reprints of selected classic contributions by Getis with chapters written by key spatial scientists. These scholars were specifically invited to react to the earlier work by Getis with an eye toward assessing its impact, tracing out the evolution of related research, and to reflect on the future broadening of spatial analysis. The organization of the book follows four main themes in Getis' contributions:

- Spatial analysis
- Pattern analysis
- Local statistics
- Applications

For each of these themes, the chapters provide a historical perspective on early methodological developments and theoretical insights, assessments of these contributions in light of the current state of the art, as well as descriptions of new techniques and applications.

Putting together a volume such as this would not be possible without the efforts of many individuals. We feel most fortunate to have been in the skilled hands of the Springer-Verlag team and in particular wish to extend our gratitude to Katharina Wetzel-Vandai and Barbara Fess for their continued support during this project and to Manfred Fischer for his editorial suggestions. We are indebted to the authors of both the original pieces as well as the new contributions and the referees. The

project benefited enormously from the technical typesetting skills of Xinyue Ye of the Department of Geography at San Diego State University and David Folch of the School of Geographical Sciences at Arizona State University, without whose dedicated efforts this volume would not have been completed.

The Spatial Analysis Laboratory at the University of Illinois Champaign-Urbana and the Department of Geography at San Diego State University both provided institutional support during the early phases of this project. The GeoDa Center for Geospatial Analysis and Computation in the School of Geographical Sciences at Arizona State University provided critical support to bring the project to closure.

Finally, we would like to dedicate this volume to Arthur Getis whose contributions have impacted so many. We feel fortunate to not only count ourselves among those, but also to call him a valued friend.

Tempe, AZ, USA *Luc Anselin*
August 2009 *Sergio Rey*

Foreword

Born in Philadelphia in 1934 Arthur Getis received his undergraduate degrees from Pennsylvania State University and his Ph.D. in Geography from the University of Washington in 1964. Until his retirement, he held the Stephen and Mary Birch Foundation Endowed Chair on Geographical Studies at San Diego State University. Prior to joining San Diego University, he was the Chairman of the Geography Department and Director of the School of Social Sciences at the University of Illinois. He has also served on the board, or as chairman of scientific societies, most notably the Regional Science Association.

Getis' research studies in social science have included work in location theory, urban geography, mathematical pattern analysis, spatial analysis, geographical information science, and most recently, clustering analysis of disease or crime. Some of this material is reflected in his books: Models of Spatial Processes – Approaches to the Study of Point, Line, and Area Patterns (with B. Boots, Cambridge University Press 1978); Point Pattern Analysis (with B. Boots, Sage Publications, 1987); The Tyranny of Data (San Diego State University Press, 1995); Recent Developments in Spatial Analysis – Spatial Statistics, Behavioral Modeling, and Computational Intelligence (with M. Fischer, Springer 1997); and Spatial Econometrics and Spatial Statistics (edited with others, Palgrave, 2003). Related work has been published in over 80 articles and in worldwide lectures. With J. Getis, he has also contributed to geographic education in the United States.

To summarize, Arthur Getis' outstanding contributions have been to the rigorous study of spatial patterns including the effect of autocorrelation and the introduction of local statistics. The latter include numerous original developments that allow the recognition of variations from place to place – in contradistinction to more conventional global analyses that are not aware of spatial detail. The introduction of these methods has revolutionized recent spatial analysis. For this, we need to thank the editors Luc Anselin and Sergio Rey.

Santa Barbara, CA, USA
January 2009

Waldo Tobler

Contents

1 Perspectives on Spatial Data Analysis 1
Luc Anselin and Sergio J. Rey

Part I Spatial Analysis

2 Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach 23
Arthur Getis

3 Spatial Statistical Analysis and Geographic Information Systems 35
Luc Anselin and Arthur Getis

4 Whose Hand on the Tiller? Revisiting "Spatial Statistical Analysis and GIS" 49
Michael F. Goodchild

5 Spatial Interaction and Spatial Autocorrelation 61
Manfred M. Fischer, Martin Reismann, and Thomas Scherngell

Part II Pattern Analysis

6 Second-Order Analysis of Point Patterns: The Case of Chicago as a Multi-center Urban Region 83
Arthur Getis

7 Second-Order Neighborhood Analysis of Mapped Point Patterns 93
Arthur Getis and Janet Franklin

8 A Class of Local and Global K Functions and Their Exact Statistical Methods 101
Atsu Okabe, Barry Boots and Toshiaki Satoh

9 Spatial Point Pattern Analysis of Plants 113
Janet Franklin

Part III Local Statistics

10 The Analysis of Spatial Association by Use of Distance Statistics 127
Arthur Getis and J. Keith Ord

11 Constructing the Spatial Weights Matrix Using a Local Statistic 147
Arthur Getis and Jared Aldstadt

12 Spatial Autocorrelation: A Statistician's Reflections 165
J. Keith Ord

13 Health Surveillance Around Prespecified Locations Using Case-Control Data 181
Peter A. Rogerson

Part IV Applications

14 Spatial Filtering in a Regression Framework: Examples Using Data on Urban Crime, Regional Inequality, and government Expenditures 191
Arthur Getis

15 Characteristics of the Spatial Pattern of the Dengue Vector, *Aedes aegypti*, in Iquitos, Peru 203
Arthur Getis, Amy C. Morrison, Kenneth Gray, and Thomas W. Scott

16 Spatial Filtering and Missing Georeferenced Data Imputation: A Comparison of the Getis and Griffith Methods 227
Daniel Griffith

17 Spatial Patterns of Fertility in Rural Egypt 235
John R. Weeks

References 257

Author Index 275

Index 283

Contributors 287

List of Tables

1.1 Samples of empirical work by Getis in epidemiology 8
1.2 Most cited articles 9
2.1 A comparison of various spatial models and the cross-product statistic 26
5.1 The log-additive spatial interaction model: parameter estimates and performance measures ($N = 12{,}432$) 75
5.2 Spatial econometric flow models based on different spatial weights matrix specifications: ML estimates using Ng and Peyton's Cholesky algorithm ($N = 12{,}432$ observations) 78
10.1 Characteristics of G_i statistics 130
10.2 Standard normal variates for $G(d)$ and $I(d)$ under varying circumstances for a specified d value 137
10.3 Spatial association among counties: SIDS rates by county in North Carolina, 1979–1984 139
10.4 Highest positive and negative standard normal variates by county for $G_i^*(d)$ and $G_i(d)$: SIDS rates in North Carolina, 1979–1984 ($d = 33$ miles) 139
10.5 Spatial association among zip code districts: dwelling unit prices in San Diego county, September 1989 142
10.6 Highest positive and negative standard normal variates by zip code district for $G_i^*(d)$ and $G_i(d)$: dwelling unit prices in San Diego county, September 1989 ($d = 5$ miles) 142
11.1 Data set descriptions 152
11.2 AIC results 159
11.3 Estimated autocorrelation coefficient values 160
11.4 Moran's Z(I) of residuals 161
12.1 Asymptotic relative efficiencies for different patterns of weights 171
12.2 Relative magnitudes of the G and LISA coefficients: major differences are in bold and moderate differences are in italics 173
15.1 Summary of clustering statistics 210
15.2 $L(d)$ values for distances 10–100 m for houses and adult mosquitoes in Maynas a* 211

15.3 $L(d)$ values for distances 10–100 m for houses and adult mosquitoes in Maynas a* 212
15.4 $\hat{L}(d)$ values for distances 10–100 for houses, adult mosquitoes, pupae, water-holding containers, positive water-holding containers in Maynas a* 215
15.5 $\hat{L}(d)$ values for distances 10–100 m for houses, adult mosquitoes, pupae, water-holding containers, positive water-holding containers in Maynas a* 215
15.6 Number of members of clusters in Maynas and Tupac Amaru in time periods a and b 217
15.7 One or more adult mosquitoes and/or pupae present in houses in Maynas and Tupac Amaru in time periods a and b 218
15.8 Spearman's rank correlations of the number of containers per house with the number of mosquitoes and pupae per house 219
15.9 $\hat{L}(d)$ values for 10 m for Maynas and Tupac Amaru for time periods a and b* 219
16.1 Sources of data for using (16.1) to construct Fig. 16.1 228
16.2 Selected model-based imputations for two selected empirical data sets containing missing values: Pennsylvania coal ash, and vandalized turnip field plots 232
16.3 Comparisons of the two spatial filter estimators 233
17.1 Fertility decline in Egypt and Menoufia, 1976–1996 245
17.2 Regression models for fertility in Menoufia, 1976 247
17.3 Regression models for fertility in Menoufia, 1986 249
17.4 Regression models for fertility in Menoufia, 1996 250
17.5 Regression models of fertility change between 1976 and 1996 252
17.6 Results for Menoufia from the 1988 and 1995 Egyptian demographic and health surveys 255

List of Figures

1.1 Annual citation patterns for Arthur Getis 8
1.2 Concept co-citation network 9
1.3 Author co-citation network 10
1.4 Journal co-citation network 11
3.1 Functions of a GIS 38
5.1 Origin-based and destination-based similarity. The flows (i, j) and (r, s) are origin-based similar in *Case A* since the origin regions i and r are contiguous spatial units, and destination-based similar in *Case B* since the destination regions j and s are contiguous spatial units 66
5.2 Flows with selected significant $G_{ij}({}^{o}\boldsymbol{W})$ statistic: (a) indicating high residual flows from neighbours of origin i to the destination region Île-de-France; and (b) indicating low residual flows from neighbours of the region Leipzig to a destination region j 76
6.1 A border correction is needed when the distance xy is greater than the distance of x to the nearest border 85
6.2 Population distribution in the Chicago region 88
6.3 A plot of $\hat{L}$ for the population of the Chicago region. The *straight line* is the mean for a Poisson process. The portion of the diagram from $t = 0$ to $t = 0.06$ is enlarged on Fig. 6.4 89
6.4 A plot of $\hat{L}$ for the population of the Chicago region. The *dashed lines* are the 95% confidence bands of the Poisson process 90
7.1 Cumulative distribution curve (*heavy line*) of $\hat{L}_i(d)$ for hypothetical tree in a square of area 1. $L_i(d)$ is the number of points within distance d of point i corrected for the boundary effect, and scaled such that $L_i(d) = d$ when $L_i(d)$ represents a pattern produced by a Poisson process in the plane. *Dashed lines* represent 0.01 significance levels around the line representing Poisson process 95
7.2 Point pattern representation of tree locations in the study area. The letters A, B, and C mark particular individual trees, which are referred to in Fig. 7.4. North is up 97

7.3 Values for $\hat{L}(d)$ over the range $0.01 \leq d \leq 0.30$. $L(d)$ is the number of points within distance d of all points i corrected for the boundary effect, and scaled such that $L(d) = d$ when $L(d)$ represents a pattern produced by a Poisson process in the plane. $\hat{L}(d)$ may be interpreted as the average for all 108 points (from Fig. 7.2) taken together. *Solid line* shows expected values given a Poisson distribution. *Solid dots* show observed values..... 97
7.4 Values for $\hat{L}_i(d)$ over the range $0.01 \leq d \leq 0.30$ when $i = A$ 98
7.5 Pattern created by assigning to each tree its $\hat{L}(d)$ value..... 99
8.1 Points of P (the *white circles*), points of Q (the *black circles*) and a disk $D_i(t)$ centered at a point of Q with radius t 102
8.2 The edge effect: (a) the first case, (b) the second case, (c) adjustment ..103
8.3 Disks are not overlapped, $t \leq t_{max}$ (a), and they are overlapped, $t > t_{max}$ (b).....105
8.4 Local Voronoi cross K function (a) and global Voronoi cross K function (b).....107
9.1 Locations of 440 Torrey pine trees (*Pinus torreyana*) in the East Grove area of Torrey Pines State Reserve, La Jolla, CA, USA (E. Santos and J. Franklin, unpublished data). *Map on left* shows tree locations scaled by size (DBH, trunk diameter at 1.3 m height), and *center map* shown the tree locations scaled by the local value of L(t) at lag of 14 m (see Fig. 9.2). Negative values shown as *squares*, positive values as *circles*. *Map on right* shown tree locations scaled by the values of local Gi* (see text) where neighborhood contiguity is based on a lag distance of 25 m (maximum nearest neighbor distance used to avoid islands). These analyses were carried out using the spatstat package in the R statistical environment (R Development Core Team, 2004) 121
9.2 Global $L(t)$ for the trees shown in Fig. 9.1 at lags of 2–100 m showing significant clumping at all scales and a peak in $L(t)$ at 10–18 m; (b) Moran's I as a measure of spatial autocorrelation of tree size (DBH, see Fig. 9.1 caption) where neighborhood contiguity is based on a lag distance of 10 m, indicating significant positive spatial association of tree size at lag 1 (10 m). These analyses were carried out using the splancs package (Rowlingson and Diggle, 1993) in the R statistical environment (R Development Core Team, 2004) 122
10.1 Sudden infant death rates for counties of North Carolina, 1979–1981 138

10.2 $Z[G_i^*(d = \text{furthest nearest neighbor} = 33 \text{ miles})]$ for SIDS rates of counties of North Carolina, 1979–1984 140
10.3 San Diego house prices, September 1989 141
10.4 $Z[G_i^*(d = \text{furthest nearest neighbor} = 5 \text{ miles})]$ for house prices of San Diego county zip code districts, September 1989 143
11.1 Random data set. Shading values are in random normal deviates 153
11.2 Two cluster data set. Shading values are in random normal deviates 153
11.3 Six cluster data set. Shading values are in random normal deviates 154
11.4 d_c's calculated for data sets in Figs. 11.1, 11.2, and 11.3. Distances are based on one unit separating centers of rook's case neighbors 155
12.1 True and test pattern for an interior cell on a regular grid 170
12.2 Normal probability plot for G statistics with two clusters of extreme values (Minitab plot)....174
12.3 Example of a partitioned study area 176
12.4 Components of spatial model building 179
14.1 There is no discernible spatial pattern of the residuals of the trial equation. One might get the impression from this that there is no spatial autocorrelation in the data 198
14.2 The dependent variable. Log Y, is high in the west and low in the east. AH other variables in the trial equation act similarly ...199
15.1 Map of Iquitos, Peru and location of the Maynas and Tupac Amaru study areas 206
15.2 Mosquitoes per house in the Maynas a study 213
15.3 Clusters of *Aedes aegypti* adults in the Maynas a study based on the number of mosquitoes in houses 214
15.4 Clusters of *Aedes aegypti* adults in the Maynas a study based on presence or absence of mosquitoes....214
15.5 Clusters of *Aedes aegypti* pupae in the Maynas a study based on presence or absence of pupae....217
16.1 Scatter plot of EM algorithm results: reported published vs. (16.1) generated estimates 229
16.2 Spherical model plots (denoted by *asterisks*) superimposed on experimental semivariograms (denoted by *solid circles*). (a) *left:* coal ash data ($n = 208$). (b) *right:* vandalized turnip field plot residuals ($n = 33$) 232
17.1 The study site of Menoufia governorate, Egypt....238
17.2 Situations improved by dasymetric mapping 243
17.3 Fertility levels in Menoufia in 1976 246
17.4 Spatial pattern of fertility in 1986 248
17.5 Spatial pattern of fertility in 1996 249
17.6 Spatial pattern of fertility change between 1976 and 1996....251
17.7 Spatial pattern of illiteracy change between 1976 and 1996 253

Chapter 1
Perspectives on Spatial Data Analysis

Luc Anselin and Sergio J. Rey

1.1 Introduction

This volume is inspired by the many contributions of Arthur Getis to the field of spatial analysis. In 2004, Arthur Getis formally retired as the Stephen and Mary Birch Foundation Chair of Geographical Studies in the Department of Geography at San Diego State University. That transition to emeritus status marked the end of a magnificent career spanning more than four decades. It started with undergraduate education in geography at Pennsylvania State University, followed by a PhD from the University of Washington in 1961. At Washington, he was part of the generation that initiated the "quantitative revolution" in geography under the tutelage of William Garrison. His graduate cohort included, among others, Brian Berry, Waldo Tobler, Duane Marble, John Nystuen, Richard Morrill and William Bunge. His academic appointments started with a position at Michigan State University, after which he moved to Rutgers University. He went on to become head of the Geography Department at the University of Illinois in 1977, and joined San Diego State University in 1989. In addition, he held many visiting scholar appointments at leading international institutions, including Cambridge University and the University of Bristol in the UK and the University of California, Santa Barbara and Harvard University in the USA.

During his career, Arthur Getis was awarded several honors and distinctions, such as the 1995 Albert Johnson Research Lecture at San Diego State University (captured in Getis, 1995c), the Walter Isard Award from the North American Regional Science Council (1997), the Robinson Lecture at The Ohio State University (1999), and the 2002 Distinguished Scholarship Award from the Association of American Geographers (AAG). In 2005, he was elected Fellow of the Regional Science Association International. He served as president of the Western Regional Science

L. Anselin (✉) and S. J. Rey
GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences, Arizona State University, Tempe, AZ 85287-0104, USA
e-mail: Luc.Anselin@asu.edu, Sergio.Rey@asu.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*, Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_1,

Association (1999) and of the University Consortium of Geographic Information Science (2002).

Arthur Getis was instrumental in the growth and increased exposure of spatial analytical research in institutional contexts in the discipline of Geography, through his leadership of the Mathematical Methods and Quantitative Methods (MMQM) specialty group of AAG and his role as driving force, secretary and organizer (jointly with Manfred Fischer) of the Commission on Mathematical Models of the International Geographical Union. The latter led to the establishment of a new specialized journal, *Geographical Systems* (now entitled *Journal of Geographical Systems*), which he co-edited until the end of 2007.

The activities of the IGU Commission on Mathematical Models included organizing several workshops and conference sessions that led to a number of journal special issues which he edited or co-edited. This includes an issue of *Geographical Analysis* (1992) and two issues of the *Papers in Regional Science* (1993, 1999). Edited volumes summarizing the state of the field include a collection of essays co-edited with Manfred Fischer on spatial statistics, behavioral modeling and computational intelligence (Fischer and Getis, 1997b), and a compendium on spatial econometrics and spatial statistics, co-edited with Jesus Mur and Henry Zoller (Getis et al., 2004b).

Arthur Getis' contributions were not limited to research in spatial analysis and regional science (on which we further elaborate in the next section), but he also felt strongly about promoting a rigorous approach to introductory geography. This is reflected in several textbooks he published, most jointly authored with Judy Getis. Classics are their *Introduction to Geography* (Getis et al., 2008), now in its eleventh edition (with the first edition in 1981), and *Human Geography* (Fellmann et al., 2008), now in its tenth edition (first edition in 1985). In addition, with Judy Getis, in 1995, he edited a regional geography text dealing with the United States and Canada, now in its second edition (Getis et al., 2001). His strong societal engagement is also reflected in five letters to the editor published in the *New York Times*.

With this volume, we aim to achieve two objectives. First, we want to honor Arthur Getis and his distinguished career and highlight its influence on the field of spatial data analysis. To this end, we have selected a small number of particularly important articles by Getis to reprint here. The second objective is to demonstrate the enduring effect of these early ideas on the current state of the art. We have therefore invited a number of leading scholars to comment on these "classics," with an eye towards the frontiers of the field. These contributions provide some novel perspectives and are intended to stimulate further research on cutting edge problems in spatial data analysis.

In the remainder of this introductory chapter, we first briefly review the main research contributions by Arthur Getis, with a particular focus on his work in spatial data analysis. We next go on to further explore and quantify the importance of these contributions and illustrate how they continue to affect current research in a range of scientific disciplines. We close with the customary outline of the remaining contents of the volume.

1.2 Getis Perspectives on Spatial Analysis

Arthur Getis's first peer reviewed article dealt with "a geographical analysis of rail freight shipments in Pennsylvania" and was published in the *The Pennsylvania Business Survey* (Getis, 1957). This interest in economic geography was also typical of much of his other early work, covering topics such as retail location and urban land use. Examples are "the determination of the location of retail activities with the use of a map transformation," which appeared in *Economic Geography* (Getis, 1963), and "retail store spatial affinities" (with Judy Getis), published in *Urban Geography* (Getis and Getis, 1968). However, even in these early studies, his interest in quantitative approaches towards understanding patterns and the application of the methodology of point pattern analysis was starting to become apparent. For example, "temporal land use pattern analysis with the use of nearest neighbor and quadrat method," which appeared in the *Annals of the Association of American Geographers* (Getis, 1964) applied the state of the art techniques of the time (nearest neighbor methods and quadrat counts) to urban land use analysis.

These initial ventures into quantitative and mathematical geography led to a wide ranging collection of writings that can be classified into four broad areas: spatial analysis (in general), pattern analysis, local spatial statistics, and empirical applications of spatial data analysis methods. We will also use these four categories as the structure for the essays included in this collection. In the remainder of this section, we briefly review a selection of Arthur Getis's main publications organized along these lines.

1.2.1 Spatial Analysis

Getis's contributions to the broad field of spatial analysis can be grouped into three specific categories: the situation of spatial autocorrelation within the context of spatial interaction theory, the linkages between GIS and spatial data analysis, and the general role of mathematical models in geographical analysis.

In "spatial interaction and spatial autocorrelation: a cross-product approach," which appeared in *Environment and Planning A* (Getis, 1991) and is included in this volume as Chap. 2, the basic analogy is outlined between several spatial autocorrelation statistics and the mathematical formalism of spatial interaction models. This builds upon his earlier work on second order statistics and is one of the precursors for a new general spatial autocorrelation statistic (G_i), subsequently further elaborated in the specific context of local spatial autocorrelation statistics (see also Sects. 1.2.2 and 1.2.3). More recently, his thoughts on the importance of spatial autocorrelation analysis in Regional Science are expressed in "reflections on spatial autocorrelation," which appeared in *Regional Science and Urban Economics* (Getis, 2007).

In a number of essays and edited volumes, the importance of the linkage between Geographic Information Systems (GIS) and analytical methods is argued. The article on "spatial statistical analysis and geographic information systems," in the *Annals of Regional Science* (Anselin and Getis, 1992) is included in this volume as

Chap. 3. It outlines a general framework to incorporate both exploratory and confirmatory spatial data analysis within a GIS. The collection of essays included in "Advances in Spatial Analysis" (Fischer and Getis, 1997a) further elaborates upon this theme and presents several empirical examples.

Specific attention to the nature of geographical data and particularly the way in which the increasing prevalence of large data sets will affect spatial analysis is given in the delightful booklet on "The Tyranny of Data" (Getis, 1995c), which includes materials from his 1995 Albert Johnson Research Lecture at San Diego State University. An extension of these ideas appeared in "some thoughts on the impact of large data sets on Regional Science," which appeared in the *Annals of Regional Science* (Getis, 1999). More recently, Getis was also a co-author of the research agenda for geographic information science of the University Consortium for Geographic Information Science (UCGIS), which outlined ideas on the research frontier for "spatial analysis and modeling in a GIS environment" (Getis et al., 2004a).

Getis has been a constant commentator on the importance of mathematical models and formal analysis in geography, both theoretical as well as applied. In 1993, he devoted a special issue of the *Papers in Regional Science* to mathematical models in geography, based on papers presented at a workshop organized by the IGU Commission on Mathematical Models (Getis, 1993a). In an thought provoking contribution to *Urban Geography*, he commented on "scholarship, leadership, and quantitative methods" (Getis, 1993b). More recently, he contributed to the volume on *Applied Geography: A World Perspective*, (Bailly and Gibson, 2004) with a chapter that outlined his views on "the role of geographic information science in applied geography" (Getis, 2004b).

1.2.2 Pattern Analysis

In 1978, Art Getis and his former student Barry Boots published an important text on "Models of Spatial Processes" (Getis and Boots, 1978), which introduced formal pattern analysis (and especially point pattern analysis) to quantitative geographers. The book reviewed a number of mathematical models that generate point patterns with specific characteristics (such as clustering or inhibition) and provided several test statistics (such as nearest neighbor tests) to assess the extent to which this was present in empirically observed point locations. About ten years later, a more accessible version appeared in the Sage series on scientific geography (the "brown" Sage series) as Boots and Getis (1987).

His early work in this area was primarily a survey of existing techniques developed in statistics and applications of these within geography. However, in the early 1980s Getis started to explore the properties of second order statistics in a number of articles, which eventually led to the local G_i and G_i^* statistics (see Sect. 1.2.3). The particular focus of second order statistics moves from the density of the points (first order) to the information contained in all the inter-point distances. The latter can be related to the structure of the covariance of point processes. This was initially outlined by Ripley (among others, in Ripley, 1976, 1981) and formally expressed in the by now familiar K and L statistics ("Ripley's K").

In two related articles, one in *The Professional Geographer*, "second-order analysis of point patterns: the case of Chicago as a multi-center urban region" (Getis, 1983, included as Chap. 6), the other in *Environment and Planning A*, "interaction modeling using second-order analysis" (Getis, 1984), Getis begins to make the move away from the pure point pattern focus in the original K statistic towards a broader context of spatial autocorrelation and spatial interaction (see also Sect. 1.2.1). Specifically, he stresses the variation of inter-point distances within specific distance bands as an indication of positive (clustering) or negative (inhibition) spatial autocorrelation (see also Getis, 1985b). This is applied to population densities in Boston and related to the identification of multiple nuclei within the urban area, suggesting an alternative to the traditional emphasis on distance to CBD.

In Getis (1984), the original L statistic is generalized to include cross-products of ratio scale variables in the numerator, in the form of the so-called $L_1(d)$ and $L_2(d)$ statistics (differing in whether or not self-similarity is included). More importantly, the statistics are framed in a context that focuses on clusters centered at a given location, thus becoming precursors to the local G_i and G_i^* statistics (Sect. 1.2.3).

Another interesting aspect of these papers is their emphasis on the importance of scale, which is further explored in the widely cited article with Janet Franklin on "second-order neighborhood analysis of mapped point patterns" in *Ecology* (Getis and Franklin, 1987, included as Chap. 7). Here, the notion of a *neighborhood analysis* is suggested which uses the L_i statistics (from Getis, 1984) to assess the extent to which clustering around a given location varies with distance. This reveals interesting patterns of heterogeneity in the clusters and further establishes the intellectual basis for the explicit formulation of the local statistics in the 1990s.

Finally, it is worth mentioning that Getis's methodological work in pattern analysis was paralleled by the development of easy to use software. The best known example of this is the so-called PPA point pattern analysis package developed with Dong Mei Chen (Chen and Getis, 1998). This provided the basis for some of the point pattern functionality in ESRI's widely distributed commercial ArcGIS software.

1.2.3 Local Spatial Statistics

Arguably Getis's most important contribution to spatial analysis is his work on local spatial statistics (for an early overview, see Getis and Ord, 1996). While the origins of these ideas can be found in some earlier articles (notably Getis, 1984, 1991), the derivation of the formal properties of the G statistics is reflected in three articles co-authored with Keith Ord.

In "the analysis of spatial association by use of distance statistics" (Getis and Ord, 1992, included as Chap. 10), the basic derivations of the moments of the statistic are provided and inference is based on a normal approximation. This is further refined in "local spatial autocorrelation statistics: distributional issues and an application" (Ord and Getis, 1995) and "testing for local spatial autocorrelation in the presence of global autocorrelation" (Ord and Getis, 2001). In the former, the initial

restriction to binary spatial weights (distance bands) is relaxed to allow the usual row-standardized weights to be used for G statistics. Also, the problem of multiple comparisons is addressed by means of a Bonferroni adjustment. The latter article introduces a new local statistic, the O_i statistic, which accounts for the global structure of autocorrelation. It is essentially a spatially dependent t-test of means within and outside an area around a candidate hot spot location. The moments of this new statistic are derived and its asymptotic normality established.

The impact of this work on the practice of spatial data analysis cannot be overemphasized. Most importantly, there was a significant shift of attention from the global to the local (e.g., Fotheringham, 1997), with a focus on the detection of local clusters and hot spots and a greater sensitivity to spatial heterogeneity. Examples include the development and widespread adoption of a general framework for local indicators of spatial autocorrelation (LISA) (Anselin, 1995) and the collection of techniques for geographically weighted regression (for an overview, see Fotheringham et al., 2002). The G statistics were applied in a wide range of empirical studies, in fields ranging from criminology and epidemiology to ecology. They were also implemented in a number of software packages, including SpaceStat, STARS, the analytical toolbox for ESRI's ArcGIS and the open source *spdep* package for R.

Apart from his own work applying the local statistics in empirical studies, Getis was also interested in two specific methodological refinements in which he saw the G statistics playing a major role: spatial filtering and the construction of spatial weights. We return to spatial filtering in Sect. 1.2.4, but include the spatial weights problem in this section, given its more fundamental methodological importance.

Two articles in *Geographical Analysis* and co-authored with Jared Aldstadt argue for the use of the G_i statistic as a means to determine the values for individual elements in the spatial weights matrix. In "constructing the spatial weights matrix using a local statistic" (Getis and Aldstadt, 2004, included as Chap. 11) and "using AMOEBA to create a spatial weights matrix and identify spatial clusters" (Aldstadt and Getis, 2006) the number of non-zero elements in each row of the weights matrix and their values are obtained as a function of the G_i statistics for increasing distance bands. This is suggested as an alternative to the standard procedure based on contiguity or geostatistical considerations. Further investigations are needed to assess the performance of this new approach in a wide range of empirical contexts.

1.2.4 Empirical Applications

In this fourth category, we illustrate some publications by Getis that are representative of his wide range of empirical interests and collaborative work with other researchers, many of whom are outside geography or regional science. We group these articles into three main categories: urban and economic geography (regional science); spatial filtering; and medical geography/epidemiology. This review is intended to be representative rather than comprehensive. Most notably, it does not include some of his recent work on crime analysis, e.g., in Getis et al. (2000).

As mentioned in the introduction, much of Getis's early empirical work dealt with questions pertaining to urban and regional systems. Even though his main focus later shifted to methodological concerns in spatial statistics, this interest in cities and regions continued throughout his career. Some examples include a study of urban population spacing (Getis, 1985c), a model of economic interdependencies among urban communities (Getis, 1989b), and, more recently, an analysis of regional patterns of affirmative action compliance costs (Griffin et al., 1996).

Getis's idea to use the G_i statistic in the construction of spatial weights matrices also led to a different approach towards modeling spatial dependence. Originally suggested in Getis (1989a), this *filtering* perspective employs the G_i statistic to eliminate spatial correlation from a variable. This then allows the statistical analysis to consider both the spatially filtered form of the variable as well as a new artificial variable that contains the spatial effects. This idea was further developed and illustrated in a number of articles and book chapters, including Getis (1990), and "spatial filtering in a regression framework: examples using data on urban crime, regional inequality, and government expenditures" (Getis, 1995a, included as Chap. 14). The Getis filtering approach contrasts with the idea of employing eigenvectors constructed from the spatial weights as additional artificial variables in the regression equation in order to eliminate spatial autocorrelation, a suggestion advanced by Griffith (e.g., Griffith, 2003). In Getis and Griffith (2002), the two approaches are compared.

The filtering approaches are currently receiving considerable attention in the literature (e.g., Griffith and Peres-Neto, 2006), although further and rigorous evaluation of its statistical (asymptotic) properties and explanatory power relative to the explicit modeling of space remains to be carried out.

A final category of empirical work by Getis deals with applications in medical geography and epidemiology. This includes articles where the main focus is on the methodology, such as the use of local statistics to assess dispersion in AIDS in California (Getis and Ord, 1998), or canine cancers in Michigan (O'Brien et al., 2000). The bulk of his articles, however, are inspired by substantive concerns, primarily related to the spread of Dengue fever. Some illustrative examples are listed in Table 1.1. Of these, Getis et al. (2003) is included as Chap. 15. This work continues to date.

1.3 Quantifying the Impact

According to the Web of Science, there are 65 articles with Getis as an author or co-author which collectively have been cited in the literature 851 times.[1] The high impact of these contributions is reflected in the average Getis paper being cited over 13 times. During the course of a typical year over 23 scientific papers cite one of

[1] The Web of Science data was current as of June 23, 2008.

Table 1.1 Samples of empirical work by Getis in epidemiology

Reference	Title	Journal/book
Getis and Ord (1998)	Spatial modeling of disease dispersion using a local statistic: the case of AIDS	*Jean Paelinck Festschrift*
Morrison et al. (1998)	Exploratory space–time analysis of reported Dengue cases during an outbreak in Florida	*American J. Trop. Med. & Hygiene*
O'Brien et al. (2000)	Temporal distribution of selected canine cancers in Michigan, USA 1964–1994	*Preventive Vet. Med.*
Getis et al. (2003)	Characteristics of the spatial pattern of the Dengue vector, Aedes Aegypti in Iquitos, Peru	*American J. Trop. Med. & Hygiene*
Morrison et al. (2004a)	Evaluation of a sampling methodology for rapid assessment of Aedes Aegypti infestation levels in Iquitos, Peru	*J. Medical Entomology*
Morrison et al. (2004b)	Temporal and geographic patterns of Aedes Aegypti (Diptera: Culicidae) production in Iquitos, Peru	*J. Medical Entomology*
Getis (2004a)	A geographic approach to identifying disease clusters	*Worldminds*

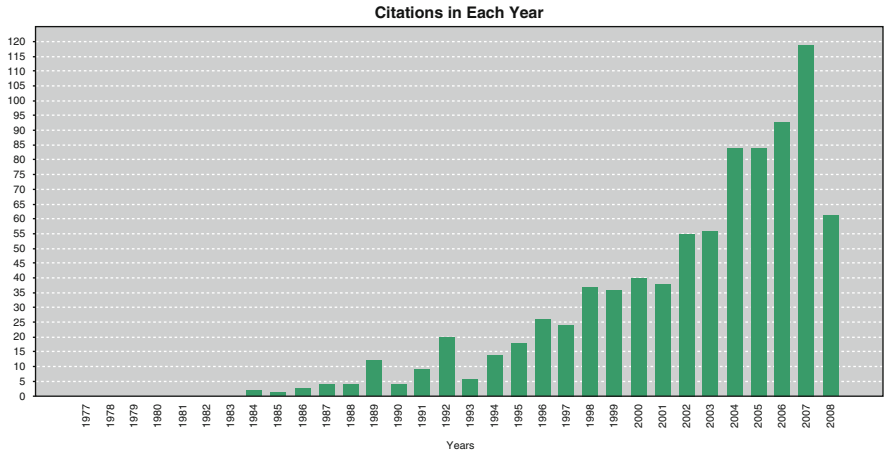


Fig. 1.1 Annual citation patterns for Arthur Getis

Getis' contributions. And, as Fig. 1.1 reveals, the impact of this work has continued to grow, even after his formal retirement in 2004.

Table 1.2 provides a listing of the highest impact papers published by Getis. The major impact of his work with Ord on the local G_i statistics is clear as their two contributions in *Geographical Analysis* account for over 45% of Getis citations over the period examined. At the same time, the mix of topics in the most cited papers includes applications and methods in fields such as ecology and public health, reflecting the broad influence of Getis' scholarship.

Table 1.2 Most cited articles

Authors	Journal	Total citations	Annual citations
Getis and Ord (1992)	Geographical Analysis	236	13.88
Ord and Getis (1995)	Geographical Analysis	149	10.64
Getis and Franklin (1987)	Ecology	114	5.18
Morisson, Getis et al. (1998)	AJTPH	44	4.00
Getis (1984)	Environment and Planning A	35	1.40

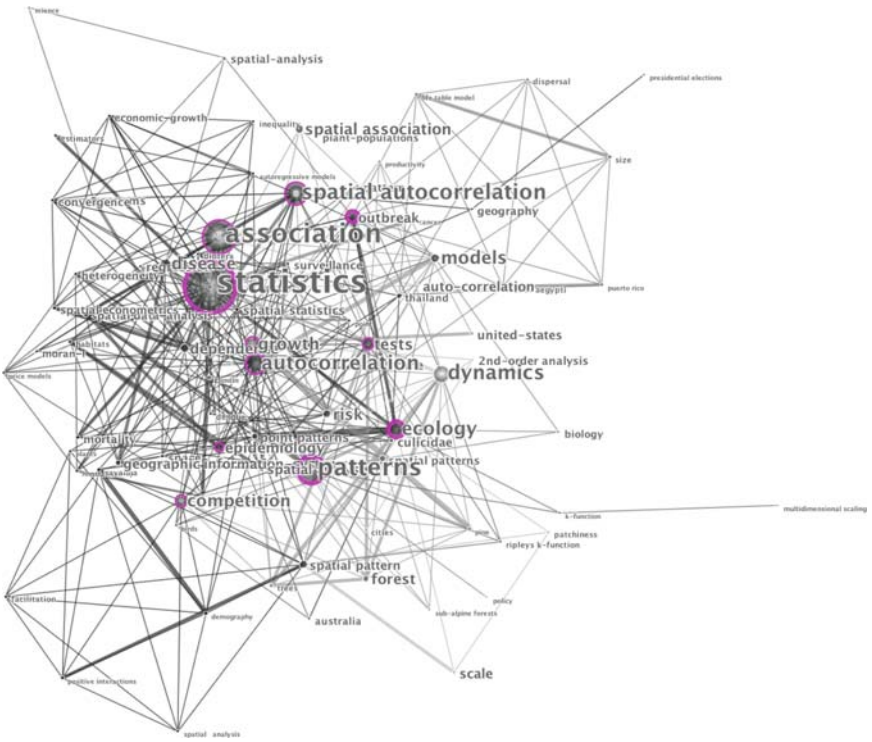


Fig. 1.2 Concept co-citation network

This impact is not only numerically deep but has been extensive in its range of fields and journals. Figure 1.2 portrays the diversity of interrelated topics treated in his work in a co-citation network. Each edge in the network combines two terms or concepts appearing in articles that have been cited by another paper which has cited one of Getis' contributions.[2] The size of each term node is a function of the number of times it is co-cited. This network portrays the footprint of the conceptual space that has been influenced by Getis' work. It is a footprint on a wide array of

[2] The citation networks were developed using the package CiteSpace II (Chen, 2006).

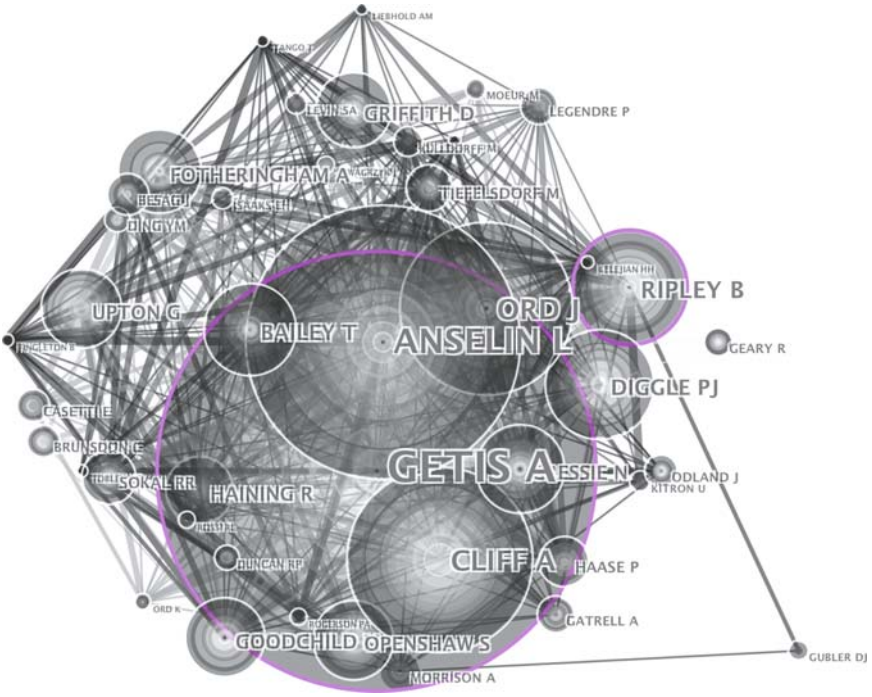


Fig. 1.3 Author co-citation network

disciplines and fundamental concepts, such as scale and patterns, interwoven with an diversity of analytical topics, including spatial autocorrelation and second-order analysis, and dealing with substantive areas ranging from habitats, to disease, to economic convergence. The central nodes of statistics, autocorrelation, patterns, and association anchor this complex and intricate influence.

A second way to explore the impact footprint is to consider the author co-citation patterns for Getis' contributions. Figure 1.3 displays this network where now nodes are individual scholars and an edge connects two authors who have been cited in a paper that originally cites a Getis paper. These edges can represent either a jointly authored contribution, or two separate pieces cited by the same work. An examination of the figure indicates that many of the contributors to this volume are prominent in this citation space, reflecting both types of scientific collaboration with Getis.

We can also examine the key journals that have been impacted by Getis. Figure 1.4 contains the co-citation network for journals (and books) that have published articles that have been co-cited by authors citing work by Getis. The dominance of the journal *Geographic Analysis* is clear, again, reflecting the importance of the two local statistics pieces as was seen in Table 1.2. Two other key journal nodes are *Environment and Planning* and *Ecology*, both of which published contributions by Getis and colleagues which are included in this volume.

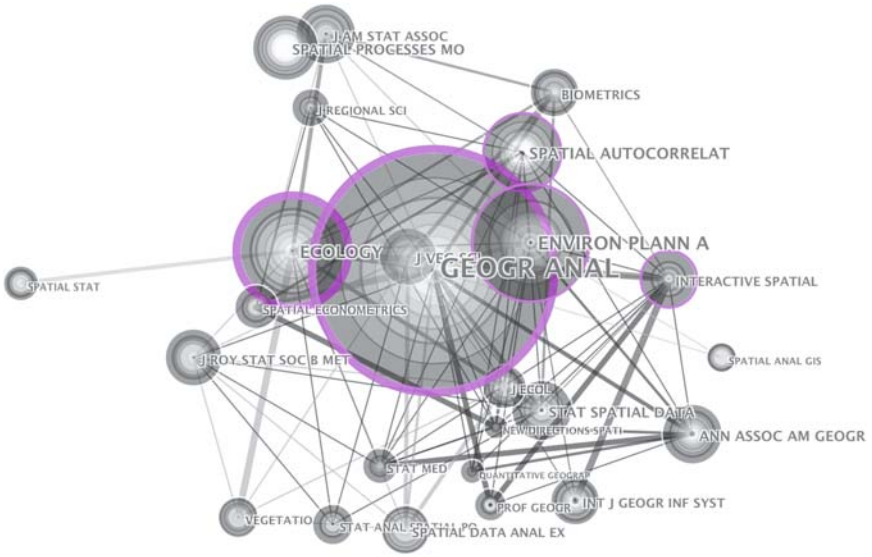


Fig. 1.4 Journal co-citation network

1.4 Outline of the Volume

As mentioned above, we have organized the volume along four broad themes that match the main contributions by Arthur Getis to spatial analysis.

1.4.1 Spatial Analysis

In Part I, we include four chapters dealing with spatial analysis in general. Chapters 2 and 3 are reprints of, respectively, Getis (1991) and Anselin and Getis (1992). The other two chapters, respectively by Goodchild and by Fischer et al. expand on these contributions.

Chapter 2, on "spatial interaction and spatial autocorrelation: a cross product approach," is an article by Getis first published in *Environment and Planning A*. Arguably, this could equally be classified in Part III on local statistics, because to our knowledge it is the first time Getis formally spells out the properties of the $G_i(d)$ local statistic. However, it presents a much broader perspective, outlining a general framework that integrates the perspective taken in the spatial interaction literature (dealing with flows of objects) with the spatial association/spatial correlation view (dealing with similarity of objects).

Getis shows how nearly all spatial autocorrelation statistics can be expressed in some way as a cross product statistic, relating the relative location of observations (e.g., through a spatial weights matrix) to the attribute similarity. This includes join

count statistics, Moran's I, Geary's c, the semivariance, the second order K statistic, and others. In addition, the "Getis model"or $G_i(d)$ statistic is included in this framework. Getis also shows how the mathematical multiplicative operation contained in spatial interaction models is formally similar to that in a cross product statistic, for example, in an origin-specific production constrained interaction model.

Getis introduces the $G_i(d)$ statistic as an extension of the second order $L_i(d)$ statistic presented in his earlier work (Getis, 1984). Formally, the new statistic is a partition of Ripley's $K(d)$ statistic for an individual location. It expresses the proportion of the value of the variable x within a given distance band from the observation. Getis also provides the initial derivation of the basic moments (mean, variance) of this local statistic under the null hypothesis of spatial randomness (using a permutation logic) and suggests some potential extensions. We revisit this in Part III.

In "spatial statistical analysis and geographic information systems," which originally appeared in the *Annals of Regional Science*, Anselin and Getis formulate some general ideas on the role of GIS as an enabling technology to define the research agenda for spatial data analysis. They situate this in the context of a broader discussion of the interface between GIS, spatial statistics and regional science.

The point of departure is a modular description of spatial analytical functionality in GIS, consisting of data selection/spatial sampling, data manipulation, exploration (data-driven analysis) and confirmation (model-driven analysis). Within such a framework, the importance of the special nature of spatial data is stressed, including scale, spatial dependence and spatial heterogeneity. They also outline how the GIS data model relates to spatial sampling, which affects the ensuing analysis in a fundamental way. While many techniques are available to tackle spatial analytical problems, the specification of a collection of generic functions of spatial analysis is identified as an unresolved research question (at the time). Existing analytical functionality can be linked with an operational GIS in a number of ways, referred to as encompassing, modular and loosely coupled.

They discuss the implementation of a framework for spatial analysis within a GIS through a cursory review of the state of the art (at the time), with a special focus on exploratory (ESDA) and confirmatory (CSDA) data analysis methods. They also identify the potential for GIS to complement the existing set of statistical methods with powerful computation intensive approaches and innovative visualization. Anselin and Getis argue for a "new" spatial analysis to emerge from a creative combination of the "old" spatial analysis with the new technologies, stressing that technological development should be led by substantive theory and methodology.

In the provocatively entitled "whose hand on the tiller? revisiting 'spatial statistical analysis and GIS'," Michael Goodchild takes a retrospective look at changes in the relationship between GIS and regional science since the early 1990s. Two themes organize the contribution. The first is the prescience of the original Anselin and Getis article in laying out some of the key methodological challenges that would confront spatial analysis and GIS.

The second theme broadens the examination to consider the relationship between academic spatial analysis and the development of commercial GIS software on the

one hand and the changes in the nature of science on the other. Goodchild agrees with the argument made by Anselin and Getis that the development of GIS software should be driven by the demands of substantive research questions. Yet, the commercial reality in the GIS world since 1992 has seen scientific research and methods representing only a tiny fraction of the overall market. This would seem to suggest that academic researchers would have little influence over trends in GIS. However, Goodchild argues that there are several reasons why this influence has actually been larger than its small market share would suggest. These relate to universities' historical function of training future researchers and scientists, critical examinations of the societal impacts of GIS technology, and academic spatial analysis being the key source of methodological innovations.

In "spatial interaction and spatial autocorrelation," Manfred Fischer, Martin Reismann and Thomas Scherngell examine the issue of spatial autocorrelation in the context of origin-destination interaction models and suggest two novel methodological advances. The first concerns the issue of the proper specification of how to specify spatial autocorrelation in the log-additive spatial interaction model, given that each observation is now associated with a pair of locations. The dyadic nature of spatial interaction data is exploited to develop a new spatial weights matrix. Based on the notion of interaction similarity, a pair of flows are considered "neighbors" if the origin locations in each of the flows are themselves geographic neighbors. Mirroring this origin-based similarity matrix is a destination based similar weight structure where two flows involving destination locations that are contiguous are considered to be neighbors in interaction space.

The second innovation is a generalization of the local G statistic to the case of spatial interaction data. Based on the interaction similarity matrix, this provides a powerful tool to apply in the exploratory analysis of flow data and to detect local interaction clusters. Both innovations are illustrated in an empirical analysis of patent data for 112 European regions.

1.4.2 Spatial Pattern Analysis

Part II contains four chapters on point pattern analysis. Chapter 6 is a reprint of an early discussion by Getis of second order point pattern analysis using Ripley's K function (Getis, 1983). The second reprint, included as Chap. 7, covers an extension of these ideas to local analysis in the much cited Getis and Franklin (1987) article that appeared in *Ecology*. Chapter 8 by Okabe et al. outlines a general class of K functions and Chap. 9 by Franklin reviews the impact of spatial point pattern analysis in plant ecology.

"Second order analysis of point patterns, the case of Chicago as a multi-center urban region" was originally published in *The Professional Geographer*. Getis approaches the study of urban population density from the perspective of point pattern analysis. Using census tract population for the Chicago area and representing each count of 10,000 people by a point, Ripley's K and L functions are applied to

detect patterns of clustering or inhibition. In these analyses, all the interpoint distances are considered, hence the reference to this approach as *second order analysis*. This is in contrast to first order analysis, which concerns itself with the density of the points as such.

A careful analysis of the Chicago example suggests both inhibition at small distances as well as clustering at about 7–9 miles, which corresponds to the mean journey to work distance at the time. Getis makes the link between these distinct patterns and the presence of multiple centers in the metropolitan area, and argues for the use of second order analysis as an alternative to the more commonly applied study of population density. Interestingly, he also points to the danger of a potential misinterpretation of these results due to the effect of scale, which is further elaborated upon in his article with Franklin.

In "second-order neighborhood analysis of mapped point patterns," Getis and Franklin approach second order analysis from a *local* perspective, by considering the variation of inter-point distances around a given point. It is suggested that the relationship between a location and all other locations surrounding it can be captured by four distance parameters: the nearest neighbor distance, the distance where heterogeneity begins, the distance where clustering becomes significant, and the distance where clustering is maximized.

The point of departure is the statistic $L_i(d)$ (Getis, 1984), a measure of the fraction of points within a given distance d of the location of interest. This is compared to the null hypothesis of a random Poisson point process. As the distance d is varied, different patterns are identified, revealing how pattern changes with scale. The analysis can be carried out for a given point, using different distances, or considering all points for a given distance. This yields maps showing the degree of clustering for different distance bands. The article illustrates this methodology with an application to the location of ponderosa pines.

In "a class of local and global K functions and their exact statistical methods" Okabe, Boots and Satoh extend the second order method introduced by Getis and Franklin to develop a broader family of K statistics. This family is composed of three pairs of K-based statistics, with each pair having a local and global form. The first pair are the *global and local cross K statistics*. These consider two sets of point patterns in which one set (P) is considered to have random locations, such as crime spots, while the second set are points whose locations are fixed as in the case of railway stations. For each fixed (or base) location i the local cross statistic $K_i(t)$ is the number of points (from the random set P) that are contained in a disc $D_i(t)$ centered on i with radius t. The global cross K function is then taken as the sum of the local functions $K(t) = \sum_{i=1}^{m} K_i(t)$.

Under the null hypothesis that the points P have a uniform distribution, they show that the local version of the statistic has a binomial distribution, yet, even though the global statistic is a simple sum of the local statistics, the derivation of the properties of the global statistic are not straightforward due to complications that arise when the discs centered around the focal base points overlap and the lens of this overlap contain points in P. This results in a double counting of the points in the calculation of the global K statistic. They show that with overlapping, non-empty

discs, the global statistic follows the univariate multinomial distribution, while in the case where there are no overlapping discs, the global statistic has a binomial distribution.

The solution to the overlapping problem is through the second pair of statistics in the family, the *local and global Voronoi cross K* statistics. These rely on a tessellation of the base points such that any overlapping lens is split by the edge of the Voronoi polygon resulting in each point in P being assigned to only one base point and disc.

The final pair of statistics in the family are the *local and global auto K functions*. In contrast to the cross statistics, the auto functions only consider one set of points and the local form of the statistic is in fact the original local K introduced by Getis and Franklin. While recognizing that the original local K function was derived by "localizing" the global K function, they demonstrate that it is entirely possible, and perhaps more straightforward, to proceed in the opposite direction and create global statistics as sums of the local statistics.

In "the spatial point pattern analysis of plants," Janet Franklin revisits the impact of the Getis and Franklin paper on the practice of spatial point pattern analysis in plant ecology, and specifically aims to determine if local statistics are being used and how. Broadly speaking, methods of point pattern analysis have been used by ecologists to relate spatial patterns to underlying processes of predation and competition, as well as to explore spatial heterogeneity and to identify clusters of individuals sharing similar characteristics. Although endogenous biological processes are expected to generate detectable spatial patterns, the application of point pattern methods to relate patterns back to processes can encounter the *equifinality* problem when multiple processes are capable of generating the same pattern.[3]

With regard to the question pertaining to the use of the local statistic in the subsequent ecological literature, a detailed examination of citation patterns reveals that although the paper was written to introduce local spatial statistics to ecologists, it has most often been cited with reference to global point patterns statistics. Franklin posits that this is because the original paper provides a clear summary of classic work in global spatial pattern analysis. The pedagogical strength of the Getis and Franklin paper in its publication history, as it originally submitted as a technical note but the editors requested a full-length paper presaging the wide utility the methods come would have in the field of ecology.

The second question taken up by Franklin is how local statistics have been used, and three general areas are identified. The first is work on methods for edge corrections that build on the adjustments presented in the original paper. Second is the use of local statistics to detect spatial segregation of individuals and species and to delineate homogeneous subareas. Third is the development of spatially explicit indices of clumping using local statistics that are in turn used in logistic regression models to test processes of survival and mortality. While these three areas represent

[3] The equifinality problem is encountered under the guise of the identification problem in econometrics, although to our knowledge the relations between the two have not been treated in the literature.

areas of methodological advancement, Franklin notes that a gap still exists between applied work and theory in ecology since the original paper has typically been cited for reasons other than the local spatial statistic it introduced.

1.4.3 Local Spatial Statistics

Part III contains four chapters dealing with a local perspective on spatial data analysis. Two are selected reprints from the series of papers by Getis and co-authors on the properties of the $G_i(d)$ and related statistics and their application in the construction of spatial weights matrices. Specifically, the classic Getis–Ord paper from *Geographical Analysis* (Getis and Ord, 1992) is included as Chap. 10, and a more recent article on spatial weights, co-authored with Aldstadt (Getis and Aldstadt, 2004, also from *Geographical Analysis*) appears as Chap. 11. Chapter 12 is a reflective essay on spatial autocorrelation analysis by Keith Ord, and Chap. 13, by Rogerson, illustrates a local approach to surveillance in spatial epidemiology.

In "the analysis of spatial association by use of distance statistics," Getis and Ord present both local and global forms of a family of the so-called G statistics and outline their formal properties. The local form dates back to suggestions in Getis (1984), whereas the global statistic has roots in the discussion of spatial autocorrelation and spatial interaction in Chap. 2.

The local G statistics, $G_i(d)$ and $G_i^*(d)$ are related to Ripley's K and show the fraction of values of a variable within a given radius d around a specific observation. They differ in whether or not the location itself is included in the calculation. The global G statistic is a generalization of this concept that includes cross products of the variates within a given radius. This measure is related to Moran's I, but differs in an important respect. Positive and significant values of the G statistics suggest a cluster of high values, whereas negative and significant values suggest a cluster of low values. In contrast, a positive Moran's I suggests similarity (either high or low) and a negative value indicates dissimilarity.

The moments of these statistics are derived under the null hypothesis of spatial randomness, implemented using a randomization approach. With the mean and variance in hand, a standardized Z value can be constructed. Its distribution can be approximated by a Gaussian distribution, which allows for statistical inference.

The new statistics and Moran's I are compared for a number of artificial spatial layouts. They are also applied in two empirical examples. One is the familiar SIDS data set for North Carolina counties popularized in the work of Cressie (1993), the other a sample of dwelling unit prices for zip code areas in San Diego county. The SIDS case is an example of a situation where global measures of spatial autocorrelation fail to be significant, whereas the local statistics suggest the existence of significant clusters. The San Diego case shows significance for both global and local spatial autocorrelation statistics.

In "constructing the spatial weights matrix using a local statistic," Getis and Aldstadt suggest a new method to compute the values for the elements of a spatial

weights matrix. This is based on the local G_i^* statistic. For each location, it is computed for increasing distance bands to determine the range beyond which no spatial autocorrelation is assumed to exist. The specific criterion used to select this *critical distance* is a decrease in the absolute value of the $G_i^*(d)$ statistic. The actual weights w_{ij} are then computed as a function of the G_i^* value for the actual distance and that for the critical distance. Weights for locations that are more than the critical distance apart are set to zero.

Since the calculation is carried out for each location in turn, this procedure allows for heterogeneity in the range and values of weights. Getis and Aldstadt also present a *local statistics model* (LSM), in which spatial structure is seen as consisting two components, one which depends on a distance effect, and one that does not. This is expressed as a regression with both a spatial weights matrix and a dummy variables on the right hand side. For those locations that do not show a distance effect, the row and column of the weights matrix is zero and the dummy variable is set to one. When non-zero elements are present in the weights row, then the dummy variable is set to zero.

This model is assessed in a small number of simulations for artificial layouts that mimic spatial randomness and two types of clustering. The new weights perform well in these examples compared to traditional contiguity based weights as well as semi-variance based weights.

In "spatial autocorrelation: a statistician's reflections" Keith Ord reminisces about his original meeting with Getis while a new member of the economics faculty at the University of Bristol in the late 1960s. It was during this time that Ord was collaborating with Andrew Cliff on statistical measures of spatial dependence in possibly irregular spatial configurations. That work was presented at a Regional Science Association meeting in London in a session shared with (now) Nobel Laureate Clive Granger. Ord reminds us that many of the questions proposed by Granger in that session some 40 years ago remain largely unanswered today and continue to present challenges to spatial modelers. These include questions about whether the spatial process is isotropic, spatially stationary and the relation of samples to populations in spatial data. Cognizant of the challenges facing spatial model development, Ord proposes the "second law of geography":

> All maps are wrong but some are useful

which can be viewed as an addition to Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things," and an adaptation of George Box's first law of statistics: all models are wrong but some are useful.

Ord's takes up spatial processes for lattice data, that is for data recorded as areas within a region. He suggests an *asymptotic relative efficiency* (ARE) measure as a guide to specification of the weight matrix. Ord uses the ARE to consider the impacts of weights misspecification under two scenarios, one in which the true matrix is the familiar rook definition of contiguity and the second the queen definition. For each case the ARE are calculated for misspecified weights matrices, using the two aforementioned definitions and a third that is a hybrid based on the notion

of isotropic dependence reflecting a directional invariance but distance-dependence structure to the spatial autocorrelation. The hybrid case is found to dominate the other two alternatives with regard to ARE.

Ord's contribution also considers questions related to local spatial statistics, particularly concerns regarding significance levels when carrying out a large number of tests. He also illuminates the differences between the G_k statistic and the local Moran I_k statistic, noting that these differences make the statistics complements rather than competitors since they each can detect different patterns in the data. Ord also examines *local estimation* in two contexts, one in which the weights are prespecified and the second in which the weights are also incorporated in the estimation. The final topic introduced is that of an anisotropic spatial lag which allows for directional dependence or asymmetry in the strength of spillovers as, for example, the case of the negative effect of a high crime rate area on housing prices in a neighboring low crime area being stronger than the positive effect on prices in the other direction.

In "health surveillance around prespecified locations using case-control data" Peter Rogerson considers the problem of monitoring data around a putative point source in order to detect as quickly as possible any change in risk that may be occurring. Known as *prospective detection*, Rogerson introduces a new the method that relies on the availability of case-control data characterized by both location and a time of diagnosis. A log-likelihood function for the cases and controls is specified which is driven by two central parameters: one (θ_1) which reflects the excess risk at the location of the putative source, and a second (θ_2) which captures the decline in risk as distance from the putative source increases.

A likelihood ratio test is then derived for conducting a single test to determine if there is both significant excess risk at the source and whether that risk declines with distance from the source. This test is then extended to a temporal context through the use of *cumulative sum methods* to derive a diagnostic for detecting *change in risk over time*. A cumulative sum of score statistics, each of the latter formed as the ratio of the log-likelihoods from before and after a change, is compared against a predefined threshold parameter. The properties of the test are then examined in a carefully developed set of simulations. The results indicate that, as expected, the detection occurs more quickly as the underlying risk increases, while they also highlight some of the challenges related to specifying parameter changes when implementing the test in practice.

1.4.4 Empirical Applications

In this final part, we include four chapters illustrating empirical applications. The two Getis reprints deal, respectively, with spatial filtering (Getis, 1995a, included as Chap. 14) and with the use of global and local spatial statistics to characterize the spatial distribution of the Dengue vector (Getis et al., 2003, included as Chap. 15). Chapter 16, by Daniel Griffith, considers a comparison of the eigenvalue

and local statistics spatial filtering methods. Chapter 17, by John Weeks, is an in-depth analysis of the spatial distribution of fertility in rural Egypt.

In "spatial filtering in a regression framework: examples using data on urban crime, regional inequality, and government expenditures," Getis refines his earlier suggestion on how to construct spatially filtered variables (outlined in Getis, 1990) and provides three detailed empirical illustrations. The rationale behind spatial filtering is to remove the inherent spatial autocorrelation from all variables in a regression specification (both dependent and explanatory variables). The new filtered variables then allow for the regression to be estimated by means of classical ordinary least squares. In addition, artificial variables containing the spatial effects can be included in the specification as well.

The main contribution of this chapter is the use of the $G_i(d)$ statistic to construct the spatially filtered variable. An important aspect of this is the choice of the optimal distance d, such that there is no remaining spatial autocorrelation beyond this distance. The original variable is decomposed into a filtered part (which contains no spatial autocorrelation) and the remainder, which represents the spatial effects due to the spatial configuration of the data.

Getis suggests four tests to assess whether the procedure works: (1) no spatial autocorrelation should remain in the filtered variable; (2) the difference between the original and filtered variable should show significant spatial autocorrelation; (3) there should be no remaining residual spatial autocorrelation in the regression; and (4) the filtered variables should be significant in the regression. This is illustrated with three empirical examples: the classic Columbus neighborhood crime data set; a study of regional per capital income in regional divisions for Turkey; and government expenditures for US states. In all instances, the filtering procedure performs satisfactorily.

Getis and co-authors illustrate a careful application of global and local spatial autocorrelation analysis in "characteristics of the spatial pattern of the Dengue vector, Aedes Aegypti, in Iquitos, Peru," which originally appeared in the *American Journal of Tropical Medicine and Hygiene*. The study is carried out using detailed household data for two distinct neighborhoods in the Amazonian city of Iquitos, Peru. These neighborhoods differ in the way water is managed, which is highly relevant to the quantification of the vector population, largely driven by the presence of water in containers.

Four variables were considered: the number of adult pupae; water holding containers; water containers positive for the presence of larvae; and water containers positive for the presence of pupae. For each of these variables, a global K function was utilized to assess clustering, and the local $G_i^*(d)$ statistic was employed to identify the locations of clusters. This was implemented by means of the PPA software package, co-developed by Getis (Chen and Getis, 1998).

A careful analysis of clusters by neighborhood and over time leads to specific recommendations for Dengue control and surveillance strategies. The results also point to the importance of spatial scale in the study of the dynamics of Dengue transmission. In this particular application, the proper scale turns out to be the household, but with a need to carry out measurement at frequent time intervals.

In "spatial filtering and missing georeferenced data imputation: a comparison of the Getis and Griffith methods," Daniel Griffith explores the use of recently developed spatial filtering methods to the case of small area estimation. Previous research had suggested that one class of filtering methods, those based on the local G_i statistic appear to lend themselves to the problem of imputation, while the eigenvector based filtering method did not. Griffith takes up this conjecture by a detailed analysis of the properties of the two different filtering methods.

By viewing these filtering approaches as special cases of the more general *Expectation Maximization* problem, Griffith derives missing data prediction equations for each of these two original formulations, and then compares these approaches using several popular datasets. The results provide two key corrections to the earlier conjecture about filtering methods in an imputation context. The first is that the eigenvector based approach can indeed be applied to impute missing georeferenced data. The second corrective is that while the G_i approach also is applicable to small area estimation problems, those applications are complicated by the constraint that the variable in question has to be nonnegative and the potential requirement for the additional estimation of threshold distances in the imputation.

In "spatial patterns of fertility in rural Egypt" John R. Weeks applies a suite of geospatial tools to examine the spatial patterns of human reproduction in a rural governorate in Egypt. Noting that existing demographic thought rarely has gone beyond the question of rural vs. urban differences in fertility behavior, Week's demonstrates that even these regional differences can mask underlying spatial heterogeneity at the finer spatial scale of the village. These patterns are uncovered through the novel combination of remote sensing imagery together with dasymetric mapping and census data to develop a spatially rich geodemographic dataset on fertility dynamics over a 20-year period.

Clusters of high-fertility areas, or "hot spots" are identified through the use of the local G_i statistics. These, in turn, are used in a spatially filtered regression model to examine the determinants of fertility change. This specification includes traditional covariates such as illiteracy, marital status, and age composition, but does so by decomposing each into a spatial component (based on the G_i statistic) and a filtered component in which the spatial autocorrelation in that covariate is removed. The results indicate, that the traditional covariates do have their expected signs, however, the spatial components have became more important over time as a predictor of fertility levels. This nicely demonstrates a basic tenet of geodemographics that behavior is a joint function of who people are and their spatial context.

Part I
Spatial Analysis

Chapter 2
Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach

Arthur Getis

This Chapter was originally published in:
Getis, A. (1991) Spatial Interaction and Spatial Autocorrelation: Across-Product Approach. Environment and Planning A 23:1269-1277. Reprinted with permission of PION Limited, London

Abstract A cross-product statistic is used to demonstrate that spatial interaction models are a special case of a general model of spatial autocorrelation. A series of traditional measures of spatial autocorrelation is shown to have a cross-product form. Several interaction models are shown to have a similar form. A general spatial statistic is developed which indicates that the relationship between the two types of models is particularly strong when the focus is on measurements from a single point.

2.1 Introduction

In casual conversation one rarely makes a distinction between those elements of our environment that are associated and those that interact. It is commonly believed that if tangible or intangible variables interact they are therefore in association with one another. Spatial scientists, however, have made in the technical literature a distinction between spatial association, which implies correlation, and spatial interaction. There is among them a deep-seated view that spatial interaction implies movement of tangible entities, and that this has little to do with spatial correlation. A literature on spatial autocorrelation has arisen that is nearly devoid of references to the literature on gravity and interaction models. Only on rare occasions will a spatial scientist use the words "spatial interaction" to refer to the ideas of the spatial associationists (Haining, 1978; Ord, 1975).

In this paper, I suggest that the family of spatial interaction models is a special case of a general model of spatial autocorrelation. The goal is to bring the two modeling "camps" together into a single group whose purpose is to develop further

A. Getis
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

spatial models in a general way. In recent reviews of the interaction model and spatial autocorrelation literature, such as in Haynes and Fotheringham (1984) and in Anselin (1988), respectively, there is little recognition of the contributions of the other group. There has not been a discussion that shows that the two types of models can be described in a general way by the same spatial model. In order to solidify the relationship I will present a statistic that I have developed with the assistance of Ord that can be interpreted as either an indicator of spatial autocorrelation or a measure of spatial interaction.

There have been a number of generalizations of gravity and spatial interaction models (Tobler, 1983; Wilson, 1970). The most recent contribution is by Haynes and Fotheringham (1984), who write the general model as

$$T_{ij} = f(V_i, U_j, S_{ij}),$$

where T_{ij} is the interaction (tangible or intangible) between i and j, V_i and U_j represent vectors of origin and destination attributes, respectively, and S_{ij} represents a vector of separation attributes. By introducing constraints and specifying the form of the attributes, one can produce a model for validation. The relationship between the dependent and independent variables is often constrained. In some instances emphasis is on V_i (origin-specific, production-constrained gravity models), whereas in others the U_j are most important (destination-specific, attraction-constrained), and in some models there is a balance between the two (doubly constrained models). Fotheringham (1983) adds a further general term to the system, C_j, which represents a vector of competition variables. As he implies, however, C_j is a refinement of and a more detailed specification for U_j.

The historic background that has led to the current understanding of spatial autocorrelation models is very much different from that of interaction models. Spatial autocorrelation modeling has had a shorter history. Interaction modeling has been active for over 100 years although it was in the late 1950s when there was a resurgence of interest that has lasted to the current time. The field had already been reinfused with the theoretical energy of Wilson in the early 1970s when Cliff and Ord (1973) presented their ground-breaking explication of the spatial autocorrelation problem based on the work of Moran (1948), Geary (1954), and Whittle (1954).

Since 1973 the development of spatial autocorrelation models has been slow and tedious. The literature gives no evidence that Moran's I model, the join-count model, and the Geary model have been replaced or modified. Considerable progress is clear, however, in the development of regression models that include one or more spatial autocorrelation coefficients. In related developments, spectral models and especially variograms (Kriging) are being used to estimate the nature of autocorrelation in spatial data.

The common elements of the various spatial autocorrelation models are (1) a matrix of values representing the association between locations and (2) values representing a vector of the attributes of the various locations. To my knowledge, only Hubert and his associates Golledge, Costanzo, and Gale (Hubert and Golledge, 1982; Hubert et al., 1981, 1985) have developed a general form for the association

of these elements. Their cross-product statistic, Γ, is written

$$\Gamma = \sum_{i,j} W_{ij} Y_{ij}, \tag{2.1}$$

where W_{ij} are elements of a matrix of measurements of spatial proximity of places i to places j, and Y_{ij} is a measure of the association of i and j on some other dimension. A slightly different form is

$$\Gamma = \sum_{i,j} W_{ij} Y_j \tag{2.2}$$

in which the relationship between the Y_i and Y_j is implicit rather than explicit as in (2.1). In this and in all subsequent formulations where we use summation signs, i does not equal j (that is, there is no self-association or self-interaction), unless otherwise indicated. In addition, in all subsequent formulations stationarity and isotropy are assumed where required. A common choice for Y_{ij} is

$$Y_{ij} = (x_i - x_j)^2, \tag{2.3}$$

where the x are the values observed for variate X_i. Clearly Y_{ij} could be some other measure of the association between i and j. For example, Hubert et al. (1981) propose $\cos(d_i - d_j)$ where d_i and d_j are angular directions at i and j. In the following paragraphs I shall identify briefly the differentiating elements of the various spatial autocorrelation models.

2.2 Cross-Product Spatial Autocorrelation Models

In this section I give a survey of the models of spatial autocorrelation. In each case attention is on the form of the model. The purpose is to show that nearly all of the models are simply just another specification of a cross-product statistic.

2.2.1 The Join-Count Models

These models require a 0,1 attribute scale. That is, some places display the attributes (1) whereas others do not (0). The Y_{ij} of the cross-product statistic differs according to the particular model of which there are three: (1) association of places with the attribute $Y_{ij} = x_i x_j$; (2) association of places with and without the attribute $Y_{ij} = (x_i - x_j)^2$; and (3) association of places without the attribute $Y_{ij} = (1 - x_i)(1 - x_j)$. The first and the third model exhibit a multiplicative form. Each of the models is constrained by allowing only a value of one for a success and zero for a failure. The model is evaluated against the expectation of the moments of X_i (see Cliff and Ord, 1973). There are no constraints on the weight matrix although in practice researchers usually choose a one-or-zero scheme to identify spatial proximity or no

Table 2.1 A comparison of various spatial models and the cross-product statistic

Model	W_{ij}	Y_{ij}	Restrictions		Scale
			W_{ij}	Y_{ij}	
Cross-product statistics					
$\Gamma = \sum \sum W_{ij} Y_{ij}$	W_{ij}	Y_{ij}	None	None	None
$\Gamma = \sum \sum W_{ij} Y_j$	W_{ij}	Y_j	None	None	None
Spatial autocorrelation models					
Joint count					
$BB = \frac{1}{2} \sum \sum W_{ij} x_i x_j$	W_{ij}	$x_i x_j$	0/1	0/1	$\frac{1}{2}$
$BB = \frac{1}{2} \sum \sum W_{ij} (x_i - x_j)^2$	W_{ij}	$(x_i - x_j)^2$	0/1	0/1	$\frac{1}{2}$
$BB = \frac{1}{2} \sum \sum W_{ij} (1 - x_i)(1 - x_j)$	W_{ij}	$(1 - x_i)(1 - x_j)$	0/1	0/1	$\frac{1}{2}$
Moran's					
$I = \frac{n \sum \sum W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{W \sum (x_i - \bar{x})^2}$	W_{ij}	$(x_i - \bar{x})(x_j - \bar{x})$	None	None	$\frac{n}{W \sum (x_i - \bar{x})^2}$
Geary's					
$c = \frac{(n-1) \sum \sum W_{ij} (x_i - x_j)^2}{2W \sum (x_i \bar{x})^2}$	W_{ij}	$(x_i - x_j)^2$	None	None	$\frac{n-1}{2W \sum (x_i - \bar{x})^2}$
Semi-variance					
$\gamma = \frac{1}{2} \sum_{i=1}^{n-h} \sum_{j=i+h}^{n} W_{ij} (x_i - x_j)^2$	W_{ij}	$(x_i - x_j)^2$	1	None	$\frac{1}{2}$
Second-order					
$K(d) = \frac{\sum \sum W_{i,j}(d) x_i x_j}{(\sum x_i)^2 - \sum x_i^2}$	$W_{ij}(d)$	$x_i x_j$	0/1	Positive	$[(\sum x_i)^2 - \sum x_i^2]^{-1}$
Getis model					
$G_i(d) = [\sum_j W_{ij}(d) x_i x_j](\sum_j x_i x_j)^{-1}$	$W_{ij}(d)$	$x_i x_j$	0/1	Positive	$(\sum_j x_i x_j)^{-1}$
Spatial interaction models					
General gravity					
$T_{ij} = k x_i^{\alpha} x_j^{\tau} W_{ij}^{-\beta}$	$W_{ij}^{-\beta}$	$x_i^{\alpha} x_j^{\tau}$	None	Positive	k
Origin-specific, production-constrained					
$T_{ij} = (x_i x_j^{\alpha} W_{ij}^{-\beta})(\sum_j x_j W_{ij}^{-\beta})^{-1}$	$W_{ij}^{-\beta}$	$x_i x_j^{\alpha}$	None	Positive	$(\sum_j x_j W_{ij}^{-\beta})^{-1}$
General spatial models					
i-to-all-j model					
$G_i = (\sum_j x_i x_j W_{ij}^{-\beta})(\sum_j x_i x_j)^{-1}$	$W_{ij}^{-\beta}$	$x_i x_j$	None	Positive	$(\sum_j x_i x_j)^{-1}$
i-to-j model					
$G_{ij} = (x_i x_j W_{ij}^{-\beta})(x_i x_j)^{-1}$	$W_{ij}^{-\beta}$	$x_i x_j$	None	Positive	$(x_i x_j)^{-1}$

Note: BB black–black joins, *BW* black–white joins, *WW* white–white joins

spatial proximity. In Table 2.1 the cross-product characteristics of the models are identified.

2.2.2 Moran's I Models

The theoretical base for these models is interval-scale observations. There are two models here, differentiated only by the procedures for the evaluation of results. Unlike the join-count models, these are essentially a Pearson product-moment correlation coefficient model altered to take into consideration the effect of a spatial weight matrix. The cross-product, Y_{ij}, is the covariance, $(x_i - \bar{x})(x_j - \bar{x})$. The weight matrix has no restrictions. As in the Pearson statistic, Moran's measurement includes a scaling factor. No doubt the popularity of the Moran statistic is because of the asymptotic normal distribution of the model as n increases (Cliff and Ord,

1973). A roughly equivalent model based on a likelihood ratio statistic is by Haining (1977).

2.2.3 Geary's c Models

The two models here are similar to Moran's models except for the way in which the cross-product attributes are written. In this case the Y_{ij} is $(x_i - x_j)^2$. This is the same as the second join-count model. The variance is a scalar, and the weight matrix is as in the Moran models. The value 1 for c implies that there is no spatial autocorrelation.

2.2.4 The Semivariance Model

The semivariance is a geostatistical measure of autocorrelation based on a lattice of evenly spaced data points. Estimation of the semivariance, $\gamma(h)$, results from the sum of multiples of the values of pairs of points that are separated by a constant spatial lag h units of distance from one another in a single direction. Because of the supposed dependence between nearby data points, as h increases one would presume that the degree of autocorrelation would decline and the variance would increase to the level of the population at large. The model gets its name from the fact that the quantity is half the expected squared difference between two values. As h increases the trend of the $\gamma(h)$ values is called a variogram, not unlike the correlogram often found in studies that use Moran's I. For Hubert's statistic the value h is the equivalent of a one-or-zero weight matrix for a specified set of pairs of points that are h distance units apart in one direction (say east to west) and the values of Y_{ij} are of the form $(x_i - x_j)^2$. The variogram can be written in cross-product form as

$$\gamma(h) = \frac{1}{2} \sum_{i=1}^{n-h} \sum_{j=i+h}^{n} W_{ij}(x_i - x_j)^2. \tag{2.4}$$

2.2.5 Second-Order Spatial Autocorrelation

In a measure of spatial autocorrelation I developed earlier (Getis, 1984) the distance between x_i and x_j, is d. The d value generates a weight matrix of ones for all pairs of points found within d of one another and gives zeroes for all other pairs of points. The result is a cumulative measure of spatial autocorrelation for each distance. The measure taken over many distances creates a cumulative correlogram. The main difference between the second-order approach and the variogram is its cumulative nature, the second-order model does not depend on a lattice of points. The model for an area of size A is given by the expression

$$K(d) = \left(A \sum_{i,j} W_{i,j} x_i x_j \right) \left[\left(\sum X_i \right)^2 - \sum X_i^2 \right]^{-1}, \tag{2.5}$$

where the elements $W_{i,j}$ of the matrix are one or zero, with a one attributed to those j within d of i, and the $Y_{i,j}$ matrix contains $x_i x_j$ pairs. The X variable has a natural origin and $x_i \geq 0$. Clearly, the cross-product statistic describes the numerator and the denominator is a scalar that describes the sum of all $x_i x_j$ pairs, revealing that the measure $K(d)$ is a proportion.

2.2.6 Spectral Analysis

Although I suspect that it is possible to squeeze a spectral view of spatial autocorrelation into a cross-product form, spectral analysis is fundamentally different from the analytical models presented above. In spectral techniques it is assumed that there is a series of frequencies making up distinct periodicities in spatial data. The mathematics for identifying the harmonics are more complicated than those embodied in cross-product analysis. Spectra result from the addition of successive harmonics of a cosine wave. Spectral analysis is an effective analytical device if one is willing to assume that spatial autocorrelation is a consequence of some sort of vibratory motion or accumulation of wave-like forces.

2.2.7 The Spatial Autoregressive Model

A first-order autoregressive model is given by

$$Y_i = \alpha + \rho \sum W_{i,j} Y_j + \epsilon_i. \tag{2.6}$$

For a spatial autoregressive interpretation ρ is the spatial autocorrelation coefficient, W_{ij} is an element of the spatial weight matrix, and ϵ is the uncorrected, normally distributed, nonspatially autocorrelated, homoscedastic error term. The $W_{i,j}Y_j$ is a spatial variable which we construct from the dependent variable itself, and the system is stationary. Thus, the model represents the spatial dependence structure of Y. This is not a model of spatial autocorrelation per se but a model of the effect of spatial autocorrelation on a dependent variable. The main difference with the models described above is that the coefficient ρ is a parameter that relates the spatial dependence form of Y with itself, whereas Moran's I, for example, is strictly a value representing the spatial autocorrelation characteristic of variable Y. In fact, the numerators of both I and ρ are the covariance.

2.2.8 A Cross-Product View of Spatial Autocorrelation

The point of the above exercise is that the numerators of the autocorrelation models are essentially cross-product statistics (see Table 2.1). The W_{ij} matrix is not constrained or, if it is, the constraint is usually because of some maximum-distance rule,

contiguity, or another condition that focuses attention on a specified set of interacting locations. In Table 2.1, the values of the Y_{ij} are entered into the equation in a multiplicative way, as a squared difference, or as a covariation. All other parts of the equations define the base or scalar for the calibration of the various statistics.

Hubert et al. (1981) imply that for testing purposes scales in the formulations are unnecessary. Scales are generally included in the various measures of spatial autocorrelation in order to satisfy assumptions that allow for statistical tests on well-known probability distributions. Hubert (1977) has developed a randomization technique of matrix manipulation that allows one to make statements of statistical significance without making distributional assumptions. Thus not only have we defined a family of cross-product statistics, but if we were to follow Hubert's advice we would use the same type of evaluation procedure for every formulation of Y_{ij}.

2.3 Interaction Models

I shall write the formulas for two common gravity and interaction models:

$$T_{ij} = kP_i^y P_j^\alpha d_{ij}^{-\beta} \tag{2.7}$$

and

$$T_{ij} = A_i O_i W_j^{\alpha_i} d_{ij}^{-\beta_i}. \tag{2.8}$$

The first is the general unconstrained gravity model where the P_i and P_j represent the magnitude of the variable under study at i and j, d_{ij} is the distance separating i and j, the exponents on the P variable are sometimes used to differentiate the effect of the origin from that of the destination. The exponent on the distance value represents the friction of distance. The k is a scalar or constant of proportion.

The characteristics of interaction measures that help differentiate them from autocorrelation measures are (1) a focus on a single ij relationship; (2) the use of exponents to adjust variables; (3) constraints to draw attention to one or more of the variables. In terms of the cross-product statistic there are significant similarities between them. In Table 2.1, (2.7) is rewritten to conform to the nomenclature of the cross-product statistic. Note that no summation sign is used in (2.7) or in Table 2.1. The focus in interaction modeling is on a single association, although the derivation of the parameters usually depends on the empirical data of all associations. The point, however, is that the form of the measure is similar to measures of spatial autocorrelation. The T_{ij} is simply one value that could be used in the development of a spatial autocorrelation statistic. The elements of a W_{ij} matrix contain the values of $d_{ij}^{-\beta}$. The Y_{ij} are simply the association values between the places i and j. As in the spatial autocorrelation statistics, the Y_{ij} are defined in any of a number of ways. The various constraints placed on the values at the i places can easily be accommodated in a cross-product statistic. Thus, the exponents that are used

in interaction models represent more advanced development than in autocorrelation models, but there is nothing standing in the way of the use of exponents to enhance spatial autocorrelation measures (Cliff and Ord, 1969).

Equation (2.8), the origin-specific production-constrained interaction model, has been rewritten in Table 2.1 to conform to the cross-product model. It is clear that even with the complexity characteristic of many interaction models, the general form remains that of a cross product.

2.4 A General Spatial Statistic[1]

The statistic developed below contains the elements of the cross-product statistic but instead of it being a summary measure over an entire set of data it focuses on a single point as in spatial interaction measures. As it is developed here, the translation from spatial autocorrelation to interaction is not without problems.

The statistic is given by the equation

$$G_i(d) = \Big[\sum_j W_{ij}(d)x_j\Big]\Big(\sum_j x_j\Big)^{-1}, \tag{2.9}$$

where W_{ij} is a one-or-zero spatial weight matrix with ones for all links defined as being within distance d of a given place i and all other links are zero. The variable X has a natural origin and is positive. The numerator is a cross product and the denominator is the sum of all the x other than x_i. If S is equal to $x_1 + \cdots + x_n$, it follows directly that

$$K(d) = \left[\sum_j x_i(S - x_i)G_i(d)\right]\left(S^2 - \sum_i x_i^2\right)^{-1} \tag{2.10}$$

so that $G_i(d)$ represents a partition of $K(d)$ to provide an index for the ith location.

Making use of a permutations argument and recognizing that the denominator is invariant under permutations, we can consider the statistic as

$$G_i = \Big[\sum_j Q_j x_j\Big]\Big(\sum_j x_j\Big)^{-1},$$

where $Q_j = 1$ if $W_{ij} = 1$, otherwise $Q_j = 0$. This means that $P(Q_j = 1)$ is equal to $W(n-1)^{-1}$ where $W = \sum_j W_{ij}(d)$. Then

[1] This section was developed with J K Ord.

$$E(G_i) = \left[\sum_j E(Q_j)x_j\right]\left(\sum_j x_j\right)^{-1} = W(n-1)^{-1} \tag{2.11}$$

and

$$E(G_i^2) = \left(\sum_j x_j\right)^{-2}\left[\sum_j x_j^2 E(Q_j^2) + \sum_{j,k} x_j x_k E(Q_j Q_k)\right]$$

so that $E(Q_j^2) = E(Q_j)$ as $Q_j = 0$ or 1, and $E(Q_j Q_k) = W(W-1)(n-1)^{-1}(n-2)^{-1}$ (that is, hypergeometric). This yields

$$E(G_i^2) = \left(\sum_j x_j\right)^{-2}\left\{\left[(n-1)^{-1}W\sum_j x_j^2 + \frac{W(W-1)}{(n-1)(n-2)}\left[\left(\sum_j x_j\right)^2 - \sum_j x_j^2\right]\right]\right\}$$

so

$$\begin{aligned} var\,(G_i) &= E(G_i^2) - E^2(G_i) \\ &= \left(\textstyle\sum_j x_j\right)^{-2}) \left[(n-1)^{-1}(n-2)^{-2}W(n-1-W)\textstyle\sum_j x_j^2 \right. \\ &\quad \left. + (n-1)^{-2}(n-2)^{-1}W(\textstyle\sum_j x_j)^2\right]. \end{aligned}$$

If we put $(\sum_j x_j)(n-1)^{-1} = Y_1$ and $(\sum_j x_j^2)(n-1)^{-1} - Y_1^2 = Y_2$, then

$$var\,(G_i) = \frac{W(n-1-W)}{(n-1)^2(n-2)}\frac{Y_2}{Y_1^2}. \tag{2.12}$$

In this paper we will not further discuss properties of $G_i(d)$ except to say that $G_i(d)$ is normal as $n \to \infty$ (from properties of sampling without replacement, that is, a Moran-type argument). In a subsequent paper (Getis and Ord, 1992), characteristics of $G_i(d)$ will be discussed for the case when normality cannot be assumed.

2.4.1 Further Development of the Statistic

The difficulty with the statistic shown in (2.9) is in its dependence on a one-or-zero weight or distance matrix. Further development of the statistic would allow i to equal j and the substitution of $d_{ij}^{-\beta}$ for W_{ij}. For example, the following formulation would replace (2.9):

$$G_i = \sum_j x_i x_j d_{ij}^{-\beta} \left(\sum_j x_i x_j \right)^{-1}, \qquad \beta > 0, \tag{2.13}$$

where $i = j$ is allowed. In (2.13) there is an obvious correspondence between both the cross-product statistic and the general form of the interaction model. The expected value would be based on the assumption that all x values were similar. Thus,

$$E(G_i) = \frac{1}{n} \sum_j d_{ij}^{-\beta}, \qquad \beta > 0. \tag{2.14}$$

As with $G_j(d)$, the new statistic G_i would have a value as follows: $0 \leq G_i \leq 1$. If i were not equal to j then the denominator of (2.14) would be $(n-1)$. Tests based on the statistic would answer the fundamental question: "are the association and the interaction between i and all j greater than chance would have it?"

A variation on (2.13) and (2.14) would focus on the single relationship between a single i and a single j. These equations are

$$G_{ij} = \frac{x_i x_j d_{ij}^{-\beta}}{x_i x_j}, \qquad \beta > 0 \tag{2.15}$$

and

$$E(G_{ij}) = d_{ij}^{-\beta}. \tag{2.16}$$

Equations (2.15) and (2.16) complete the merger of correlation and interaction formulations.

2.4.2 *Interpretation of $G_i(d)$*

In order to test hypotheses, for example, if all x_i are set to one, the pattern of x_j represents a condition of no spatial autocorrelation. In this case, the null hypothesis is: there is no difference (and thus no spatial autocorrelation) among the x_j within distance d of i. By substituting a one for each x_j, we find (2.9) and (2.12) become

$$E[G_i(d)] = \frac{W}{n-1} \tag{2.17}$$

and

$$E[varG_i(d)] = \frac{(n-1-W)^2}{(n-1)^2(n-2)}. \tag{2.18}$$

The estimated $G_i(d)$ is found by solving (2.19) by using the observed x_j values. If

$$Z = \frac{G_i(d) - E[G_i(d)]}{\{E[varG_i(d)]\}^{1/2}}$$

is positively or negatively greater than some specified level of significance, then positive or negative spatial autocorrelation are obtained. A large positive Z implies that large values (values above the mean x_j) are spatially associated. A large negative Z means that small x_j are spatially associated with one another.

When $G_i(d)$ represents a measure of interaction, the model is expanded from (2.9) to

$$G_i(d) = \left[\sum_j W_{ij}(d)x_i x_j\right]\left[\sum_j x_i x_j\right]^{-1}. \tag{2.19}$$

A null hypotheses might call for interaction no greater (or less) than one might expect when all x_j are equal. The expectations are as in (2.17) and (2.18). Rejection of the null hypothesis would indicate that there is greater (or less) interaction than expected.

2.5 Conclusion

In verbal terms, the key words differentiating the two types of models are *interaction* and *association*. The interaction implied in gravity models refers to the possible *movement* of elements at i to or from places j. In the spatial autocorrelation model, the *link* between i and j is a correlation in the sense of places having common or different specified characteristics. As the development of the spatial autocorrelation model has a statistical origin, one usually considers association as having positive or negative statistical significance. For interaction models, statistical significance is less important and prediction is more important. For interaction modelers, interest is in the flow between places, whether or not the flow are greater or less than those predicted by a normal random variable model. In this paper we were able to show that the cross-product statistic of Hubert et al. (1981) allows for a unification of the two types of models. This was accomplished by means of the development of a spatial autocorrelation statistic that serves as a measure of spatial interaction as well. An advantage to the approach taken here is that the way is now paved for the development of statistical tests on interaction theory.

Acknowledgements The author would like to thank James O Huff, Norbcrt Oppenheim, and J Keith Ord for their valuable advice.

Chapter 3
Spatial Statistical Analysis and Geographic Information Systems

Luc Anselin and Arthur Getis

This Chapter was originally published in:
Anselin, L., Getis, A. (1992) Spatial Statistical Analysis and Geographic Information Systems. The Annals of Regional Science 26(1): 1992.

Abstract In this paper, we discuss a number of general issues that pertain to the interface between GIS and spatial analysis. In particular, we focus on the various paradigms for spatial data analysis that follow from the existence of this interface. We outline a series of questions that need to be confronted in the analysis of spatial data, and the extent to which a GIS can facilitate their resolution. We also review a number of exploratory and confirmatory techniques that we feel should form the core of a spatial analysis module for a GIS.

3.1 Introduction

Space plays a central role as an organizing concept in regional science. It is therefore to be expected that the analysis of spatial data and the specialized techniques that this requires have received considerable attention in the research literature. The emphasis of this research has been on theoretical and methodological aspects, such as the role of spatial dependence and spatial heterogeneity, the effect of spatial scale, and the development of estimation methods for spatial process models (for a review, see Anselin, 1988). However, as pointed out by Anselin and Griffith (1988), the dissemination of these results to the practice of data analysis in empirical work has been rather limited. In part, this has undoubtedly been due to the lack of an easy

L. Anselin (✉)
School of Geographical Sciences, Arizona State University, Tempe, AZ, USA
e-mail: Luc.Anselin@asu.edu

A. Getis
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

Formerly published in The Annals of Regional Science (1992) 26:1.

and effective way to explicitly incorporate the spatial aspects of data. This problem is now largely eliminated, due to recent advances in the technology of geographic information systems (GIS). In spite of this, the effectiveness and importance of GIS and spatial information systems as an enabling technology is only slowly becoming recognized in regional science (Nijkamp and Rietveld, 1984; Nijkamp, 1988, 1990; Anselin, 1990). In particular, some fundamental questions related to the role played by GIS technology in defining the research agenda for spatial data analysis have not received the attention they deserve. We thought it timely to identify a number of important issues that pertain to the interface between spatial statistical analysis, GIS and regional science. Our objective is not so much to review the recent literature, but to outline and discuss alternative viewpoints, to summarize the current state of the art in spatial statistical analysis and how it relates to GIS, and to suggest directions for future research.

The focus of attention in GIS tends to be on the display, organization and simple manipulation of information in spatial data bases. As a result, most commercial GIS implementations are rather limited in what they offer in terms of statistical tools for the analysis of spatial data. This lack of analytical capacities of a GIS is by now a familiar complaint in the research literature (e.g., Goodchild, 1987; Burrough, 1990; Couclelis, 1991) and several efforts have been initiated to alleviate this situation (Abler, 1987). In those, spatial data analysis and spatial statistics are often perceived as playing a central role among the components of the analysis function in a GIS (HMSO, 1987; Gatrell, 1987; Goodchild and Brusegard, 1989; Bailey, 1990; Openshaw, 1990; Csillag, 1991; Goodchild et al., 1992). There are two aspects to this. First, there is the incorporation of spatial statistical techniques as part of the toolbox provided with a GIS, by adding statistical functions to the menu of GIS capacities, or by providing an easy link between a GIS and a statistical package. A second, and potentially more interesting aspect is the extent to which statistical and even spatial statistical techniques are appropriate for use with a GIS, and the resulting need to develop new "spatial" analysis tools. In this paper, we focus on both aspects.

We start by discussing the interface between GIS and spatial analysis, and the various paradigms for spatial data analysis that follow from this. Included is a section on the special qualities of spatial data. We next outline a series of questions that need to be confronted in the analysis of spatial data, and the extent to which a GIS can facilitate their resolution. We also briefly review a number of exploratory as well as confirmatory techniques that in our opinion should form the core of a spatial analysis module for a GIS. We close with some remarks about the role of GIS and spatial analysis in regional science research in general.

3.2 Interfacing Spatial Analysis and GIS

Traditionally, geographic information systems are considered to perform four basic functions on spatial data: input, storage, analysis, and output (Goodchild, 1987). Of these, analysis has received least attention in commercial systems. Typically, a

variety of map description and manipulation functions are defined by commercial vendors as being "spatial analysis" but they have little bearing on the use of this concept in the academic community (Couclelis, 1991). In the GIS literature spatial analysis has become narrowly defined. For example, Gatrell (1987) defines it as "the application of statistical methods to the solution of geographical research questions" (see similar uses of the concept in Openshaw, 1990; Openshaw et al., 1991; Ding and Fotheringham, 1991; Goodchild et al., 1992). Clearly, to support research in regional science (as a spatial information system or a spatial decision support system) a large set of techniques should be included under the rubric of spatial analysis, such as location-allocation and other operations research methods, urban and regional modelling, and spatial demographics.

We give a highly simplified schematic representation of the interaction between the four basic functions of GIS in Fig. 3.1. At one end of the graph is "reality," at the other the "user," concerned with policy or theory development. In between are the four functions, input (data model and measurement), storage (of data values and their location and topology), analysis (data selection, manipulation, exploration and confirmation), and output (display). Our focus is on the analysis functions and their interface with the storage function. The latter is typically associated with a relational database. In a GIS, this database not only contains information on value, but also on the location and spatial arrangement (topology) of observational units. The way in which reality is measured and structured into a spatial data base is determined by the data model, which has important implications for the types of spatial analyses that can most effectively be carried out (Peuquet, 1984, 1988; Goodchild, 1992). We return to this point below.

In the analysis module we distinguish between four important functions: one is the selection or sampling of observational units from the data base and the choice of the proper scale of analysis. The other three functions consist of increasing degrees of abstraction from the data. We call them data manipulation, exploration and confirmation. In Fig. 3.1 they are represented on the same level to illustrate the property that each of these can be considered as a self-contained module of spatial analysis, followed by output (display). However, in an idealized framework of spatial analysis, there would be a natural progression from data manipulation to exploratory analysis and confirmatory analysis, obviously with multiple feed-backs between the modules.

In our framework, data manipulation encompasses the partitioning, aggregation, overlay and interpolation procedures needed to convert the selected information into meaningful maps and surfaces. Most of these techniques represent what is understood by "spatial analysis" in commercial GIS, and they form some of the more powerful aspects of the technology in terms of the flexibility in changing observational units. Under data exploration we classify inductive approaches to elicit insight about pattern and relations from the data, without necessarily having a firm pre-conceived theoretical notion about which relations are to be expected. We could also call this "data-driven" analysis (Anselin, 1990) to stress the emphasis on "letting the data speak for themselves" (Gould, 1981). The final module is then

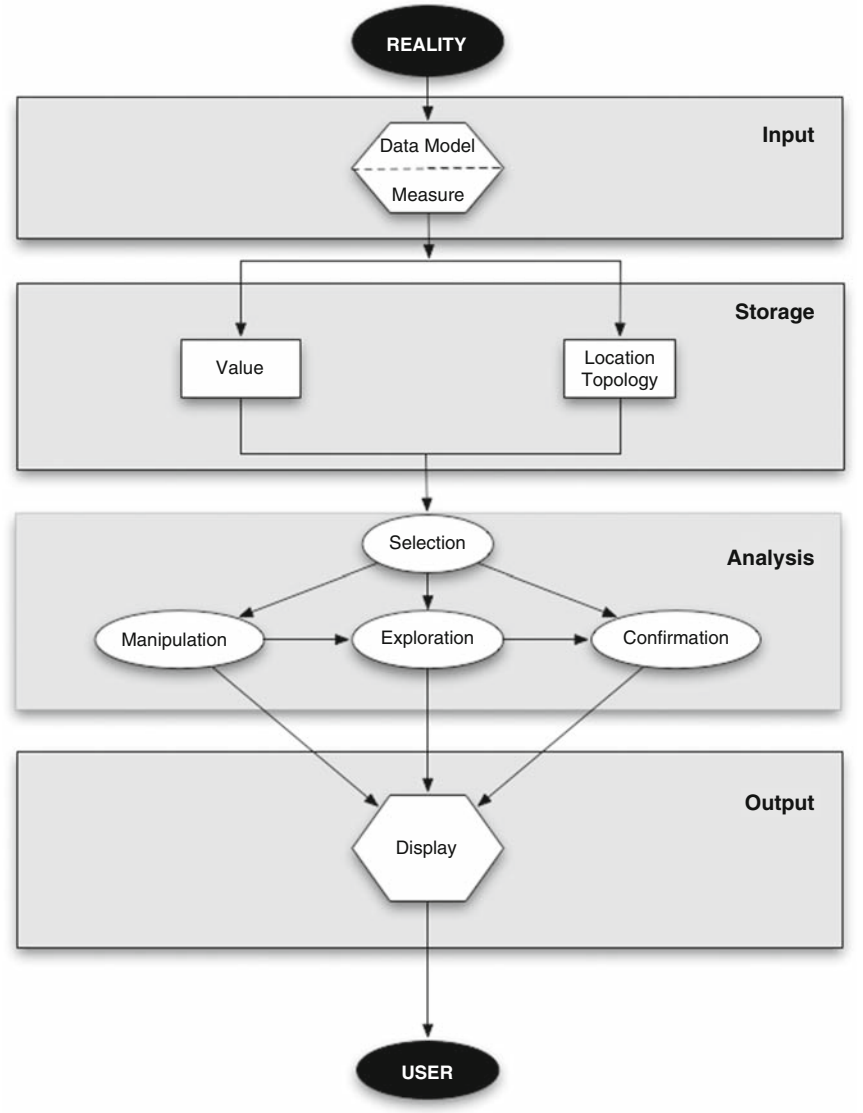


Fig. 3.1 Functions of a GIS

confirmatory analysis, where the point of departure is a theoretical notion or model ("model-driven" analysis in Anselin, 1990). This would include most of the "traditional" techniques of spatial data analysis, such as hypothesis tests, estimation of spatial process models, simulation and prediction. In principle, the type of model implemented in this module could be anything where space (location, region) is a

relevant element, and thus could encompass a wide range of urban, regional and multiregional models from the toolbox of the regional scientist.

The way in which the analysis function in a GIS (as we perceive it) is linked to the database has been forged differently. Logically, there are three ways in which this can be done:

1. Fully integrate all spatial analysis within the GIS software
2. Construct modules of spatial analysis that efficiently link with the GIS and effectively exploit the "spatial" information in the database
3. Leave the GIS and spatial analysis as two separate entities and simply import and export data in a common format between the two

The third approach is really a non-solution, since it ignores the distinctive characteristics of a spatial data base for use in spatial data analysis. Nevertheless, it seems to be the approach most taken in practice, due to the problems with proprietary data formats in commercial GIS and the limited facilities of often awkward macro languages (for an extensive discussion, see Kehris, 1990a; Bivand, 1990). Examples of this strategy are the joint use of GRASS and S for exploratory data analysis in Farley et al. (1990) and Williams et al. (1990); the combination of SPANS and SYSTAT to carry out stepwise regression in Bonham-Carter et al. (1988); and the use of ARC/INFO and BMDP for logistic regression in Warren (1990).

The second approach is similar to the so-called "modular" design in integrated regional modelling and consists of developing self-contained modules for various types of spatial analyses. These modules are then linked to the specific data structures used in a commercial GIS. They are thus not "generic," but limited to a particular combination of GIS and technique. Among many examples are the work of Walker and Moore (1988) on combining GIS and other statistical packages and the linkage of ARC/INFO and GLIM in Flowerdew et al. (1991). Most of these modules are written and compiled separately and access the data structure of the GIS by means of proprietary library functions. In general, the use of the GIS macro facilities is avoided, given its poor performance in terms of speed (Bivand, 1990; but see Ding and Fotheringham, 1991, for an example of the use of the ARC/INFO AML macro language to construct a measure of spatial association). Even though this second approach links a statistical package to a GIS, it is generally limited to simple descriptive measures, such as univariate measures of spatial association (e.g., Kehris, 1990b; Ding and Fotheringham, 1991; for an exception, see Bivand, 1990).

Finally, the first strategy ("encompassing" in the terminology of integrated modelling) is basically non-existing, due to the lack of analytical capabilities in most commercial GIS (partial exceptions are SPANS, IDRISI and GIS-PLUS/TRANSCAD). It is most closely approximated by the idea behind the "spatial analysis toolkit" of Openshaw and associates, if it were not that spatial analysis is limited to a small number of generic functions in their approach (Openshaw, 1990; Openshaw et al., 1991). In contrast, our vision of the spatial analysis function in a GIS is much wider and would include at least all of the "traditional" techniques. The determination of an unambiguous set of "generic" functions of spatial analysis is an important and still largely unresolved question.

3.2.1 The Nature of Spatial Data

The implementation of a generally useful spatial analytical capability within a GIS can be achieved once it is recognized that the solution to spatial problems inevitably must be based on the special character of spatial data. From an analytical point of view, it is more than the fact that data are spatially referenced that differentiates spatial data from other types of data. When the data are spatially referenced one must go beyond Tukey for spatial exploration and beyond standard statistics and econometrics for confirmatory analysis.

What is it that makes spatial data special? Anselin (1990) explains in considerable detail why one must treat spatial data differently than other types of data. In essence, the point is that spatial effects complicate any straightforward understanding of spatial data. "Spatial effects" has two interrelated meanings. The first is that embodied in Tobler's First Law of Geography (Tobler, 1979a), where "everything is related to everything else, but near things are more related than distant things." This simply implies that we should expect stronger relationships within and among variables that are sampled at places that are spatially near to one another rather than far from one another. The more troublesome second meaning, however, is that because of the size and configuration of spatial units we find relationships within or among variables that are due as much to the nature of the spatial units as to the nature of the variables being studied. The first type of spatial effect can be handled, for the most part, with conventional data analytical procedures, but not the second. Since all spatial data are subject to the second effect, one must take it into account when devising systems for analysis.

Spatial effects can be divided into two types: dependence and heterogeneity. Spatial dependence refers to the relationship between spatially referenced data due to the nature of the variable(s) under study and the size, shape, and configuration of the spatial units. The smaller the spatial units, the greater the probability that nearby units will be spatially dependent. If units are spatially long and narrow, the chances of spatial dependence with nearby units will be greater than if the units are more compact. Spatial heterogeneity occurs when there is a lack of spatial uniformity of the effects of spatial dependence and/or of the relationships between the variables under study. A dependence structure that is inconsistent across the study area lacks homogeneity. In a sense, then, spatial heterogeneity can be thought of as a special case of spatial dependence. It represents a complex realization of the nature of the variable(s) under study and the effects of the size, shape, and configuration of spatial units.

3.3 GIS and Perspectives on Spatial Data Analysis

Spatial analysis ranges from simple description to full-blown model-driven statistical inference. As outlined in Anselin (1990) many different perspectives can be taken towards spatial data analysis. For the purposes of our discussion, we will

classify techniques into exploratory (or data-driven), and confirmatory (or model-driven), although in practice many techniques incorporate aspects of both (Haining, 1990a).

In an exploratory data analysis (EDA, Tukey, 1977) the data are used in an inductive fashion to gain new insights. To date this is by far the most common approach towards data analysis using GIS, although it has taken at least three different forms.

In one, represented by the work of Openshaw and associates (Openshaw, 1990; Openshaw et al., 1987, 1988, 1990, 1991), the role of the analysis is limited to pure description and indication of potentially interesting spatial patterns. In fact, Openshaw goes so far as to reject most of the traditional spatial analytical methods of the type outlined in Berry and Marble (1968). Furthermore, the role of the "spatial analysis" function in GIS is restricted to pure description of map pattern, without explanation, since "what is causing the pattern is not a subject matter for the geographer" (Openshaw, 1990, 158). His geographical analysis machine (GAM) and geographical correlates exploration machine (GCEM) are computation intensive approaches to elicit patterns in the data. In that sense, they can be considered to be examples of exploratory data analysis. However, the lack of indication of "significance" and the admitted possibility that the patterns could be spurious are a far cry from the usual interpretation of EDA (Tukey, 1977; Mosteller and Tukey, 1977). In addition, while most EDA exploits the high dimensionality of data (using various clustering and graphical cross tabulation methods), Openshaw's examples so far pertain to univariate and bivariate situations. The extent to which these "machines" can be made operational and cost-effective to address more complex research questions remains to be resolved.

In a second approach to combining GIS with EDA, data are exported from a GIS into a standard statistical package for analysis (typically, S; Becker et al., 1988). This stands in sharp contrast to Openshaw's rejection of such a linkage as worthless and "an irrelevant distraction" (Openshaw et al., 1991, 788). The types of analyses that are carried out use standard EDA tools, such as box plots, Chernoff faces, Tukey stars, scatterplot matrices and hierarchical clustering (e.g., Farley et al., 1990; Williams et al., 1990). Although such techniques are very useful in generating insight into patterns and potential associations, they are a-spatial. Moreover, to the extent that measures of fit and tests of significance are included (e.g., as in added variable plots; Haining, 1990b) the presence of spatial dependence (and/or spatial heterogeneity) can easily lead to spurious conclusions. One comes close to "spatial" analysis in the work of Kehris (1990b) and Ding and Fotheringham (1991), who provide a link between a GIS and specialized routines to compute measures of spatial association. However, this is still fairly rudimentary, and a true "spatial" EDA, or ESDA (exploratory spatial data analysis) does not yet exist (see also Anselin, 1990).

A third approach consists of some recent developments in the use of statistical graphics, where the "map" is included as one of a series of dynamically linked graphs (for an extensive discussion of this concept, see Cleveland and McGilI, 1988). This is typified by the work of Haslett and associates (Haslett et al., 1990, 1991; Stringer and Haslett, 1991; see also MacDougall, 1991) on interactive graphic environments (SPIDER and REGARD) that combine a map, histogram and

scatterplot view of the data, as well as various lists. Selection of any subset of the observations in one of the dynamically linked windows affects the representation in all other windows. This allows for an intuitive and visual impression of the correspondence between value association (scatterplot) and locational association (map), although no quantification of the latter is provided. In a sense, the focus is on spatial heterogeneity (regional differentiation) rather than on spatial dependence. Ideally, of course, both should be included in a framework for ESDA.

In contrast to the recent flurry of research activity in exploratory data analysis and GIS, very little has been achieved in terms of model-driven or confirmatory analysis. Most applications are non-spatial applications of regression analysis and fail to exploit the information on the topology of the observations that is contained in a GIS. As pointed out in Anselin and Griffith (1988) and Haining (1990a), the general problem is one of lack of software to carry out the complex and nonlinear estimation and inference for spatial process models. A number of recent advances have been made in software development for spatial data analysis in the form of libraries of macro routines for commercial statistical packages (e.g., Griffith, 1988b; Griffith et al., 1990; Bivand, 1990, 1991). A self-contained spatial data analysis software package, SpaceStat, is introduced in Anselin (1991). So far, however, the linkage between this software and a GIS is very limited (e.g., the work of Bivand, 1990, which exports data from ARC/INFO into a SYSTAT module).

3.4 The GIS Data Model and Spatial Statistics

As defined by Goodchild (1992), the data model implicit in a GIS is the "discretization" of geographical reality necessitated by the nature of computing devices. Commercial GIS can be classified as following either a raster (or grid) or vector data model, i.e., a regular or irregular tesselation of the plane (see also Peuquet, 1984, 1988). The raster or vector structure defines the spatial unit of observation that can be used in spatial analysis. In the former, the unit is the grid (or other regular tesselation) and all points within the grid are assumed to take on the same value. This is an implicit form of spatial sampling. Clearly, if the grid does not exactly correspond to the spatial arrangement of values in the underlying process there will be an inherent tendency for spatial dependence. Similarly, if the scale of the grid cell has an imperfect match with the scale of the process studied, various types of misspecification may result, often called ecological fallacy or the modifiable areal unit problem (MAUP).

When a vector structure is used, the choice of the points, lines and polygons that will be represented, their spatial resolution and spatial arrangement are also an implicit form of spatial sampling. Similar to the raster approach, homogeneity is assumed within the point, line or areal unit of observation. For the latter in particular, this may only be a crude approximation and spatial dependence as well as scale problems are likely to be present.

The implied spatial sampling is also a component of alternative conceptualizations of data models, such as the distinction between a so-called "field view" ("infinite sets of tuples approximated by regions and segments," Goodchild, 1992), and an "object view" ("planes littered with objects," Goodchild, 1992). For the former it necessitates a choice of the size of the region and their relative spatial arrangement (the so-called container view of GIS, Couclelis, 1991), for the latter the selection of which "objects" will be included in the database.

It is important to keep this in mind, since this sampling process structures the database and precedes any sampling the analyst may want to carry out (the data selection module in Fig. 3.1). It is often dictated by administrative or policy (or political) concerns which may or may not be founded on "accepted" theoretical concepts of the time. Examples are the delineation of administrative regions which pre-determine the collection of many socio-economic data. In a sense then, even though spatial analysis may be exploratory, the data that are available and the way in which they are collected and arranged are often constrained by the accepted theoretical knowledge of the time (which variables are important, etc.) and its implications for spatial resolution (or, rather, the lack of interest in spatial resolution).

Obviously, this sampling will lead to a sampling error and the resulting problem of accuracy in spatial data bases (see Goodchild and Gopal, 1989, for an overview). The error in spatial databases pertains both to value (the usual problem) as well as to location and spatial arrangement (topology). As a result, what we perceive as "observations" can be conceptualized as a mixture of signal (truth) and noise (error), or, more precisely, as either a sample from an unknown population or a realization of a stochastic process. The objective of spatial analysis is to elicit information about the signal, taking into account the fact that noise is present. The presence of this "error" does not preclude a statistical methodology (and its associated inference) as argued by Openshaw (1990), but is in fact the very essence for its need.

3.5 Implementation Issues

In the implementation of a framework for spatial analysis within a GIS many issues can be addressed by means of familiar techniques. These techniques do not necessarily fit neatly within our classification of spatial analysis into the four modules of data selection, data manipulation, exploratory and confirmatory analysis (Fig. 3.1), but many methods are important in more than one module. In order to make our discussions less abstract, we next review a number of ways in which specific techniques would be incorporated into our framework. It is important to keep in mind that this will only give a general flavor of what we envisage as a general purpose spatial analysis system, since a detailed inventory of techniques is beyond the scope of this paper. Also, much remains to be addressed, and many tricky methodological problems have not yet found a satisfactory solution.

Most of the decisions made about the selection, manipulation, and analysis of spatial data can be thought of as strategies designed to avoid, specify, or account

for the effects of spatial dependence. The data available in a GIS are rarely referenced in spatial units that are appropriate for final analysis. For example, pixel data, which are highly spatially dependent, must be aggregated for land use studies. In the data selection process (the first analysis module in Fig. 3.1), the nature of the data dependence should be evaluated before a representative sample can be designated.

In a well-known study, Openshaw and Taylor (1979) summarize the results of extensive experimentation in which scale changes radically altered the correlative and autocorrelative relationships among variables. Arbia (1989) claims that it is the spatial autocorrelation, or the dependence of nearby spatial units on one another, that is responsible for changes in summary measures as scale is changed. If units are summed into larger units, the mean increases, the covariance increases, and the correlation decreases in absolute value in proportion to the change in the size of the units. In all but a few circumstances, however, the variance increases in relation to the changed size of units and to the correlation between specified neighboring units. Immediately it becomes clear that statistical tests will be affected by the chosen scale(see also Haining, 1991). This being the case, the selection of an appropriate sample is a crucial decision to which a great deal of attention must be given. This is particularly important, since it often is not clear whether the so-called modifiable areal unit problem is indeed an artifact of a particular data set, as is typically assumed, or instead should be attributed to the use of an improper model and/or technique, as argued by Tobler (1989).

If is difficult to predict how the moments of a spatial sample will change with changing scale in all but the simplest circumstances, that is, when the specification of the relationship between spatial units is simple, and therefore, not particularly interesting. In addition, when spatial units are of unequal size, weighting schemes to "equalize" them must be arbitrary and, as a consequence, one must settle for a range of test results rather than a specific value. It is clear that any multi-purpose GIS must be capable of assisting the data selection process by containing flexible clustering and aggregation algorithms.

The manipulation of spatial data (the second spatial analysis module in Fig. 3.1) may result in the creation or smoothing of a surface or the partition of data units into polygons. These types of operations rest to a large extent on the evaluation of the degree of spatial dependence present in the data. The creation of a surface by interpolation is based on the nature of trends or regularities in the data. Filtering a complex surface into a smooth one is essentially an exercise in specifying a structure for spatial dependence. In order to carry out these operations, a GIS might contain a number of measuring devices that evaluate dependence. Various cross product statistics (Hubert et al., 1981) such as Moran's I, Geary's c, the variogram, and Getis and Ord's G are all helpful in this regard (Cliff and Ord, 1981; Haining, 1990a; Cressie, 1985; Getis and Ord, 1992). In addition, smoothing techniques can be based on spectra (Rayner, 1971), trend surfaces, spatial adaptive filtering (Foster and Gorr, 1986), and smooth pycnophylactic interpolation (Tobler, 1979b), to name only a few commonly used methods.

For the creation of partitions, meaningful criteria should be based on the dependence structure of the spatial data under investigation. The techniques mentioned

in the last paragraph can be used for this purpose as can clustering algorithms. Similarly, Thiessen polygons and associated tesselation techniques are often-used partitioning devices (Boots, 1985).

Perhaps of greatest importance for the preparation and manipulation of spatial data for further analysis is the need to fill a surface with estimates of variable values when data are missing. For example, a GIS may contain data at points when the analytical interest is in areas. This problem of missing spatial data has received considerably attention (for a review, see Griffith et al., 1989) and many techniques have been implemented in operational GIS, e.g., based on kriging (Cressie, 1986; Davis, 1986; Oliver and Webster, 1990).

As pointed out before, the precise allocation of techniques to the exploratory spatial data analysis (ESDA) and confirmatory spatial data analysis (CSDA) modules is not always clear (the third and fourth spatial analysis modules in Fig. 3.1), although there are some major differentiating characteristics between the viewpoints taken in each (see Anselin, 1988; Haining, 1990a). Suffice it to say here that ESDA is that phase of analysis in which spatial patterns and structures are revealed, hypotheses proposed and models suggested. In contrast, CSDA includes the entire roster of techniques and methodologies for hypothesis testing, the determination of confidence intervals, estimation, simulation, prediction and the assessment of model fit. In ESDA one searches for structure and association, while in CSDA one evaluates the evidence. As Haining (1990a) points out, one alternates in the application of the two aspects of spatial data analysis, similar in spirit to the idea behind EDA advanced by Tukey (1977).

The various elements of ESDA include those which aid in the identification and description of patterns and variables, elicit the characteristics of variables and patterns, help determine the extent of data dependence and heterogeneity. In addition, ESDA should also allow for simple modeling, especially so that residuals can be evaluated and the selection of a "best" subset of explanatory variables can be determined.

A wide array of techniques are available for ESDA. These include the standard tools of EDA and statistical graphics, such as box plots, star plots, Chernoff faces, etc., as well as many of the measures mentioned above. In addition, pattern recognition devices such as those discussed in the artificial intelligence and spatial statistics literatures are highly relevant here, e.g., as outlined in the work of Ahuja and Schachter (1983); Pielou (1977); Ripley (1981); Boots and Getis (1987). However, the "spatial" aspects of ESDA have to date not been fully developed. In this respect, approaches that blend the analytics of the traditional techniques with the computing power and interactive graphics of some of the recent developments could show great promise.

In addition to the predominantly non-parametric approach taken in traditional EDA, one often also needs to know moments, errors, and other parametric characteristics of samples and surfaces at different scales. For example, the parameters of simple linear regression, trend surfaces, periodicities, semi-variograms and correlograms are often useful. Directional statistics and spatial ANOVA are tools that could be included in any exploratory analytical module. In addition, categorical variables

are often mapped by GIS users. Thus, logit analyses of overlapping variables would prove useful in the exploratory stage of analysis.

It is here that the distinction between ESDA and CSDA becomes difficult. Indeed, the standard tools of CSDA consist of estimation algorithms for a wide range of specifications, both linear and nonlinear. The spatial aspects of such analysis are often identified with the field of spatial econometrics, i.e., "the collection of techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models" (Anselin, 1988, 7). In essence this boils down to four broad categories of methods: (1) diagnostics for the presence of spatial dependence and spatial heterogeneity in regression analysis (this includes ANOVA and trend surface models as special cases); (2) methods to estimate and obtain inference (e.g., based on maximum likelihood, instrumental variables or bootstrap estimators) for various types of regression models for cross-sectional and space–time data that explicitly take into account spatial effects (e.g., spatial process models); (3) methods to estimate and obtain inference that are robust to the presence of spatial effects; (4) spatial measures of model validity. Although much methodological progress has been made in these areas, a number of very tricky issues remain to be resolved, such as the issue of spatial dependence in models with limited dependent variables (e.g., logit, probit and Poisson regression models), the discrimination between spatial dependence and spatial heterogeneity, nonstationarity in models for space–time data, edge effects, etc. (Anselin, 1990). To some extent then, the implementation of CSDA in a spatial analysis system is constrained by the state of the art, which to date is still unsatisfactory to be able to answer the range of questions faced by the users of a GIS.

3.6 GIS, Spatial Analysis and Regional Science

The fundamental issue in carrying out spatial analysis with a GIS is whether the "observations" (in the GIS) contain sufficient information to extract the signal and control for the noise. This is not necessarily satisfied, even though the technological sophistication of a GIS and the large size of many databases may give the opposite impression. The more one knows (or assumes) about the signal, the easier it will be to falsify preconceived notions and/or generate new hypotheses. Sometimes additional information can be obtained by combining observations on many different indicators (variables) and at many locations or spatial scales, but often even this is not sufficient. As is well known, in many instances different processes can generate observationally equivalent patterns of values. Failure to distinguish "significant" patterns or to gain insight into underlying causal relationships should not imply a rejection of the statistical methodology. Instead, more data (e.g., in the time dimension) and/or better theoretical notions may be needed. The statistical methodology provides one with a set of tools to assess the extent to which this is the case. This set of tools should not be used to the exclusion of others, but the GIS technology allows it to be complemented with powerful computation intensive approaches and

innovative visualisation. A creative combination of the "old" spatial analysis with these new technologies, to form a "new" spatial analysis (similar to the change in perspective generated by the "new" urban economics of the 1970s) has not yet been achieved.

To suggest that the recent developments in GIS have already transformed the way in which spatial analysis is carried out in the field of regional science is clearly an overstatement. Although some embrace the new technology as an innovative means to look at the world in a different way, others tend to dismiss it as just another set of fancy color graphics. The opportunity in the use of GIS is that it indeed has made previously prohibitive computationally intensive and highly visual ways of spatial analysis accessible at reasonable cost. The challenge to the GIS field is that it has not yet been able to furnish or incorporate the types of analytical tools that are needed to answer the questions posed by regional scientists. Some would argue that those are the wrong questions and that using the existing GIS tools will lead to different and more interesting questions. Our position is that the technology should be led by theoretical and methodological developments in the field itself. Does this require an abandonment of the traditional spatial analysis? Clearly, approaches that were inspired by the lack of computational and graphical resources have now become redundant, but a considerable number of fundamental insights into the nature of spatial structure, spatial dependence and spatial processes remain relevant. An effective integration of these perspectives with the new technology may go a long way toward convincing researchers in regional science and other social sciences that the special role of space which underlies the essence of the field merits its own analytical toolbox. We suggest that spatial statistical analysis should play a central role in this toolbox.

Acknowledgements Anselin's research was supported in part by Grant SES-8721875 from the National Science Foundation and by the National Center for Geographic Information and Analysis (NSF Grant SES-8810917).

Chapter 4
Whose Hand on the Tiller? Revisiting "Spatial Statistical Analysis and GIS"

Michael F. Goodchild

Abstract Anselin and Getis argue in their 1992 paper "Spatial statistical analysis and GIS" that the development of a toolbox of spatial analytic techniques should be directed by the scientists whose work defines the need for such a toolbox. The field of GIS has changed fundamentally since 1992 as a result of new technical developments, including a general move away from the map as the defining metaphor, the influence of the Internet and the World Wide Web, and changes in the practice of software engineering. Science as a whole has also changed, towards a more collaborative model that is more dependent on computational infrastructure. The impacts of space on the methodology of science are also better understood. The Anselin and Getis paper was remarkably prescient in its identification of the major issues that continue to affect the relationship between spatial analysis and GIS. Institutional issues continue to frame the relationship between GIS and spatial analysis, and are best addressed through partnerships.

4.1 Introduction

In 1992 Luc Anselin and Art Getis published "Spatial statistical analysis and GIS" in the Annals of Regional Science (Anselin and Getis, 1992). The paper was one of a number of explorations of the relationship between spatial analysis and geographic information systems (GIS) that appeared at about that time (Burrough, 1990; Ding and Fotheringham, 1992; Fotheringham and Rogerson, 1994; Goodchild, 1987; Goodchild et al., 1992; Openshaw et al., 1990), driven perhaps by a perception that the evident growth in GIS as a large and complex software application might eventually benefit science, by providing improved access to many of the tools that researchers had developed over the previous decades. Indeed GIS was being widely

M. F. Goodchild
National Center for Geographic Information and Analysis and Department of Geography, University of California, Santa Barbara, CA, USA
e-mail: good@geog.ucsb.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*, Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_4,

hailed in this light (Abler, 1987), and it seemed only a matter of time before virtually all of the known methods of spatial analysis would be available in a single, massive, widely available toolbox.

The Anselin and Getis paper was distinguished from others in the genre by its focus on regional science, and by its strongly stated belief that future developments in GIS should be driven by scientists in the substantive fields of application, including regional science:

> Some would argue that using the existing GIS tools will lead to different and more interesting questions. Our position is that the technology should be led by theoretical and methodological developments in (regional science) itself. (Anselin and Getis, 1992, p. 30)

In other words, future developments in GIS should be directed by those most familiar with the kinds of questions it was ideally able to answer; the idea that the GIS tail might wag the regional science dog was clearly not as attractive.

Approximately a decade and a half has elapsed since the paper was written, and the landscape of GIS has been changed almost beyond recognition. In this paper I attempt to bring the discussion up to date, to answer some of the questions raised by the authors, and to pose new ones that seem to have arisen recently – and to do so within a somewhat larger context that includes geography and other social and environmental sciences in addition to the regional science of the original paper. The next section provides a brief and I hope accurate summary of the main arguments of the Anselin and Getis paper. This is followed by a review of events and trends within the field of GIS since 1992, and then by a review of major trends affecting science, and particularly the social and environmental sciences. The final section of the paper updates the 1992 comments on the role of regional scientists in directing the development of GIS, by suggesting specific parts that substantive scientists can play in the evolving saga of software development and support.

4.2 Synopsis of the 1992 Arguments

The model that underpins the Anselin and Getis discussion is particularly elegant in the way it combines widely accepted organizing principles from both GIS and spatial analysis. It is reproduced in Fig. 3.1. It defines GIS operations in four classes – input, storage, analysis, and output – following many extensive discussions of GIS functionality (see, e.g., Maguire, 1991; Maguire and Dangermond, 1991). The analysis function has been seen as the most important by several authors (see, for example, Cowen, 1988), who have argued that it represents the vital transformation of data into useful information, making visible what might otherwise be invisible to the user. The authors then introduced a four-way classification of analysis, into selection, manipulation, exploration, and confirmation. The sharp distinction between analysis and display that was implicit in the classification of GIS was softened somewhat, as display was seen as inherent in each of manipulation, exploration, and confirmation. Selection encompassed sampling and other aspects of what

today might be termed ontology, or in statistical terms support, while manipulation included many of the lower-grade operations of GIS such as buffering, spatial joins, and point-in-polygon operations, that are nonetheless termed analysis by many GIS developers (see, e.g., Mitchell, 1999).

The authors identified several software strategies that were used circa 1992 to deliver the functionality of spatial analysis using GIS. Some techniques were included directly in the GIS itself as core functions, though it was recognized that this strategy was not always the fastest computationally. Separate modules could be constructed that performed the more elaborate forms of spatial analysis, and coupled to the GIS for purposes of data input, storage, and display, either through the exchange of files or through access to a common database (Nyerges, 1993).

Anselin and Getis adopted the distinction between exploratory and confirmatory analysis, while accepting that the distinction was often somewhat blurred. Exploratory analysis was seen as data driven and inductive, while confirmatory analysis was theory driven and deductive. Exploratory spatial data analysis (ESDA) was a new and exciting field in 1992, building on the improved interactive graphics capabilities that became available first in the Macintosh and in Unix workstations in the 1980s, and later in the PC. Exploratory analysis also included the data-driven approach being advocated by Openshaw (Openshaw et al., 1987, 1988, 1990), which saw the search for pattern as an essential activity that could be essentially independent of any theoretical framework. Today we might find echoes of the same strategy in the use of techniques borrowed from artificial intelligence, including neural nets and self-organizing maps (Fischer, 1997; Fischer and Leung, 1998; Skupin and Hagelman, 2005).

The fourth section of the paper looked at the relationship between GIS data models and spatial statistics, making the point that discretization of space was an essential step in any representation, and that it impacted the results of analysis in ways that were largely out of the analyst's control. The Modifiable Areal Unit Problem (MAUP) had caught the attention of many researchers in the 1980s, who were dismayed to discover how much their results were affected by the choice of areal units, and therefore by decisions made in statistical agencies that were far from neutral in this respect. The impact of such scale-related effects was discussed at greater length in the fifth section of the paper, titled Implementation Issues, which emphasized the importance of spatial dependence in determining the effects of scale change, and in confounding any attempt to use spatial methods in a confirmatory mode.

In their concluding section, Anselin and Getis argued that to serve the needs of regional science, GIS needed to be complemented with powerful computational intensive approaches and innovative visualisation. A creative combination of the "old" spatial analysis with these new technologies to form a "new" spatial analysis has not yet been achieved. It was "clearly an overstatement" to suggest that "the recent developments in GIS have already transformed the way spatial analysis is carried out in the field of regional science." "An effective integration (of computationally intensive approaches and innovative visualization) with the new technology may go a long way toward convincing researchers in regional science and other

social sciences that the special role of space which underlies the essence of the field merits its own analytical toolbox."

4.3 GIS Since 1992

The term GIS had rather different connotations in 1992 than it does today. New technologies have driven a substantial restructuring of the field, as has massive growth in numerous application areas. In 1992 GIS connoted a single, monolithic software package running on a stand-alone workstation or perhaps a local-area network, and analogous to Microsoft Word or Excel. Its purpose was to relieve its user of tasks that would be too tedious, repetitive, time-consuming, complex, or inaccurate if performed by hand. Analysis of maps and map data (Maling, 1989) has all of these characteristics, and a technology that promised apparently effortless analysis at the speed of light was clearly attractive. In 1992 several vendors offered such packages, including Intergraph, MapInfo, ESRI, Wild, Caliper, and Tydac. Some product differentiation was evident, between large, expensive packages targeted at corporations, government departments, and universities, and small, cheaper packages designed for single users. Several packages were available from the academic community, including Idrisi from Clark University, and in general these offered a more sophisticated range of analytic tools but had more severe limitations in terms of speed and capacity. Finally GRASS offered a large number of useful analytic functions in an open-source package originally developed by the US Army Corps of Engineers.

Many factors have contributed to a changing perspective on GIS over the past 15 years. First, early developments in GIS were built on the map as the primary source of input, and the first applications of commercial GIS were accordingly in areas heavily dependent on maps, such as resource management and forestry. By the early 1990s, however, it had become clear that much could be gained by adding geographic references to the records contained in the otherwise non-spatial but massive databases of utility companies, marketing companies, and other commercial sectors. The leading database vendors, including Oracle and Informix, developed extensions to handle such spatially enabled records and to support simple queries, such as "select for me all of the hotels within 10 miles of this airport" – and the major GIS vendors responded with products of their own, such as ESRI's ArcSDE.

Second, the Internet became the dominant network, and the World Wide Web emerged as a dramatically effective application. In the second edition of their survey of GIS, Longley et al. (1999) admit to having missed completely the impact that this would have on the field when they wrote the conclusion of their first edition in 1991 (Maguire, 1991). Data had always been something of an Achilles heel to GIS, because the conversion of paper maps to digital records was difficult to automate and frustratingly error-prone. But the Internet opened an apparently unlimited potential for sharing of digital geographic data, first with such early applications as ftp and WAIS (Nebert, 1993) and later with Web applications. Today many

GIS users have long forgotten the tedium of map digitizing, and instead rely on a vast array of clearinghouses, data archives, digital libraries, and data warehouses to supply the basic data on which GIS depends. The state of the art is perhaps best represented by the Geospatial One-Stop (http://www.geodata.gov), a US Government effort to provide a geo-portal, a single point of entry to a vast, distributed network of geographic data sources. At the same time, much effort has gone into developing the standards and protocols needed to achieve interoperability between suppliers and users of data, with their many different data formats and GIS software products. The Open Geospatial Consortium (http://www.opengeospatial.org) has developed and promulgated many standards, and the Federal Geographic Data Committee (http://www.fgdc.gov) continues to have a very effective influence. The term National Spatial Data Infrastructure was coined in the early 1990s (Mapping Science Committee, 1993) to describe a vision of an interoperable, networked future, and was authorized by Presidential Order in 1994. Similar efforts are under way in many other countries, and within the European Union (http://inspire.jrc.it).

This networking of GIS data access has had a fundamental effect on the software, requiring that it contain the data conversion routines needed for interoperability, as well as the tools to support search for data over a distributed network. It has become essential to support metadata, the descriptive catalog entries that now allow researchers to specify needs and to search, evaluate, and retrieve suitable data sets. Increasingly, students of GIS find themselves learning as much about the management of geographic data as about the software that runs in their local machine, and there have been calls to drop the S and to refer to the field simply as GI.

The Internet has had a further influence in allowing the processing steps that underlie GIS to occur at locations remote from the user. Client-server systems divide the processing tasks between the user's own local machine, the client, and a remote and probably more powerful machine known as the server. In the extreme, all processing occurs on the server and the client is reduced to a "dumb terminal" or a simple Web browser such as Microsoft Explorer. This arrangement is often favored by government agencies, which make limited GIS services available in this way and thus avoid having to distribute copies of their data. Many Web sites now offer mapping and simple forms of analysis via servers, allowing users to visualize data, make summary extracts, and even perform simple statistical analyses. The GIS software industry now offers a range of server-based products, such as ESRI's ArcIMS, to provide the necessary services, and MapServer (http://mapserver.gis.umn.edu) is a popular public-domain product. The term GIServices is sometimes used to distinguish such client-server configurations from the more traditional GISystem.

In principle any GIS function, and any form of spatial analysis, could be offered as a GIService. In practice a limited number of simple GIS functions are now available as commercial or public services, including geocoding (the task of converting street addresses to coordinates, e.g., http://www.travelgis.com/geocode/), wayfinding (the task of generating driving directions from an origin to a destination, e.g., http://www.mapquest.com), gazetteer lookup (the task of converting a placename to coordinates, e.g., http://www.alexandria.ucsb.edu), location analysis (e.g., a demographic analysis of the neighborhood of a potential retail site, offered by many

market research companies), and spatial search (e.g., search for hotels within a given distance of an airport, http://www.expedia.com). Some of these are based on a viable business model, and it seems that the functions most likely to be offered as GIServices are those that (1) depend on access to a large and rapidly changing database that individuals would not be able to keep up to date, and (2) require a level of complexity of analysis that would be beyond the average user. There is clearly potential here for providing the kinds of exploratory and confirmatory spatial analysis discussed by Anselin and Getis as GIServices, but to date there appear to be no obvious examples in practice.

Third, the past 15 years have seen a radical rethinking of practice in software engineering. The single, monolithic package of 1992 has largely been replaced by re-usable software components that can be mixed and matched for specific applications. Microsoft's COM/OLE and .Net environments allow software components to be mixed across boundaries that once seemed impenetrable, such as between Excel and ESRI's ArcGIS (Ungerer and Goodchild, 2002), allowing applications to combine functions from both under the direction of scripts written in standard languages such as Visual Basic or Python. Instead of a single package, vendors now offer a variety of extensions for specific purposes, leading to a growing sense of segmentation in the GIS market. This sense has been reinforced by the advent of object-oriented data modeling (Zeiler, 1999; Arctur and Zeiler, 2004), which allows the basic data objects of a GIS (points, lines, and areas) to be specialized for particular application domains.

Today the old sense of GIS as a well-defined type of software has been significantly eroded. Some vendors have chosen to adopt descriptions that better reflect their target application domains, while others emphasize the ability to customize a range of products to specific needs. An adjective seems more appropriate than a noun in this new more complex world, and in recent years the term geospatial appears to have gained some traction, as evidenced by the recent renaming of what was the Open GIS Consortium and by extensive restructuring within the US Geological Survey.

The geospatial world of today is clearly a much broader domain of data, tools, services, and concepts than the limited GIS world of 1992. Many of the statistical packages now include limited support for spatial analysis, and an extensive set of geospatial tools can be found in Matlab. Links have been constructed between GIS packages and simulation environments such as Stella and Repast, and what were previously considered functions exclusive to GIS, such as simulated fly-by, are now readily available in Google Earth and other geo-browsers. A social or environmental scientist needing tools to support spatial analysis now has a vast array of options, many of which would no longer involve anything recognized as a GIS. The idea examined by Anselin and Getis in 1992, of whether the particular needs of spatial analysis justify the development of a special toolbox, no longer seems as relevant – the importance and special nature of spatial analysis is clearly demonstrated by the vast array of data, tools, and services that are now available, whether or not they are labeled as GIS.

Moreover, the days of monolithic software environments are now over, and it is no longer appropriate to envision a day when all forms of spatial analysis will be available in a single toolbox. Just as geo-portals provide a single point of entry to a complex, distributed array of data, it seems appropriate to envision a day when a single point of entry will provide access to a set of distributed, interoperable tools and services. Already some sites provide searchable directories of tools (see, e.g., http://www.csiss.org), and some geo-portals provide searchable directories of GIServices in addition to data (see, e.g., http://www.geographynetwork.com). But as yet there are no universal standards for catalogs of tools (Crosier et al., 2003), or standards for interoperability. Thus the necessary supporting infrastructure for such a vision still needs substantial work.

4.4 The Broader Context

4.4.1 Cyberinfrastructure

The topics discussed in the previous section – the need for interoperability, access to distributed data, and GIServices – are representative of much broader trends within computation and within the infrastructure it provides for science. Various names have been given to this new, distributed form of computing, including cyberinfrastructure, and much has been written about its potential. In the UK the Economic and Social Research Council has made substantial investments in building a cyberinfrastructure for social science through the e-Social Science initiative (http://www.ncess.ac.uk), and in the USA the National Science Foundation (NSF) has recently established an Office of Cyberinfrastructure (http://www.nsf.gov).

One of the most definitive documents in this general trend is the report of the Report of the Blue-Ribbon Advisory Panel on Cyberinfrastructure, generally known as the Atkins report after the study committee's chair (Atkins et al., 2003). It argues persuasively that cyberinfrastructure can not only improve the ability of scientists to do what they already do, but also underpin a new kind of science that is more computationally intensive, more collaborative, and more visual than before. The report distinguishes between two traditional kinds of science, one inductive and data driven and the other deductive and theory driven, and argues that cyberinfrastructure can enable a third kind that has elements of both, and that relies heavily on simulation. This vision has clear echoes of the comments made by Anselin and Getis in 1992, regarding the need for computationally intensive tools and better methods of visualization, and it reflects trends that have been evident within the social and environmental sciences for at least the past decade. It could be argued that this is a case of the computational tail wagging the substantive dog. But the Atkins report clearly leaves the role of defining needs to the substantive sciences and to the fundamental questions they need to answer.

In principle one might expect GIS to provide an ideal environment for this third kind of science. It is digital, and it allows rules and functions representing process to be simulated on the data contained in its databases. But in practice, as Anselin and Getis noted in 1992, the macro or scripting languages used by GIS to program complex tasks have had a reputation for being awkward to use and slow to execute. Traditional GIS software has been designed for the comparatively leisurely pace of analysis, rather than the computationally intensive pace of simulation.

However this situation has changed significantly in the past decade. Certain GIS packages, notably GRASS and the PCRaster package produced by the University of Utrecht (http://pcraster.geo.uu.nl) have been designed specifically to support simulation, and reference has already been made to recent efforts to link GIS to simulation environments such as Repast. Moreover, the comments in the previous section suggest that it is no longer useful to ask whether some somewhat arbitrarily delimited type of software known as GIS can or cannot perform simulation – rather, the broader set of geospatial tools able to support simulation is now clearly rich and powerful (Maguire et al., 2005).

Nevertheless, the community identified with GIS remains somewhat limited in its perspective, and has not yet built the kinds of bridges to larger communities focused on simulation, in domains such as atmospheric science, geophysics, or oceanography, all of which are clearly embedded in geographic space. Moreover, simulation technology is very advanced in domains that deal with other spaces, including aeronautical and structural engineering. As yet, no GIS vendor supports the representation of partial differential equations, either as finite difference or finite element approximations, and there is little support for simulation of processes in three spatial dimensions.

4.4.2 New Methodologies

Cyberinfrastructure is only one manifestation of other, more fundamental trends in science. It argues for the use of information and communication technologies (ICT) to support collaboration between scientists, reflecting a general trend away from the single-investigator style of science to a more cooperative mode in which teams of specialists combine their expertise to solve complex problems. The software systems needed to support massive simulation of complex systems, such as those that underlie global climate models, are far too elaborate for any one person to know completely. To most scientists, computational tools will have to be black boxes, defined by their inputs and outputs rather than by their contents, and no one individual will be able to meet the traditional standard of scientific reporting: to provide sufficient detail to allow another scientist to replicate the experiment.

Unfortunately this problem is all too well known to scientists who use computational tools developed and marketed by the commercial sector, and particularly geospatial tools. The GIS industry has been driven largely by commercial applications, and science has always been only a small fraction of its market. Standards

of documentation that are adequate for commercial applications may well fail the scientific test of providing sufficient detail for replicability. For example, it is frequently impossible to determine exactly how certain GIS functions operate, or to determine the degree of noise introduced in certain operations such as raster-vector conversion or projection change.

The widespread use of GIS is raising other issues of a methodological nature. Anselin and Getis refer to spatial dependence and spatial heterogeneity as the two defining characteristics of spatial data, noting the problems that the former causes for many methods of statistical analysis. Tobler's First Law of Geography (Tobler, 1970) asserts that positive spatial dependence is endemic in geographic data, in obvious violation of the independence assumption of many statistical tests, and ensuring that any significance test that results in the acceptance of a null hypothesis of no spatial dependence is almost certainly making a Type II statistical error.

Spatial heterogeneity raises even greater methodological issues, because it suggests that any attempt to find universal principles that apply everywhere on the Earth's surface is fundamentally problematic. Instead, analysis should focus on estimating and interpreting the inevitable variation in parameters, adopting a methodological position that is somewhere between the traditional nomothetic and idiographic extremes. So-called local or place-based analysis is more consistent with this position, and is now represented by a range of techniques that includes Anselin's LISA (Anselin, 1995) and the geographically weighted regression of Fotheringham et al. (2002).

Substantial progress has been made in building geospatial tools to support spatial analysis in the presence of both spatial dependence and spatial heterogeneity over the past 15 years. ESRI's ArcGIS supports a range of geostatistical techniques through its Geostatistical Analyst, and extensions are also available for the analysis of point patterns and other spatial statistical tests. Anselin's own GeoDa (http://geoda.uiuc.edu) has been developed as a stand-alone package but using standard GIS data structures, and has proven very popular (by March 2006 over 10,000 copies had been downloaded). The issues raised by Anselin and Getis in 1992 are now rapidly becoming the basis of standard practice.

4.5 Whose Hand on the Tiller?

Anselin and Getis concluded with comments about the need for regional scientists to play a central role in directing the future of GIS. As noted in the previous section, GIS has been a largely commercial product, and its development has been driven by its market, where the emphasis has been on such applications as forestry and utility management rather than on science. More effort has gone into providing rapid responses to simple queries from massive databases than into the kinds of sophisticated spatial analysis demanded by the social and environmental sciences.

The commercial nature of GIS is both a blessing and a curse, of course. On the one hand it is doubtful if GIS could have survived commercially based on the

science market alone, and the existence of large commercial applications has provided access by scientists to a well-engineered, well-supported range of products. The distinction between commercial and science applications is also quite blurred, with many companies and agencies making use of the more sophisticated analytic functions of GIS and many universities using GIS to maintain inventories of their own physical plant. On the other hand, academics are traditionally leery of commercial motives, and as noted in the previous section there are clear differences between modes of operation in the commercial world and the norms of the scientific method.

Moreover, many significant contributions have been made by the non-commercial GIS sector, by such products as Idrisi, PCRaster, and GRASS. But these always run the risk of being attacked as unfair competition, and the early support of GRASS by the US Army Corps of Engineers had to be withdrawn, at least in part for this reason.

Nevertheless, there are several reasons why the academic sector continues to exert a greater amount of influence on the future of GIS than its significance as a market would suggest. First, academics are also educators, and have a strong influence on the knowledge and predilections of future generations. A commercial GIS vendor looking to long-term success will clearly want to curry favor with the academic sector, by discounting software and providing other forms of support. If academics feel the need to educate students in particular techniques of spatial analysis then commercial GIS vendors may well choose to support those techniques, whether or not the market for them is viable. Sophisticated features also add to the perceived attractiveness of a product, whether or not they are actually used, just as many customers demand features in other products that they will never learn to use.

Second, the academic sector has a well-recognized duty to reflect and comment on all aspects of human society, and commentary on GIS and its impacts has become a significant subject of scholarship in geography, planning, and related disciplines. Academics have pointed to the importance of uncertainty, and the inherent vulnerability of GIS when uncertain results are used to regulate the use of land (Goodchild and Gopal, 1989); and they have questioned many of the assumptions underlying GIS analysis (Pickles, 1995). While the impacts of these critiques may not be apparent in GIS software products, they have undoubtedly altered the context in which GIS is used.

Finally, the academic sector is the primary source of fundamental innovation in GIS. The research domain variously known as geographic information science, geomatics, geovisualization, or geoinformatics has grown rapidly in the past decade, and has spawned journals, conferences, and organizations (http://www.ucgis.org). GIScience research has led to new insights into such fundamental issues as scale, new methods of representation that go far beyond the traditional GIS data models, and new techniques for addressing uncertainty. Such research increasingly provides the framework for new geospatial standards, and for improvements in user interface design.

4.6 Conclusion

In hindsight, Anselin and Getis provided a remarkably prescient analysis of the issues surrounding GIS and spatial analysis in the early 1990s. They accurately anticipated the growing need for tools that were more computationally intensive, with better support for advanced visualization, in a clear call for what today has become a much more broadly based interest in cyberinfrastructure. They also recognized the importance of spatial dependence and spatial heterogeneity as the defining characteristics of geographic data, and their implications in the need for new forms of analysis and new supporting tools.

The software landscape has changed dramatically since 1992. GIS is no longer a monolithic, stand-alone application, but instead encompasses a range of product types and a range of new functions, and new, broader terms such as geospatial have been adopted to try to capture this new complexity. The idea of a single toolbox has been replaced by the concept of interoperable software modules, operating in a mix of hardware architectures that ranges from the hand-held PDA (personal digital assistant) through the desktop workstation to the remote server. New techniques of search allow researchers to discover and access these modules quickly, though much work remains to be done on appropriate catalogs and methods of description.

Like almost all of us, Anselin and Getis missed the massive transformations caused by the Internet and the Web that began in 1993 and still continue today, and the impacts that these transformations are having on the practice of science. They were correct, however, in arguing that it should be the basic questions addressed by science, and the needs of scientists for techniques to answer those questions, that should drive future development of tools. The GIS steam-engine has always been driven by a range of applications, only some of which are concerned with fundamental science questions, and the needs of researchers have always had to battle with the needs of other users. Anselin's own GeoDa is an excellent example of a feasible strategy in this environment – a package for sophisticated spatial analysis that is designed to interoperate with standard GIS products, but is designed and programmed by a team of scientists. But it is difficult for the academic marketplace to provide an income stream, and such projects must therefore be funded by grants, an erratic source at best. Software produced by academic teams is rarely engineered to the same standards as commercial software, and it is difficult to provide the same kinds of support.

In short, while much has happened in the decade and a half since the publication of the Anselin and Getis paper, its institutional context remains much as it was in 1992. The commercial GIS sector is large and growing, and able to produce and support complex "industrial-strength" software. The academic sector has only a limited ability to influence the commercial sector's directions, in integrating the kinds of tools needed to support research, and in ensuring adherence to the norms of the scientific method. As in many other fields, it is clear that innovative partnerships represent the best way forward, integrating the work of the two sectors and allowing their different objectives to be harmonized.

Chapter 5
Spatial Interaction and Spatial Autocorrelation

Manfred M. Fischer, Martin Reismann, and Thomas Scherngell

Abstract The objective is to combine insights from two research traditions, spatial interaction modelling and spatial autocorrelation modelling, to deal with the issue of spatial autocorrelation in spatial interaction data analysis. *First*, the problem is addressed from an exploratory perspective for which a generalisation of the Getis–Ord G statistic is presented. This statistic may yield interesting insights into the processes that give rise to spatial association between residual flows. *Second*, the log-additive spatial interaction model is extended to spatial econometric origin-destination flow models consistent with an error structure that reflects origin, destination or origin-destination autoregressive spatial dependence. The models are formally equivalent to conventional spatial regression models. But they differ in terms of the data analysed and the way in which the spatial weights matrix is defined.

5.1 Introduction

Spatial econometric theory and practice have been dominated by a focus on object data. In economic analysis these objects correspond to economic agents with discrete locations in space, such as addresses, census tracts and regions. In contrast, spatial interaction or flow data pertain to measurements each of which is associated with a link or a pair of origin-destination locations that represent points or areas in space. While there is a voluminous literature on spatial autocorrelation with a typical focus of interest in the specification and estimation of models for cross-sectional object data, there is scant attention paid to its counterpart in spatial interaction data. For example, there is no reference to spatial interaction data in any of the commonly cited spatial econometric or statistic texts, such as Anselin (1988), Griffith (1988a) or Cressie (1993).

In contrast, there is the field of spatial interaction modelling which has a long and distinguished history that has led to the emergence of three major schools

M. M. Fischer (✉), M. Reismann, and T. Scherngell
Vienna University of Economics and Business, Wien, Austria
e-mail: manfred.fischer@wu.ac.at

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*,
Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_5,

of analytical thought: the macroscopic school based upon a statistical equilibrium approach (see Wilson, 1967), the microscopic school based on a choice-theoretic approach (see Sen and Smith, 1995) and the geocomputational school based upon the neural network approach that perceives spatial interaction models as universal function approximators (see Fischer, 2002). In these schools there is a deep-seated view that spatial interaction implies movement of tangible entities such as persons and commodities or intangible ones such as information and knowledge across space, and that this has little to do with spatial association (Getis, 1991).

The focus in this chapter is on the spatial autocorrelation problem in spatial interaction data analysis. The objective is to combine insights from both research traditions, spatial interaction modelling and spatial autocorrelation modelling. *First*, we address the problem from an exploratory perspective, and present a generalisation of the Getis–Ord G statistic. This statistic may yield interesting insights into the processes that give rise to spatial association between residual flows in that it enables to detect local non-stationarity. *Second*, we shift the attention to the model driven mode of analysis,[1] and extend the log-additive model of spatial interaction that has served as the workhorse in spatial interaction analysis, to a general class of spatial econometric origin-destination flow models consistent with an error structure that reflects origin and/or destination autoregressive spatial dependence. These models represent not only extensions of the spatial interaction models but also extensions of the spatial regression models introduced by Anselin (1988), Griffith (1988a) and others. The paper derives the log likelihood function for these models and suggests a computational approach that relies on sparse matrix Cholesky algorithms to efficiently compute the maximum likelihood estimates. An example using patent citation data that capture knowledge flows across 112 European regions serves to illustrate the way the G_{ij} statistic and the spatial regression origin-destination flow models might be applied.

5.2 The Classical View on Spatial Interactions

Spatial interaction data represent phenomena that may be described in their most general terms as interactions between actors and opportunities distributed among some relevant geographic space. Such interactions may involve movements of individuals from one location to another. Interactions may also involve flows of knowledge as captured by means of patent citations. Here inventors may be the relevant actors, and the possible receivers of knowledge may be considered as the relevant opportunities.

[1] This draws heavily on previous work by Fischer et al. (2006a,b).

5.2.1 The General Spatial Interaction Model

Suppose we have a spatial system consisting of n regions. Flows, Y_{ij}, are observed between each pair (i, j) of regions where i $(i = 1, \ldots, n)$ denotes the origin region and j $(j = 1, \ldots, n)$ the destination region of interaction. The Y_{ij} are assumed to be independent random variables. They are sampled from a specified probability distribution dependent upon some mean, say μ_{ij}. Let y_{ij} denote the observed flows and assume that no a priori information is given on the row and column totals of the flow matrix $[y_{ij}]$. Then this so-called *unconstrained* spatial interaction problem may be solved by modelling the observed flows y_{ij}, according to a statistical spatial interaction model of the general form

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \tag{5.1}$$

where $E[Y_{ij}] = \mu_{ij}$, ε_{ij} is an error term about the mean, and μ_{ij} is specified as a function of covariates measuring the characteristics of origin regions, destination regions, and their separation:

$$\mu_{ij} = C\ A_i\ B_j\ F_{ij}. \tag{5.2}$$

A_i is called origin factor that characterises the origin region i, B_j destination factor that characterises the destination region j, and F_{ij} the separation function that measures separation between i and j. It is implicitly assumed that A, B and F are positive, and that the factors A and B are independent of F. μ_{ij} is the expected mean interaction flow for a given separation configuration defined by F_{ij}. C denotes a constant of proportionality.[2]

Various specific models can be derived from (5.2) specifying A, B and F appropriately. It is general practice to represent the variables A_i and B_j as power functions of the form

$$A_i = A(a_i, \alpha_1) = a_i^{\alpha_1}, \tag{5.3}$$
$$B_j = B(b_j, \alpha_2) = b_j^{\alpha_2}, \tag{5.4}$$

where a_i and b_j denote some appropriate origin and destination variables. α_1 and α_2 are parameters to be estimated. The separation function F_{ij} that constitutes the very core of spatial interaction models may be specified as

[2] Note that if the origin totals of $[y_{ij}]$ are a priori given, C has to be replaced by an origin specific constant C_i that is given by $C_i = y_{i\bullet}\,[A_i \sum_j B_j F_{ij}]^{-1}$, so that $\sum_j \mu_{ij} = y_{i\bullet}$ is guaranteed. If the destination totals are a priori given, C has to be replaced by a destination specific constant C_j that is given by $C_j = y_{\cdot j}[B_j \sum_i A_i F_{ij}]^{-1}$ which ensures $\sum_i \mu_{ij} = y_{\cdot j}$. The first case is called the production constrained case of spatial interaction and the second the attraction constrained case. In the production-attraction constrained case $y_{i\cdot}$ and $y_{\cdot j}$ are given. Generally, it is assumed here that $C = C_i C_j$, where C_i is dependent on all the C_j, and vice versa (see Fischer and Reggiani, 2004 for more details).

$$F_{ij} = \exp\left[\sum_{k=1}^{K} \beta_k \, d_{ij}^{(k)}\right] \tag{5.5}$$

with $d_{ij}^{(k)}$ representing a measure of separation between i and j. The β_k are unknown parameters. There are various ways to estimate the parameters $\alpha_1, \alpha_2, \beta_1, \ldots, \beta_K$. Maximum likelihood and least squares are among the most commonly used (see Sen and Smith, 1995 for a general discussion).

5.2.2 The Log-Additive Model of Spatial Interaction

From the positivity of the factors A, B and F follows that spatial interaction models defined by (5.1)–(5.5) can equivalently be expressed as a log-additive model[3] of the form

$$y(i,j) = \alpha_0 + \alpha_1 a(i) + \alpha_2 b(j) + \sum_{k=1}^{K} \beta_k \overset{(k)}{d}(i,j) + \varepsilon(i,j), \tag{5.6}$$

where $y(i,j) \equiv \ln \mu_{ij}$, $\alpha_0 \equiv \ln C$, $a(i) \equiv \ln a_i$, $b(j) \equiv \ln b_j$, $d(i,j) \equiv d_{ij}$ and $\varepsilon(i,j) \equiv \varepsilon_{ij}$. Under the assumption that $a(i)$ and $b(j)$ are measured without error and that the error terms $\varepsilon(i,j)$ are independent identically distributed with zero mean and constant variance,[4] we obtain the ordinary least squares estimator, say $\hat{\boldsymbol{\gamma}}$, for $\boldsymbol{\gamma} = (\alpha_0, \alpha_1, \alpha_2, \beta_1, \ldots, \beta_K)^T$ as the solution to the matrix equation

$$(\boldsymbol{X}^T \boldsymbol{X})\hat{\boldsymbol{\gamma}} = \boldsymbol{X}^T \boldsymbol{y}, \tag{5.7}$$

where

$$\boldsymbol{X}^T = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ a(1) & \cdots & a(1) & \cdots & a(n) & \cdots & a(n) \\ b(1) & \cdots & b(n) & \cdots & b(1) & \cdots & b(n) \\ \overset{(1)}{d}(1,1) & \cdots & \overset{(1)}{d}(1,n) & \cdots & \overset{(1)}{d}(n,1) & \cdots & \overset{(1)}{d}(n,n) \\ \vdots & & \vdots & & \vdots & & \vdots \\ \overset{(k)}{d}(1,1) & \cdots & \overset{(k)}{d}(1,n) & \cdots & \overset{(k)}{d}(n,1) & \cdots & \overset{(k)}{d}(n,n) \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ 1 & \cdots & n & \cdots & N-n+1 & \cdots & N \end{bmatrix} \tag{5.8}$$

[3] Note in some cases $y_{ij} = 0$ indicating the absence of flows from i to j. This leads to the so-called zero problem since the logarithm is then undefined. There are several pragmatic solutions to this problem with adding a small constant to the non-zero elements of $[y_{ij}]$ being widely used. In this contribution we have decided to add 0.08 in such cases.

[4] This assumption implies that the individual flows, $y(i,j)$, from origin i to destination j are independent from each other and that interaction flows between any pairs of regions are independent from flows between any other pairs of regions. A violation of this assumption leads to spatial autocorrelation or heterogeneity.

and

$$\boldsymbol{y}^T = [y(1,1), \ldots, y(1,n), \ldots, y(n,1), \ldots, y(n,n)] \tag{5.9}$$

given that the $N = n^2$ vector $\boldsymbol{d}^{(k)} = [\overset{(k)}{d}(1,1), \ldots, \overset{(k)}{d}(n,n)]^T$ is the vectorised form of the n-by-n separation matrix $[d_{ij}^{(k)}]$, $\boldsymbol{a} = [a(1), \ldots, a(1), \ldots, a(n), \ldots, a(n)]^T$ and $\boldsymbol{b} = [b(1), \ldots, b(n), \ldots, b(1), \ldots, b(n)]^T$ are N-by-1 vectors, appropriately indexed over the $N = n^2$ values. The N-by-1 vector $\varepsilon = [\varepsilon(1,1), \ldots, \varepsilon(n,n)]^T$ denotes the vectorised form of $[\varepsilon_{ij}]$.

Depending upon the assumptions made about the variance-covariance matrix of ε, the estimators derived from (5.7) may or may not be efficient. But (5.7) is an unbiased equation (Durbin, 1960) in the sense that

$$E[\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\gamma}}] = \boldsymbol{X}^T\boldsymbol{X}E[\hat{\boldsymbol{\gamma}}] = \boldsymbol{X}^T E[\boldsymbol{y}] = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\gamma}, \tag{5.10}$$

where $E[.]$ denotes the expectation operator. From (5.10) we see that

$$E[\hat{\boldsymbol{\gamma}}] = \boldsymbol{\gamma} \tag{5.11}$$

provided that $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists. That is, the data must not be perfectly collinear. This result holds whatever dispersion matrix, $\sigma^2\boldsymbol{V}$, is postulated for the disturbance, $\boldsymbol{\varepsilon}$.

A violation of the assumptions made may lead to two separate problems: (1) spatial autocorrelation among the $\boldsymbol{X}$-variables, and (2) spatial autocorrelation among the residuals, $\boldsymbol{\varepsilon}$. Both problems may well arise, but neither implies the other. If (1) holds, this will affect the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, or $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$ in general, and thus the variance estimates of the coefficients. If (2) holds, then the basic assumption of a scalar dispersion matrix for the disturbances, $\boldsymbol{\varepsilon}$, is violated, that is $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V} \neq \boldsymbol{I}$. Thus, there will be an extra $\boldsymbol{V}$ matrix in the expressions, and generalised rather than ordinary least squares should be used. If $\boldsymbol{V}$ is unknown, as is generally the case, then some form of iterative generalised least squares should be performed. In this case the parameter estimates will be consistent, but not necessarily unbiased. But this is true for every regression problem, and the residuals should be tested for spatial autocorrelation (Cliff et al., 1974).

The problem of modelling spatially autocorrelated residuals in spatial interaction models has been largely neglected so far.[5] This may be because spatial interaction models are more complex than linear regressions for object data, and each region is associated with several values as an origin and/or destination so that specification of the autocorrelation structure is less obvious. In the next section we suggest spatial weights structures that enable to model dependence between origin-destination pairs in a fashion consistent with conventional spatial autoregressive models.

[5] There are only very few exceptions, most notably the studies by LeSage and Pace (2005); Bolduc et al. (1992, 1995); Brandsma and Kelletaper (1979).

5.3 Spatial Autocorrelation Among Flows

While the conventional notion of spatial autocorrelation in a cross-sectional spatial regression context involving a sample of n regions relies on a n-by-n spatial weights matrix, in a spatial interaction context where the $\boldsymbol{y}$ vector reflects flows between origin-destinations we need to extend the notion of spatial autocorrelation to a concept of spatial connectivity between origin-destination pairs of regions. In analogy to the conventional case of object data, we loosely define spatial autocorrelation among flows, say $y(i,j)$, between origin-destination pairs (i,j) of regions as coincidence of $y(i,j)$-value similarity with what may be termed interaction similarity, i.e., similarity of flows (i,j) and (r,s) in the four-dimensional space $\{i,j;r,s|i,j,r,s = 1,\ldots,n\}$. A crucial issue in this definition of spatial autocorrelation is the notion of interaction similarity, or the determination of those dyads for which the values of the random flow variable are correlated. Such dyads may be referred to as "neighbours." A convenient way to define interaction similarity is by means of a four-dimensional spatial weights matrix that defines for each dyad (i,j) a relevant "neighbourhood set."

5.3.1 Specification of a Spatial Weights Matrix

A spatial weights matrix, say $\boldsymbol{W}^* = [w^*(i,j;r,s)]$, in a spatial flow context is a N-by-N positive matrix which expresses for each dyad (i,j) those dyads (r,s) that belong to its neighbourhood set as non-zero elements. Formally, $w^*(i,j;r,s) > 0$ when (i,j) and (r,s) are neighbours, and $w^*(i,j;r,s) = 0$ otherwise. By convention, the diagonal elements of the weights matrix are set to zero. The specification of which elements are non-zero is a matter of considerable arbitrariness. But this is true for the case of area data as well.

In this contribution we distinguish between origin-based and destination-based similarity. In the first case, flows (i,j) and (r,s) are similar if origin regions i and r are neighbours (see *Case A* in Fig. 5.1), while in the second case flows (i,j) and (r,s) are considered to be similar if destination region s is an element of the

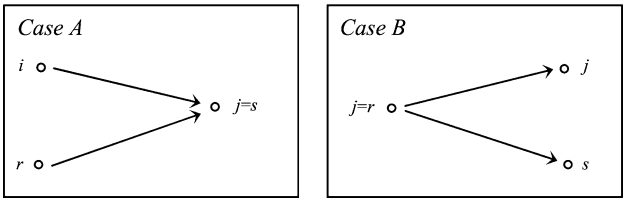


Fig. 5.1 Origin-based and destination-based similarity. The flows (i,j) and (r,s) are origin-based similar in *Case A* since the origin regions i and r are contiguous spatial units, and destination-based similar in *Case B* since the destination regions j and s are contiguous spatial units

neighbourhood set of destination region j (see *Case B* in Fig. 5.1). Intuitively, it seems plausible that forces leading to flows from a region i to a particular destination j may create similar flows from neighbours to this origin region to the same destination region j, as well as forces leading to flows of an origin i to a destination j may create similar flows to neighbouring destinations (see LeSage and Pace, 2005).

Formally, we define the following N-by-N binary spatial weights matrix ${}^{o}\boldsymbol{W}^{*} = [{}^{o}w^{*}(i,j;r,s)]$ to capture origin-based spatial dependence:

$$ {}^{o}w^{*}(i,j;r,s) = \begin{cases} 1 & \text{if } j = s \text{ and } w_{ir} = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.12} $$

where w_{ir} is the element of a conventional n-by-n first order contiguity matrix that defines whether the origin regions i and r are contiguous or not:

$$ w_{ir} = \begin{cases} 1 & \text{if } i \neq r, \text{ and } i \text{ and } r \text{ have a common border,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.13} $$

This spatial weights matrix specifies an origin-based neighbourhood set for each origin-destination pair (i,j). An element ${}^{o}w^{*}(i,j;r,s)$ defines an origin-destination pair (r,s) as being a "neighbour" of (i,j) if the origin regions i and r are contiguous spatial units, and $j = s$. By convention, an origin-destination pair (i,j) is not a neighbour to itself so that the diagonal elements are zero. It is convenient to work with a row-standardised form of ${}^{o}\boldsymbol{W}^{*}$. In order to achieve this, each element of the matrix has to be divided by the respective row sum so that the row elements of the standardised matrix ${}^{o}\boldsymbol{W}$ sum to one:

$$ {}^{o}w(i,j;r,s) = \frac{{}^{o}w^{*}(i,j;r,s)}{\sum\limits_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} {}^{o}w^{*}(i,j;r,s)}. \tag{5.14} $$

In analogy, we define a row-standardised destination-based N-by-N spatial weights matrix ${}^{d}\boldsymbol{W} = [{}^{d}w(i,j;r,s)]$ in which we capture destination-based dependence as follows:

$$ {}^{d}w(i,j;r,s) = \frac{{}^{d}w^{*}(i,j;r,s)}{\sum\limits_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} {}^{d}w^{*}(i,j;r,s)} \tag{5.15} $$

with

$$ {}^{d}w^{*}(i,j;r,s) = \begin{cases} 1 & \text{if } i = r \text{ and } w_{js} = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.16} $$

and

$$w_{js} = \begin{cases} 1 & \text{if} j \neq s, \text{ and } j \text{ and } s \text{ have a common border,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.17}$$

The spatial weights matrix ${}^{o}\boldsymbol{W} + {}^{d}\boldsymbol{W}$ specifies origin-to-destination dependence in which case it is assumed that flows from origin region i to destination region j are accompanied by similar flows from neighbours of region i to neighbours of region j.

Given n regions and, thus, N observations, it takes five steps to generate a spatial weights matrix of the form ${}^{o}\boldsymbol{W}$, ${}^{d}\boldsymbol{W}$ or ${}^{o}\boldsymbol{W} + {}^{d}\boldsymbol{W}$:

1. Partitioning the surface into Voronoi diagrams or Thiessen polygons
2. Constructing the polygon topology
3. Generating the n-by-n binary first-order contiguity matrix $[w_{ir}]$ or $[w_{js}]$
4. Producing the N-by-N binary spatial weights matrix $\boldsymbol{W}^* = [w^*(i,j;r,s)]$
5. Standardising the N-by-N matrix $\boldsymbol{W}^*$ to arrive at the N-by-N row-standardised spatial weights matrix $\boldsymbol{W}$

Steps (1) and (2) are computationally complex, but many GIS packages provide functions for these tasks. The other steps can easily be performed by means of standard tools. Note that the spatial weights matrix $\boldsymbol{W}$ is an extremely large, sparse matrix. A sample of $n = 100$ regions, for example, would result in a $\boldsymbol{W}$-matrix of dimension N-by-N where $N = 10{,}000$. Only a very small portion (generally less than 1%) of the elements is non-zero.

5.3.2 The Generalised Getis–Ord Statistic

Recently a number of statistics, called local spatial statistics, have been developed for object data. They identify the association between a single value in each region and its neighbours. These statistics are well suited to identify the existence of pockets or "hot spots" and to assess assumptions of stationarity. Prominent examples are provided by the family of G statistics introduced by Ord and Getis (1995) to allow for non-binary spatial weights matrices and non-positive values.

Although the Getis–Ord's G_i statistic was defined in the context of scalar observations in each region, it is easily generalised to flow data (see Berglund and Karlström, 1999). Let (r,s) denote the flow from origin region r to destination region s, and $e(r,s)$ the residual flow associated with the origin-destination pair (r,s). Let, moreover, $[w(i,j;r,s)]$ be a spatial weights matrix. Then it is straightforward to define the flow autocorrelation statistic $G_{ij}(\boldsymbol{W})$ as[6]

[6] Including the residual flow from i to j defines the G^*_{ij} in analogy with the G^*_i statistic. The null hypothesis appropriate for the G_{ij} statistic requires that $e(i,j)$ be excluded from the summation, while the null hypothesis appropriate for the G^*_{ij} statistic requires that $e(i,j)$ itself be summed together with the values of the "neighbouring" (r,s) dyads.

$$G_{ij}(\boldsymbol{W}) = \frac{\sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} w(i,j;r,s)\, e(r,s)}{\sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} e(r,s)}. \tag{5.18}$$

In this and in all subsequent formulations where we use summation signs, (r, s) does not equal (i, j) so that there is no self-interaction.

The estimated $G_{ij}(\boldsymbol{W})$ is found by solving (5.18) using (5.6) to obtain $e(r, s) = y(r, s) - \hat{y}(r, s)$. The statistic can be transformed to a standard variate which asymptotically follows a normal distribution. The standardised z-value is obtained in the usual manner as

$$z[G_{ij}(\boldsymbol{W})] = \frac{G_{ij}(\boldsymbol{W}) - E[G_{ij}(\boldsymbol{W})]}{\{var[G_{ij}(\boldsymbol{W})]\}^{\frac{1}{2}}}. \tag{5.19}$$

If $z[G_{ij}]$ is positively or negatively greater than some specified level of significance, then positive or negative spatial autocorrelation is obtained. A large positive $z[G_{ij}]$ implies that the residual flow from i to j is surrounded by relatively large residual flows from r to s whereas a negative $z[G_{ij}]$ indicates that the flow is surrounded by relatively small residual flows.

Under the assumption of a normal error, the expected value and the variance of the statistic are given as

$$E[G_{ij}(\boldsymbol{W})] = \frac{\sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} w(i,j;r,s)E[e(r,s)]}{\sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} e(r,s)} = \frac{1}{N-1}W(i,j) \tag{5.20}$$

and

$$var[G_{ij}(\boldsymbol{W})] = \frac{W(i,j)[N-1-W(i,j)]}{(N-1)^2(N-2)}\left[\frac{Q_2(i,j)}{Q_1(i,j)^2}\right], \tag{5.21}$$

where

$$W(i,j) = \sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} w(i,j;r,s), \tag{5.22}$$

$$Q_1(i,j) = \frac{1}{N-1} \sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} e(r,s), \tag{5.23}$$

and

$$Q_2(i,j) = \frac{1}{N-1} \sum_{\substack{r,s=1 \\ (r,s)\neq(i,j)}}^{N} e(r,s)^2 - Q_1(i,j)^2. \tag{5.24}$$

Note that the G_{ij} statistic is formally equivalent to the G_i statistic. It differs from the latter in terms of the data analysed and the manner in which the spatial weights matrix is defined. As it is true for object data analysis, the flow statistic G_{ij} is a convenient exploratory tool to identify the existence of pockets or "hot spots" and to assess assumptions of stationarity.

5.4 A Spatial Econometric View on Spatial Interactions

One way to deal with the issue of spatially autocorrelated errors is to respecify the log-additive model of spatial interaction by modelling spatial error dependence with an autoregressive error structure. The resulting error covariance will be non-spherical, and thus OLS estimates while still unbiased will be inefficient.

5.4.1 The General Spatial Econometric Model of Origin-Destination Flows

To improve the precision of inference and the prediction accuracy we introduce a spatial error structure into spatial interaction models. The resulting models may be viewed as an extension of the conventional spatial regression models described in Anselin (1988). Different spatial processes lead to different error covariances. The most common specification in conventional spatial regression models is a first order autoregressive spatial process in the error terms. Using this specification results into the following general spatial econometric origin-destination flow model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{5.25}$$

i.e., a log-additive spatial interaction model (as defined in Sect. 5.2.2) with a N-by-1 error vector $\boldsymbol{\varepsilon}$ given by

$$\boldsymbol{\varepsilon} = \rho \boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{\eta}, \tag{5.26}$$

where $\boldsymbol{y}$ denotes the N-by-1 vector of observations on the interaction variable, $\boldsymbol{X}$ is the $(N, K+3)$-matrix of observations on the explanatory variables including the origin, destination and separation variables, and the intercept. $\boldsymbol{\gamma}$ is the associated $(K+3)$-by-1 parameter vector, $\boldsymbol{\eta}$ a N-by-1 vector of independent identically distributed random errors with zero mean and equal variances. Usually we shall take them to be normally distributed so that

$$\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}). \tag{5.27}$$

$\boldsymbol{W}$ is a row-standardised N-by-N-spatial weights matrix as specified in the previous section. All diagonal elements of the matrix are zero by construction. ρ is a scalar parameter that reflects the magnitude of spatial dependence and is typically referred to as the spatial autoregressive parameter. It is assumed that $|\rho| < 1$. If $|\rho| > 1$, the model would be explosive and non-stationary.

If $|\rho| < 1$ and $\boldsymbol{I} - \rho\boldsymbol{W}$ is non-singular, then

$$\boldsymbol{\varepsilon} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\eta} \tag{5.28}$$

follows from (5.26). Thus, $E[\boldsymbol{\varepsilon}] = 0$ and $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \boldsymbol{\Omega}(\rho)$ where

$$\boldsymbol{\Omega}(\rho) = \sigma^2\boldsymbol{V}(\rho) \tag{5.29}$$

with

$$\boldsymbol{V}(\rho) = [(\boldsymbol{I} - \rho\boldsymbol{W})^T(\boldsymbol{I} - \rho\boldsymbol{W})]^{-1}. \tag{5.30}$$

To ensure the variance-covariance matrix $\boldsymbol{\Omega}(\rho)$ is positive definite and, thus, non-singular, the autocorrelation parameter ρ has to be within its feasible range $\rho \in]\lambda_{\text{min}}^{-1}\lambda_{\text{max}}^{-1}[$, where λ_{min} and λ_{max} are the smallest and largest eigenvalues of $\boldsymbol{W}$, respectively, with $\lambda_{\text{min}} < 0 < \lambda_{\text{max}}$ (Hepple, 1995). Since the row sums of $\boldsymbol{W}$ are bounded uniformly in absolute value by one, the Perron–Frobenius theorem (Cox and Miller, 1965, p. 120), tells us that $\lambda_{\text{max}} = 1$ and $-1 \leq \lambda_{\text{min}}$, so that we have the restriction of $|\rho| < 1$ for the stationarity of the spatial origin-destination flow models of type (5.25) with an error structure specification given by (5.26).

Different Model Specifications. The general model defined by (5.25)–(5.26) leads to three specifications that are of specific interest in this contribution. These are derived from the following spatial weights matrices:

1. $\boldsymbol{W} = {}^o\boldsymbol{W}$ results in a model specification which reflects origin-based autoregressive spatial error dependence.
2. $\boldsymbol{W} = {}^d\boldsymbol{W}$ leads to a model specification which reflects destination-based autoregressive spatial error dependence.
3. $\boldsymbol{W} = {}^o\boldsymbol{W} + {}^d\boldsymbol{W}$ generates a model form which reflects autoregressive spatial dependence at both origins and destinations.[7]

5.4.2 The Log Likelihood Function and Maximum Likelihood Estimation

Given $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, the log-likelihood function for ρ, $\boldsymbol{\gamma}$ and σ^2 is

$$\mathfrak{L}(\boldsymbol{\gamma}, \rho, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) + \ln|\boldsymbol{A}(\rho)| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma})^T[\boldsymbol{A}(\rho)]^T\boldsymbol{A}(\rho)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}) \tag{5.31}$$

[7] In this case we implicitly assume that there is a lack of separability between the impacts of origin and destination interaction effects in favour of a cumulative impact.

with

$$\boldsymbol{A}(\rho) = \boldsymbol{I} - \rho \boldsymbol{W}, \tag{5.32}$$

where $|\boldsymbol{A}|$ is the determinant of $\boldsymbol{A}$. The log-likelihood function can be maximised with respect to ρ, $\boldsymbol{\gamma}$ and σ^2 simultaneously to obtain the maximum likelihood (ML) estimates. This optimisation can be difficult if the number of explanatory variables is large. Alternatively, the ML estimates can be obtained from the concentrated likelihood function. First, (5.31) is solved for the values of $\boldsymbol{\gamma}$ and σ^2 to maximise $\mathfrak{L}$, conditional on ρ. These are

$$\tilde{\boldsymbol{\gamma}}(\rho) = (\boldsymbol{X}^T[\boldsymbol{A}(\rho)]^T\boldsymbol{A}(\rho)\boldsymbol{X})^{-1}\boldsymbol{X}^T([\boldsymbol{A}(\rho)]^T\boldsymbol{A}(\rho))^{-1}\boldsymbol{y}, \tag{5.33}$$

$$\tilde{\sigma}(\rho)^2 = \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\gamma}}(\rho))^T[\boldsymbol{A}(\rho)]^T\boldsymbol{A}(\rho)(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\gamma}}(\rho)). \tag{5.34}$$

The concentrated likelihood function is then obtained by substituting (5.33) and (5.34) into (5.31):

$$\mathfrak{L}(\rho) = K + \ln|\boldsymbol{A}(\rho)| - N\ln\left[(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\gamma}}(\rho))^T[\boldsymbol{A}(\rho)]^T\boldsymbol{A}(\rho)(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\gamma}}(\rho))\right], \tag{5.35}$$

where K is a constant not depending on ρ. The concentrated likelihood function is maximised with respect to ρ. ML estimates of $\boldsymbol{\gamma}$ and σ^2 ($\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\sigma}}^2$, respectively) are found by substituting the optimal value of ρ into (5.33) and (5.34). Since (5.35) has only one parameter, its optimisation can be performed with a more sophisticated optimisation technique or with a simple one-dimensional search over $(\lambda_{\min}^{-1}, 1)$.

The major difficulty in numerical maximisation of the concentrated likelihood function is the necessity of evaluating the N-by-N log-determinant of $\boldsymbol{A}$ at each step. The evaluation becomes computationally intensive when N is not small. To minimise the computational burden Ord (1975) suggested to exploit the log-determinant of $\boldsymbol{A}$ in terms of the eigenvalues $\lambda_i (i = 1, \ldots, n)$ of the spatial weights matrix $\boldsymbol{W}$:

$$\ln|\boldsymbol{A}| = \sum_{i=1}^{N}\ln(1 - \rho\lambda_i). \tag{5.36}$$

The advantage of (5.36) to compute the log-determinant is that the $\{\lambda_i | i = 1, \ldots, n\}$ can be determined once and for all at the outset of the optimisation process, and not repeatedly at each of the iteration steps. But the eigenvalue approach to computing the log-determinant still leaves the researcher with the task of determining the eigenvalues of the N-by-N spatial weights matrix. Unless $\boldsymbol{W}$ has a particular structure, this task is typically very challenging especially if n and, hence, N is large. $\boldsymbol{W}$ is a large sparse N-by-N matrix. A sparse matrix is one that has only a very small proportion of non-zero elements. Unfortunately, common procedures for sparse eigenvalue problems, such as rank-one modification or band-peeling, have limited appeal since the required structure is unrealistic for spatial weights matrices (see Smirnov and Anselin, 2001).

Matrix factorisation procedures for sparse matrices in general and sparse matrix Cholesky factorisation techniques in particular provide very powerful procedures to quickly evaluate the log-determinant of $\boldsymbol{I} - \rho \boldsymbol{W}$. For a row standardised spatial weights matrix, such as ${}^{o}\boldsymbol{W} + {}^{d}\boldsymbol{W}$, transferred to symmetric form, the Cholesky factorisation consists of solving

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{T}, \tag{5.37}$$

where $\boldsymbol{L}$ is a lower triangular matrix, referred to as the Cholesky factor of $\boldsymbol{A}$. Since the determinant of a triangular matrix only involves the diagonal elements, the log-determinant is easily computed as

$$\ln |\boldsymbol{A}| = \sum_{i=1}^{N} \ln l_{ii} \tag{5.38}$$

with $l_{ii} (i = 1, \ldots, N)$ denoting the diagonal elements of $\boldsymbol{L}$. Cholesky factorisation is very efficient when the sparse structure of $\boldsymbol{A}$ is exploited by reordering rows and columns to yield a sparse factor matrix $\boldsymbol{L}$ while preserving the numerical characteristics of $\boldsymbol{A}$. Good reordering techniques can reduce the complexity of the factorisation for $\boldsymbol{A}$ from $O(N^3)$ down to $O(N^2)$ (see Smirnov and Anselin, 2001). This approach puts the maximum likelihood solution of spatial econometric flow models into the computational reach for larger origin-destination interaction systems.

5.5 An Empirical Example

To illustrate the way the G_{ij} statistic and the spatial econometric origin-destination flow models might be applied, patent citation data are used. Such data recorded in patent documents are widely recognised as a rich and fruitful source for the study of the spatial dimension of innovations and technological change[8] (see, for example, Jaffe and Trajtenberg, 2002; Fischer et al., 2006b).

5.5.1 The Data

We use interregional patent citation flows as the dependent variable in the models. The data specifically relate to citations between European high-tech patents. By European patents we mean patent applications at the European Patent Office assigned to high-tech firms located in Europe. High-technology is defined to

[8] Each patent contains highly detailed information on the invention itself, the technological area to which it belongs, the inventors, the assignee and the technological antecedents of the invention. Because patents record the residence of the inventors they are an invaluable resource for studying how knowledge flows are affected by the geography.

involve the ISIC-sectors aerospace (ISIC 3845), electronics-telecommunication (ISIC 3832), computers and office equipment (ISIC 3825) and pharmaceuticals (ISIC 3522). Self-citations, i.e., citations from patents assigned to the same firm, have been excluded, given our interest in pure externalities as evidenced by *interfirm* knowledge spillovers.

It is well known that the observation of citations is subject to a truncation bias, because we observe citations for only a portion of the "life" of an invention. To avoid this bias in the analysis we have established a five-years-window to count citations to a patent.[9] The observation period is 1985–1997 with respect to cited patents and 1990–2002 with respect to citing patents. The sample used in this contribution is restricted to inventors located in $n = 112$ regions, generally NUTS-2 regions, covering the core of "Old Europe" including Germany (38 regions), France (21 regions), Italy (20 regions), the Benelux countries (24 regions), Austria (8 regions) and Switzerland (1 region), resulting into N = 12,432 interregional flows.[10]

Subject to caveats relative to the relationship between citations and spillovers, these data allow us to identify and measure spatial separation effects to interregional knowledge spillovers in this interaction system of 112 regions. Our interest is focused on $K = 3$ measures: $\boldsymbol{d}^{(1)}$ is a N-by-1 vector that represents geographic distance measured in terms of the great circle distance (in kilometers) between the regions represented by their economic centres, $\boldsymbol{d}^{(2)}$ is a N-by-1 country dummy variable vector that represents border effects measured in terms of the existence of country borders between the regions.

As we consider the distance effect on interregional patent citations it is important to control for technological proximity between regions, as geographical distance could be just proxying for technological proximity. To do this we use a technological proximity index s_{ij} that defines the proximity between regions i and j in technology space. We divide the high-technology patents into 55 technological subclasses following the International Patent Code classification system. Each region is assigned a (55, 1)-technology vector that measures the share of patenting in each of the technological subclasses for the region. The technological proximity index s_{ij} between regions i and j is given by the uncentred correlation of their technological vectors. Two regions that patent exactly in the same proportion in each subclass have an index equal to one, while two regions patenting only in different subclasses have an index equal to zero. This index is appealing because it allows for a continuous measure of technological distance by the transformation $d_{ij} = 1 - s_{ij}$. Appropriate ordering leads to the N-by-1 vector $\boldsymbol{d}^{(3)}$.

The product $A_i B_j$ in (5.2) may be interpreted simply as the number of distinct (i, j)-interactions which are possible. Thus, it is reasonable to measure the origin

[9] For details on data construction see Fischer et al. (2006b). The trouble is that to obtain citations by any one patent application in year t, one needs to search the references made by all patent applications after year t. This is called the inversion problem that arises due to the fact that the original data on citations come in the form of citations made, whereas we need dyads of cited and citing patents to construct interregional patent citations flows.

[10] Note that intraregional flows are left out of consideration. In the case of cross-regional inventor teams the procedure of multiple full counting has been applied.

factor in terms of the number of patents in the knowledge producing region i in the time period 1985–1997, and the destination factor in terms of the number of patents in the knowledge absorbing region j in the time period 1990–2002 to produce the N-by-1 vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

5.5.2 Application of the G_{ij} Statistic

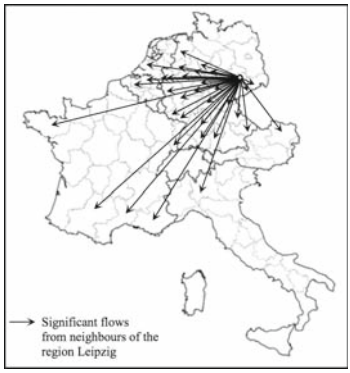
We briefly illustrate the application of the G_{ij} statistic as a tool for identifying non-stationarities, using ${}^{o}\boldsymbol{W}$ as spatial weights matrix and error flows generated by the conventional log-additive spatial interaction model (5.6). The parameter estimates and their associated probability levels are summarised in Table 5.1, along with some performance measures. The estimated coefficients indicate that the origin, destination and separation variables are highly significant with appropriate signs of the coefficients. The results provide clear evidence that geographical distance is important, but less so than national borders. Most important is technological proximity. This suggests that interregional knowledge flows seem to follow particular technological trajectories, and occur most often between regions that are located close to each other in both technological and national spaces.

Table 5.1 The log-additive spatial interaction model: parameter estimates and performance measures ($N = 12{,}432$)

	Ordinary least squares estimation
Parameter estimates (p-values in brackets)	
Constant [α_0]	−4.851 (0.000)
Origin variable [α_1]	0.594 (0.000)
Destination variable [α_2]	0.562 (0.000)
Geographical distance [β_1]	−0.181 (0.000)
Country border [β_2]	−0.592 (0.000)
Technological distance [β_3]	−2.364 (0.000)
Performance measures	
Adjusted R^2	0.563
Log likelihood	−21,024.128
Sigma square	1.723

Notes: The spatial interaction model is defined by (5.6) where the standard assumptions for least squares estimation hold. $\boldsymbol{a}$ is measured in terms of the log number of patents (1985–1997) in the knowledge producing region i, $\boldsymbol{b}$ in terms of the log number of patents (1990–2002) in the knowledge absorbing region j, $\boldsymbol{d}^{(1)}$ represents geographic distance measured in terms of the great circle distance (in kilometers) between the economic centres of the regions i and j, $\boldsymbol{d}^{(2)}$ border effects measured in terms of the existence of country borders between regions i and j, and $\boldsymbol{d}^{(3)}$ technological distance measured in terms of the technological proximity index s_{ij}. Model performance is given in terms of the adjusted R^2, the log likelihood and sigma square (the error variance).

(a) High residual flows from neighbours of origin i to the destination region Île-de-France



(b) Low residual flows from neighbours of the region Leipzig to a destination region j

Fig. 5.2 Flows with selected significant $G_{ij}({}^{o}\boldsymbol{W})$ statistic: (a) indicating high residual flows from neighbours of origin i to the destination region Île-de-France; and (b) indicating low residual flows from neighbours of the region Leipzig to a destination region j

An origin-destination pair (i, j) with a significant $G_{ij}({}^{o}\boldsymbol{W})$ statistic indicates that there is local non-stationarity in flows from a neighbourhood of origin region i to destination region j. A closer look at the G_{ij} statistic scores reveals that some destination regions and some origin regions have many significant statistics of the same sign. Figure 5.2 provides a visualisation for two specific cases. Figure 5.2 represents the case where Île-de-France is the destination region, and Fig. 5.2 the case where Leipzig is the origin region. These two figures clearly indicate that there is spatial non-stationarity in flows when these regions are the destination and the origin, respectively.

Given regions with many significant $G_{ij}({}^{o}\boldsymbol{W})$ origin or destination factors in the spatial interaction model may be misspecified. We apply the $G^{*}_{i\bullet}({}^{o}\boldsymbol{W})$ and $G^{*}_{\bullet j}({}^{o}\boldsymbol{W})$ statistics[11] which may be – as Berglund and Karlström (1999) have shown – considered as test statistics for local non-stationarities in the origin and destination factors, respectively. Examining the $G^{*}_{i\bullet}({}^{o}\boldsymbol{W})$ and $G^{*}_{\bullet j}({}^{o}\boldsymbol{W})$ statistics we find that the regions Île-de-France and Leipzig indeed have significant high $G^{*}_{\bullet j}$ and significant low $G^{*}_{i\bullet}$ statistics, respectively. This might indicate that the significant $G_{ij}({}^{o}\boldsymbol{W})$ statistics for these regions could be attributed to non-stationarity in the destination and origin factors. The heterogeneity in the residual flows is confirmed by a Breusch–Pagan test. Its value of 568.1 is fairly significant ($p = 0.000$).

[11] For the definition of the G^{*}_{ij} statistic see (5.18) together with Footnote 6. The subscript dot signifies that a sum is taken with respect to the subscript replaced by the dot.

5.5.3 ML Estimates of the Origin-Destination Spatial Econometric Flow Models

The pattern of residuals examined by means of the G_{ij} statistic clearly indicates spatial autocorrelation, so that it makes sense to fit the models given by (5.25)–(5.26). This section reports the ML estimates of the three spatial econometric model specifications that reflect origin, destination and origin-destination spatial dependence of flows, respectively. We used the *spdep package*[12] running on a Sun Fire V250 with 1.28 GHz and 8 GB RAM to create the spatial weight matrices from polygon contiguities, and the *errorsarlm* procedure based on Ng and Peyton's (1993) sparse matrix Cholesky algorithm to generate the ML estimates for the models. Using this algorithm, computation of the maximum likelihood estimates of the spatial econometric models that reflect origin, destination and origin-destination dependence of flows required only between 56 and 836 s, a remarkably short time considering that each iteration required calculating the determinant of a 12,432-by-12,432 matrix.

Table 5.2 contains the parameter estimates of the three model specifications and their associated log likelihoods. Moving from the log-additive spatial interaction model to the flow model reflecting spatial dependence at the origins (destinations) raises the log likelihood from $-21{,}024.13$ to $-20{,}676.15$ and $-20{,}516.01$, respectively. This is to be expected given the indication of the Lagrange multiplier test for spatial error dependence.[13] It is clear that least squares which ignores spatial dependence and assumes residual flows to be independent produces a much lower likelihood function value. Capturing the dependences greatly reduces the residual variance and strengthens the inferential basis affiliated with the models. Moving further to the model specification that reflects spatial dependence at both origins and destinations raises the log likelihood further to $-20{,}212.01$.

The ML estimates display the expected signs, as the ordinary least squares estimates do. The estimates reported in Table 5.2 are not significantly different from each other, and, moreover, lie within the 95% confidence limits of the least squares estimates. So, in accordance with spatial econometric theory, mere spatial dependence in disturbances does not impact the point estimates, just the precision of the parameters. Turning to the spatial autoregressive parameter, we see that the estimates are highly significant. They clearly point to origin-based, destination-based and origin-to-destination-based spatial dependence. The strength of dependence for destinations seems to be slightly more important than that for origins. But the estimate for the spatial autoregressive parameter reflecting dependence based on interaction between origin and destination neighbours is clearly most important.

Turning to the spatial autoregressive parameter ρ, we see that the estimates are highly significant. They clearly point to origin-based, destination-based and

[12] Source package: spdep 0.3-17 which may be retrieved from http://cran.r-project.org/src/contrib/Descriptions/spdep.html.

[13] Its value is at 833.6 and 1,254.8 respectively, when using ${}^{o}\boldsymbol{W}$ and ${}^{d}\boldsymbol{W}$, and fairly significant in both cases ($p = 0.000$).

Table 5.2 Spatial econometric flow models based on different spatial weights matrix specifications: ML estimates using Ng and Peyton's Cholesky algorithm ($N = 12{,}432$ observations)

	Model specifications based on		
	${}^{o}\boldsymbol{W}$	${}^{d}\boldsymbol{W}$	${}^{o}\boldsymbol{W} + {}^{d}\boldsymbol{W}$
Parameter estimates (p-values in brackets)			
Constant [α_0]	−7.041 (0.000)	−6.817 (0.000)	- 4.658 (0.000)
Origin variable [α_1]	0.598 (0.000)	0.576 (0.000)	0.593 (0.000)
Destination variable [α_2]	0.548 (0.000)	0.560 (0.000)	0.553 (0.000)
Geographical distance [β_1]	−0.194 (0.000)	−0.223 (0.000)	−0.224 (0.000)
Country border [β_2]	−0.641 (0.000)	−0.600 (0.000)	−0.651 (0.000)
Technological distance [β_3]	−2.395 (0.000)	−2.040 (0.000)	−2.183 (0.000)
Spatial autoregressive parameter [ρ]	0.311 (0.000)	0.365 (0.000)	0.613 (0.000)
Performance measures			
Log likelihood	−20,676.142	−20,516.006	−20,212.013
Sigma square	1.595	1.541	1.442
Computational time (s)	62	56	836
Diagnostics (p-values in brackets)			
Likelihood-ratio test	3,695.970 (0.000)	1,016.243 (0.000)	1,624.23 (0.000)
Moran's I	−0.008 (0.929)	−0.014 (0.990)	−0.006 (0.939)

Notes: The origin-destination models are defined by (5.25)–(5.26). $\boldsymbol{a}$ is measured in terms of the log number of patents (1985–1997) in the knowledge producing region i, $\boldsymbol{b}$ in terms of the log number of patents (1990–2002) in the knowledge absorbing region j, $\boldsymbol{d}^{(1)}$ represents geographic distance measured in terms of the great circle distance (in kilometers) between the economic centres of the regions i and j, $\boldsymbol{d}^{(2)}$ border effects in terms of the existence of country borders between i and j, and $\boldsymbol{d}^{(3)}$ technological distance in terms of the technological proximity index s_{ij}. The origin-based spatial weights matrix ${}^{o}\boldsymbol{W}$ is defined by (5.14), the destination-based spatial weights matrix ${}^{d}\boldsymbol{W}$ by (5.15), while the origin-destination-based weights matrix ${}^{o}\boldsymbol{W} + {}^{d}\boldsymbol{W}$ by (5.14)–(5.15). Model performance is measured in terms of the log likelihood, sigma square (the error variance) and computational time in seconds (running on a SunFire V 250 with 1.28 GHz and 8 GB RAM)

origin-to-destination based spatial dependence. The strength of dependence for destinations seems to be slightly more important than that for origins. But the estimate for the spatial autoregressive parameter reflecting dependence based on interaction between origin and destination neighbours is clearly most important.

5.6 Conclusions and Outlook

The chapter has illustrated the importance of proper handling spatial interaction data where spatial dependence is present. The generalised Getis–Ord statistic G_{ij} appears to be a powerful tool in exploratory spatial interaction data analysis that

yields interesting insights into the processes that give rise to spatial association between residual flows in that it enables detection of local non-stationarities.

Spatial econometric origin-destination flow models extend conventional spatial interaction models to deal with the problem of spatial autocorrelation among the residuals and to examine the role of spatial dependence in flows. These models are formally equivalent to conventional spatial regression models. But they differ in terms of the data analysed and the way in which the spatial weights matrix is defined. The chapter has suggested a computational approach for maximum likelihood estimation that relies on sparse matrix Cholesky algorithms to efficiently compute the maximum likelihood estimates and to test for origin-based, destination-based and origin-to-destination-based spatial dependence. This approach makes ML estimates practical for larger spatial interaction systems and yields reasonable computing times. An example using patent citation data that capture knowledge flows between 112 European regions has served to illustrate the discussion. The ML estimates results have shown the importance of incorporating spatial dependence in the estimation of spatial interaction relationships.

Acknowledgements The authors gratefully acknowledge the grant no. 11329 provided by the Jubiläumsfonds of the Austrian National Bank. Many thanks go to Roger Bivand (Norwegian School of Economics and Business Administration) for solving an issue with procedure *errorsarlm* and sparse matrices in the *R spdep package*.

Part II
Pattern Analysis

Chapter 6
Second-Order Analysis of Point Patterns: The Case of Chicago as a Multi-center Urban Region*

Arthur Getis

This Chapter was originally published in:
Getis, A. (1983) Second-Order Analysis of Point Patterns: The Case of Chicago as a Multi-Center Urban Region. The Professional Geographer 35:73-80. Reprinted with permission of Taylor & Francis, Philadelphia, ©The Association of American Geographers

Abstract A comprehensive approach to the analysis of point patterns demonstrates the usefulness of second-order methods by exploring population distribution in the Chicago region. The methods are based on the development of a distribution of all interpoint distances representing the total covariation in a pattern. Clustering and inhibition models are explored with regard to the population pattern. Some evidence supports a multi-center city hypothesis for the region.

In the past, most point pattern work has relied on either quadrat sampling or the analysis of nearest neighbor distances. The former technique, while lending itself to spatial modeling, has had limited usefulness because significance tests cannot verify or reject the spatial character of the supposed process. Also, the sampling technique commonly used, a lattice of regions, has many drawbacks, including the strong possibility of a violation of the required independence assumption. Furthermore, results depend on the size of the sampling units. When border effects are taken into account, nearest neighbor techniques have been useful for tests on a Poisson-process model, but they tend to be inadequate to test other models. Study is usually restricted to only the first few nearest neighbor distances.

The first attempts at a comprehensive analysis of point patterns were made by Bartlett (1963, 1964), who proposed the use of a two-dimensional spectrum of *all*

* Getis, Arthur (1983). Second-order analysis of point patterns: the case of Chicago as a multi-center urban region, *The Professional Geographer*, 35, 73–80, reprinted with permission.

A. Getis
Department of Geography, University of Illinois, Urbana-Champaign, IL, USA
e-mail: arthur.getis@sdsu.edu

the interpoint distances of a spatial point pattern. The difficulty with his method lies in the need to smooth the spectral estimate with an arbitrary smoothing function. Nonetheless, the method provides considerable promise, especially when the assumption of spatial stationarity (see below) can be sustained (Cliff and Ord, 1975). In geography, Tobler (1969) and Rayner and Golledge (1972, 1973) used spectral analysis to advantage for the examination of town spacing.

The use of a spectrum of distances allows for the development of tests at many different scales. Perhaps it was the challenge of the need for a comprehensive statistic that led Ripley to introduce second-order techniques (Ripley, 1976, 1977). These have since been modified and extended by Ripley (1979a, 1981), Diggle (1979), Besag (1977a), and others. Cliff and Ord (1981) were the first in geography to discuss second-order theory, although Glass and Tobler, as early as 1971, considered city spacing in this general context. In a related development, a useful model of clustering by Strauss (1975) is also based on interpoint distances.

6.1 Second-Order Theory

The motivation for the development of second-order theory is that a single measurement or even several measurements from a point or location to other points is not sufficient to summarize a set of point pattern data (Ripley, 1976, 1977, 1979a). The object is to find a cumulative distribution function based on all distances between pairs of objects. Since all interpoint distances taken together represent the total *covariation* in a set of points, we consider the analysis of the distribution of these distances as the study of the second moment or second-order analysis. Because attention is focused on the exploration of the arrangement of sets of points rather than on specific point locations, it follows that stochastic processes are the vehicle for analysis (Ripley, 1979a, p. 55). It must be borne in mind in the analysis we assume a known generating mechanism is responsible for the configuration of points. The method is designed for tests on hypotheses that are culled from our knowledge of point processes.

Perhaps the major limitation of the method to be discussed is that the assumption of stationarity must be met (Cliff and Ord, 1975). The assumption can be thought of as being divided into two parts: homogeneity and isotropy. The first implies that the surface on which the objects (represented by points) are contained is uniform, so that objects are not denied sites for their location. For example, a mountainous region is not uniformly able to contain towns. Second, the pattern must be isotropic; that is, there must be no directional biases in the data. Although it is possible to make adjustments so that the stationarity assumption can be met, the need to make corrections might limit the usefulness of second-order analysis.

If we can assume stationarity, then following Ripley (1977) we define a quantity $\lambda K(t)$ as the expected number of additional points within distance t of an arbitrary point. λ is the density of points per unit area and is estimated by N/A, where N is the number of points in the sample and A is the area. $K(t)$ is a non-negative increasing function. The empirical cumulative distribution function of points to all other points within distance t in a region is $F(t) = (\sum 1)/N(N - I)$,where the

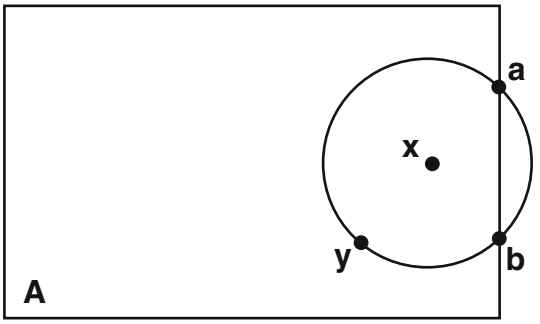


Fig. 6.1 A border correction is needed when the distance xy is greater than the distance of x to the nearest border

sum is over all ordered pairs of points not more than t apart. The difficulty with this function is that border effects and the size and shape of the region introduce bias. Ripley demonstrates that it is possible to produce an unbiased estimator by weighting pairs of objects (Ripley, 1981, p. 159). Thus,

$$\hat{K}(t) = A\frac{\sum k(x,y)}{N^2}, \tag{6.1}$$

where A is the area of the region under consideration and $\sum k(x,y)$ is the sum of the weights associated with each of the ordered pairs of points labeled x and y. In order to find the weight, consider the value $1/k(x,y)$ as the ratio of ayb to the entire circumference, as seen in Fig. 6.1.

Thus, when the border plays a strong role, that is, when the ratio $1/k(x,y)$ is low, the resulting value of $\hat{K}(t)$ will be increased. The weights make $\hat{K}(t)$ an unbiased estimator for $K(t)$ as long as t is less than the circumradius of the region studied. For a square of unit area, $\hat{K}(t)$ would be unbiased up to $t = 0.707$.

Besag (1977a) further developed Ripley's statistic by suggesting that the plot of $\hat{K}(t)$ be made linear for the Poisson process (Ripley, 1981, p. 160). The value $L(t)$ straightens $K(t)$, and Ripley (1979a, p. 63) shows that its mean is t

$$\hat{L}(t) = \sqrt{\frac{\hat{K}(t)}{\pi}} \tag{6.2}$$

and variance is $\frac{1}{2}\pi N^2$ a Poisson process, thus allowing for statistical tests of significance up to the point of biasedness.

Before analyzing the Chicago data we shall briefly discuss two second-order models. The first is based on a clustering process while the second pertains to an inhibition (or evenness) process. Both can be discussed using the distribution func-

tion of distances. Also, we will show that clustering and inhibition can be considered simultaneously.

6.1.1 Clustering Model

The second-order clustering model is much like the models discussed under the title "generalized" in Getis and Boots (1978). Such models as the negative binomial, Neyman Type A, and the Polya–Aeppli are based on a multiplicative process involving a certain distribution of offspring related to another distribution of progenitors. Diggle (1981) shows that the distribution function $K(t)$ can be interpreted as the sum of a Poisson progenitor function of intensity ρ and a random number (Poisson variate) of offspring per parent positioned in a radially symmetric (Gaussian) way around the parent with common dispersion parameter σ. The totality of offspring is

$$K(t) = \pi t^2 + \rho^{-1} \left\{ 1 - \exp \left[-t^2/(4\sigma^2) \right] \right\}. \tag{6.3}$$

The first term of (6.3) represents the progenitors and is a Poisson variable. The second term accounts for the offspring and thus represents the clustering over and above that expected in a Poisson process. Thus, $K(t) > \pi t^2$ implies clusters of points scattered around parent points.

6.1.2 Inhibition Model

An inhibition model depicts the disinclination of objects to be near one another. Following Matern's technique of identifying a radius from a point within which no other point is observed (Matern, 1971), the second-order model would be one where little or no covariation in the data is present. The $\hat{L}(t)$ function is close to or at 0 for t values representing dispersal in the data. A regularity in the inhibition process (evenness) would yield a function less than the Poisson $\hat{L}(t)$ throughout the range of unbiasedness.

6.1.3 Clustering with Inhibition

It is possible to observe clustering when inhibition is present. For example, towns may be clustered together, but they do not occupy the same space. Strauss (1975) provides us with a way of modeling this type of clustering/inhibition. From theory or observation, we identify some distance t as our indicator representing points that are close. A parameter ν measures the clustering tendency and can be estimated by finding the cumulants of Y, where Y is the number of pairs of points whose distance apart is less than t. The quantities a, b, and s can be estimated from the cumulants (defined below) of Y by

$$\begin{aligned} a &= \kappa_1 - 2\kappa_2^2/\kappa_3, \\ b &= \tfrac{1}{4}\kappa_3/\kappa_2, \\ s &= 8\kappa_2^3/\kappa_3^2, \\ \text{and } \hat{\nu} &= \frac{1}{2b} - \frac{s}{2(y-a)}. \end{aligned} \tag{6.4}$$

The observed number of pairs of points within distance t of all points is y. In a Poisson pattern with no inhibition, $\hat{\nu} = 0$ and $var(\nu) = (1-2b\hat{\nu})^2/2b^2 s$. A positive value for $\hat{\nu}$ implies clustering, while a negative value indicates inhibition.

In the two-dimensional case where we define a region as having unit area, we have $D = \pi t^2$,

$$\begin{aligned} \kappa_1 &= E(Y) + \tbinom{N}{2} D, \\ \kappa_2 &= var(Y) = \kappa_1(1-D), \\ \kappa_3 &= \kappa_2(1-2D) + N^{(3)}(0.5865 - D)D^2, \\ \kappa_4 &= \kappa_2(1-6D(1-D)) + N^{(3)}D^2(D(0.4596 - D) \\ &\quad + 6(0.5865 - D)(1-2D)), \end{aligned}$$

where $N^{(s)} = (N)(N-1)\cdots(N-s+1)$. Technical problems associated with cumulants are discussed in Strauss (1975).

The choice of t depends on our understanding of the processes bringing about clustering. As I will show in the example below, there are several values of t that are directly related to assumptions about the population distribution in urban areas. Strauss provides a test for any t, although the optimal test obtains when the value of t corresponds to the maximum variance (κ_2).

6.2 The Population Pattern of Chicago

In order to demonstrate constructively the use of second-order methods, I have attempted to develop further the notion of the multi-center city (see Griffith, 1981 and Plane, 1981 for recent analytical development of this idea). For many years, urban population density decline analysis has been used to test empirically single center urban land use theory. The multi-center city hypothesis requires, however, some other forms of analytical routines. I suggest that second-order methods may offer a useful approach for testing some spatial aspects of cities that are assumed to have more than one center of high population density. I am using data on the residential population distribution in the Chicago metropolitan region.

Suppose that each of the major centers of the multi-center city is a workplace focus and that peoples' residences cluster around those centers according to some journey-to-work distance principle (see, e.g., Getis, 1969). What results in most cases is a city comprised of overlapping population clusters. These clusters cannot

be identified exactly, because of the overlap, but they can be modeled to some extent by second-order methods.

Before the analysis can be carried out, three problems must be addressed. The first concerns the level of generalization, that is, the number of points used to represent the population. It is obvious that analysis of a population distribution map for a large urban area must utilize something less than one point per person. Also the greater the ratio of points to people, the more accurate is the representation of the population. Both the realism of the map desired and the value of computer time and labor must be weighed in selecting a level of generalization. For every ten-fold increase in the number of points used (N), there is a 100-fold increase in computer time needed to find $\hat{L}(t)$.

Chicago census tracts are used as the fundamental data base. These are small enough for a careful but uncrowded and realistic placement of points to represent the population distribution. I decided on one point for each 10,000 people, using data from the 1970 census. Any reasonably accurate representation of the point pattern will yield similar results for a clustering hypothesis, though not for an inhibition hypothesis. In the case of inhibition, any result will vary directly with the level of generalization simply because the more points there are, the closer they will be to one another. Thus, hypotheses dealing with inhibition must be carefully constructed.

The second problem concerns the assumption of stationarity. No city occupies a completely homogeneous site, therefore no point set representing population satis-

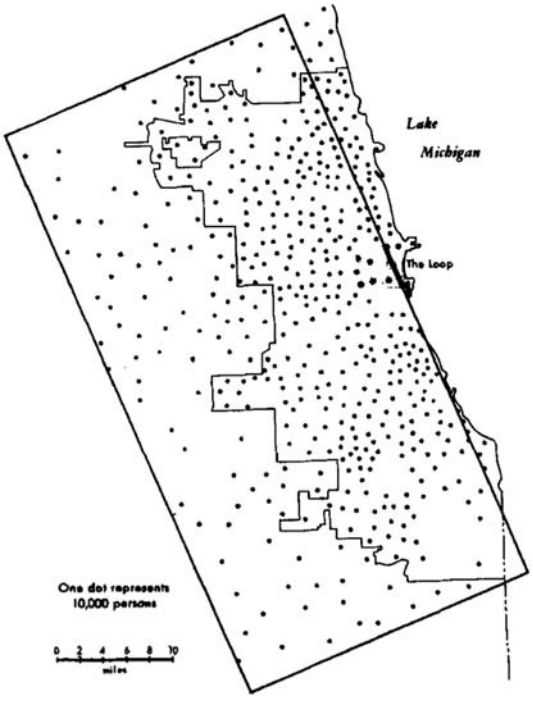


Fig. 6.2 Population distribution in the Chicago region

fies the assumption. Steps such as map transformations may be taken to correct for heterogeneity. In the Chicago region, physical features affecting population distribution are generally absent except for Lake Michigan. The lake affects population distribution along the shoreline, especially to the north of the CBD (the Loop). The strip of clustering along the north shore was eliminated by constructing a rectangular study area that has as its eastern edge Lake Michigan south of the Loop and an arbitrary line about 1 mile inland north of the Loop (see Fig. 6.2). Eliminating that strip reduces the likelihood of a directional bias in the data but retains isotropic properties.

The third problem concerns the bias that might result from failure to consider the region's boundaries, a problem that can be handled in at least two ways. The first is to employ the boundary condition discussed above, that is, use the factor $1/k(x, y)$ in the formula for $K(t)$. One might also consider mapping the study area onto a torus and thus eliminating boundaries altogether. In practice, this procedure means that the original study area is reproduced eight times; the reproductions are placed along the borders of the study area so as to completely surround it. Adopting this strategy solves both the boundary problem and the problem of the absence of settlement in Lake Michigan. The reflection along the eastern edge expands the study area into the shape of a square and effectively allows us to define $A = 1$ and t values to correspond to distances within the square of unit area. One mile is equivalent to $0.032t$. In the area under study, there are 422 points representing more than four million people.

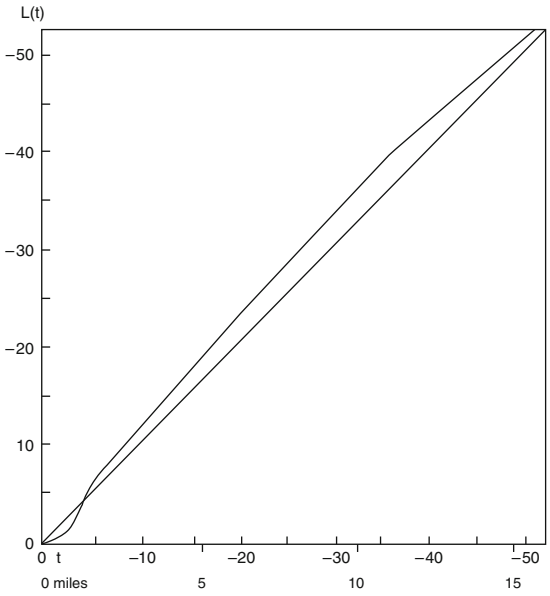


Fig. 6.3 A plot of $\hat{L}$ for the population of the Chicago region. The *straight line* is the mean for a Poisson process. The portion of the diagram from $t = 0$ to $t = 0.06$ is enlarged on Fig. 6.4

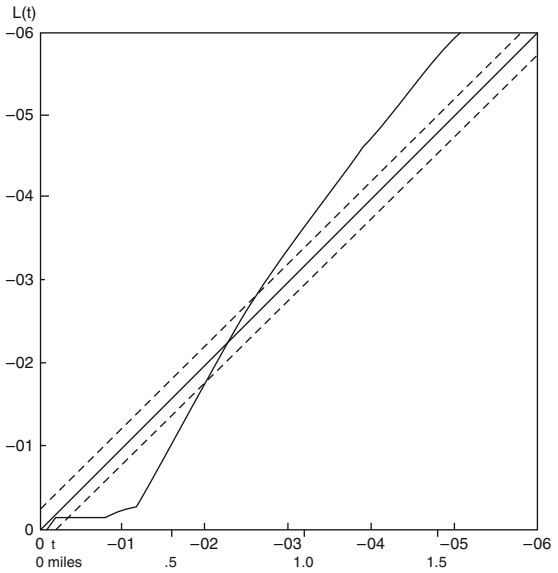


Fig. 6.4 A plot of $\hat{L}$ for the population of the Chicago region. The *dashed lines* are the 95% confidence bands of the Poisson process

$\hat{L}(t)$ for the 1970 map of Chicago's population is shown in Fig. 6.3. Recall that $\hat{L}(t)$ is the linearized cumulative distribution function of pairs of points; the Poisson process model is represented by the straight line. Several noteworthy features appear. First, there is evidence of an inhibition effect within 1 mile; second, there is a smooth curve above the diagonal representing clustering. The clustering appears to increase with t to about 7–9 miles, at which point there is a movement toward the Poisson expectation. At longer distances (greater than 15 miles) the pattern becomes much less clustered and the curve has very few irregularities. The pairs of points increase in distance from each other at a fairly constant rate, implying a certain spatial regularity to the clustering.

An enlarged view of the first 4 miles of interpoint distances is shown in Fig. 6.4. Also shown is the 95% confidence band around the Poisson process expectation. One can identify the inhibition effect to a distance of 0.3 miles; by 0.9 miles the clustering tendency is pronounced. Here, too, one observes a very smooth curve.

The Chicago Area Transportation Study (1980) reports that the mean journey-to-work distance for the Chicago SMSA for 1970 was close to 7.00 miles. Applying the Strauss model for the comparable t value of 0.225, we get the following:

$$
\begin{aligned}
N &= 844,\\
D &= 0.159,\\
y &= 73{,}194 \text{ (number of pairs of points within } t),\\
\kappa_1 &= 56{,}579,\\
\kappa_2 &= 47{,}580,\\
\kappa_3 &= 6{,}509{,}939,\\
\kappa_4 &= 27{,}236{,}366,\\
\hat{\nu} &= +0.014.
\end{aligned}
$$

Note that $\hat{\nu}$ is positive, indicating clustering within the distance t. Using a normal test $(y - \kappa_1)/\sqrt{\kappa_2}$, we see that $\hat{\nu}$ is 76 positive standard deviational units away from the Poisson expectation. A check of the data reveals that the maximum variation in $\hat{L}(t)$ occurs at $t = 0.230$ (or 7.11 miles). The 7-mile journey-to-work distance appears to encompass the greatest degree of clustering in the Chicago area. This value does not necessarily link the mean journey-to-work distance to the spacing of urban centers, but it does suggest that such a hypothesis is plausible. Further empirical evidence suggests that in the Chicago metropolitan area relatively short journey-to-work distances (less than 10 miles) focus on a few widely-spaced, important work centers (Continental Illinois National Bank, 1978; Getis, 1985a).

The configuration shown on Fig. 6.3 results from more than one cluster of points. A single cluster of points with a lower density of dispersed points would tend to yield a curve that would dip below the Poisson expectation for medium and long distances. More than one cluster forces the curve above the diagonal for most of its length. Nearest neighbor distances would not capture this important difference. The curve's peakedness depends on the proportion of points that are very close to one another. In this example, a rather modest peak in the range 7–9 miles implies that the clustering is not intense. During earlier periods in Chicago's history the peakedness was probably more pronounced. Thus, I suggest that a new type of population density analysis might be available by means of second-order methods.

A second model based on an inhibition distance of 0.5 miles ($t = 0.015$) yields

$$
\begin{aligned}
N &= 844,\\
D &= 0.0007069,\\
y &= 77,\\
\kappa_1 &= 251.46,\\
\kappa_2 &= 251.28,\\
\kappa_3 &= 426.27,\\
\kappa_4 &= 250.56,\\
\hat{\nu} &= -1.689.
\end{aligned}
$$

In this case $\hat{\nu}$ is negative, indicative of inhibition within the distance t. Since $\hat{\nu}$ is 11 negative standard deviational units from the Poisson expectation, I conclude that

the inhibition is pronounced, as expected. As was mentioned previously, this is a function of the scale of analysis and is not a true population effect.

6.3 Conclusions

The advantages of second-order analysis seem to lie in the help it gives for the development of plausible models of the distribution of geographic phenomena. This is in addition to the possible analytical value of having a complete description-of interpoint distances. Both of these advantages should not be undervalued since $\hat{L}(t)$ provides more opportunities for insight than do a few summary numbers.

In the Chicago population example, I have shown the nature of population clustering and tentatively conclude that a 7–9 mile distance includes maximum clustering. This approach and finding suggest there are numerous other urban characteristics worth exploring, including inferences about population density, population patterns over time, and population pattern differences among cities of different sizes and cultures.

Acknowledgements I would like to thank Marc Armstrong for preparing the computer program used in this research and for his critical comments.

Chapter 7
Second-Order Neighborhood Analysis of Mapped Point Patterns

Arthur Getis and Janet Franklin

This Chapter was originally published in:
Getis, A., Franklin, J. (1987) Second-Order Neighborhood Analysis of Mapped Point Patterns. Ecology 68:473-477. Reprinted with permission of The Ecological Society of America, Washington DC

Abstract A technique based on second-order methods, called second-order neighborhood analysis, is used to quantify clustering at various spatial scales. The theoretical model represents the degree of clustering in a Poisson process from the perspective of each individual point. The method is applied to point location data for a sample of ponderosa pine (*Pinus ponderosa*) trees, and shows that heterogeneity within the forest is clearly a function of the scale of analysis.

7.1 Introduction

In any study where spatial data or pattern analyses are required, the appropriate scale for analysis must be chosen. The choice is often arbitrary. Scale is usually defined as the ratio of map distance to the real world distance it represents (Robinson et al., 1984). As scale changes, so does the level of resolution, and new spatial patterns emerge. Theory or subject matter should guide the selection of an appropriate scale, but often researchers need to look at pattern at a number of scales. Spatial pattern has both intensity, the extent to which density varies in space, and grain, the distance over which density is perceived to vary (Pielou, 1977, 155–156).

Workers in a number of disciplines have attempted to find methods for identifying parameter changes that take place when scale is made to vary. Perhaps *blocking*, or contiguous quadrat analysis, is the most common method used for examining

A. Getis (✉)
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

J. Franklin
School of Geographical Sciences and Urban Planning, and School of Life Sciences, Arizona State University, USA
e-mail: Janet.Franklin@asu.edu

grain of pattern. The study area is covered by an array of N quadrats. These quadrats are combined into larger quadrats in a systematic way. Various guidelines have been proposed for selecting the optimum quadrat size, the most common of which is an analysis of variance (Moellering and Tobler, 1972; Grieg-Smith, 1983).

Spectral analysis, as a method for selecting scales, has been satisfactorily applied to studies where blocking or hierarchically defined units are used. In addition, data transects have been studied as continuous spectra for scale effects (for a review of the literature see Ripley, 1981). Rayner (1971), following the lead of Bartlett (1950), demonstrates how pattern may be studied using two-dimensional spectra. Although more complicated than the technique discussed here, an important advantage is that it allows for an assessment of the effect of orientation on pattern and scale.

While one of the authors was engaged in a remote sensing study of canopy reflectance for a ponderosa pine (*Pinus ponderosa*) forest (Franklin et al., 1985), we developed a technique for describing both intensity and grain of tree spatial patterns simultaneously at a number of scales. The method, second-order neighborhood analysis, is a variation on second-order analysis of point patterns (Ripley, 1977, 1981; Diggle, 1983; Getis, 1984). Second-order analysis is designed to test randomness hypotheses, often based on the Poisson distribution, by examining the proportion of total possible pairs of points in Euclidean space whose pair members are within a specified distance of each other. The analysis is second order because it is the variation rather than the mean of distances that is being studied. The technique discussed below, while similar to second-order analysis, differs in that consideration is given only to those pairs of points having as one of its members a given point i. This method depends on relatively large amounts of digitized point data, from aerial photographs or maps, where coordinates can be accurately recorded.

7.2 The Model

Getis' model (1984) has the form

$$\hat{L}_i(d) = \left[A \sum_{j=1}^{n} k_{ij} / \pi(n-1) \right]^{\frac{1}{2}}, \tag{7.1}$$

where Σk_{ij} is the summation over all points that are within distance d of point i, and it includes a boundary correction where required. If for a given neighborhood point j the specified distance d is more than the distance between i and j, then the pair (k_{ij}) counts as 1 (unless the boundary correction is required); otherwise k_{ij} counts 0. The value A is the area of a rectangular region, and $n-1$ represents all possible pairs of points having i as a pair member. Taking into consideration the circular area centered on point i, for convenience, π and the square root are included in order to make $L_i(d)$ linear with respect to d and to have $L_i(d) = d$ when $L_i(d)$ represents a pattern produced by a Poisson process in the plane.

The boundary correction is as follows: if the distance between i and j is greater than the distance between i and the nearest boundary (e_1), instead of the value 1 for k_{ij}, substitute

$$k_{ij} = [1 - \cos^{-1}(e_1/d)/\pi]^{-1}. \tag{7.2}$$

If the distance between i and j is greater than the distance to both of two boundaries (e_1, e_2), use

$$k_{ij} = \{1 - [\cos^{-1}(e_1/d) + \cos^{-1}(e_2/d) + \pi/2]/2\pi\}^{-1}. \tag{7.3}$$

The boundary correction is based on the assumption that the region outside of the boundary in the vicinity of the distance measurement has a spatial pattern similar to the nearby areas within the boundary. If this assumption cannot be accepted, then results must be exclusively for the areas within A greater than d from all boundaries (see Getis, 1984 for further discussion of the boundary problem).

The form of the analysis can best be depicted by a diagram. Figure 7.1 shows a curve describing the typical values of $\hat{L}_i(d)$ for a given i in a somewhat clustered forest in a square of area 1. The horizontal axis represents d; that is, at any distance from a tree designated as i we can identify an $\hat{L}_i(d)$ value. The diagonal represents $L_i(d)$ values for a pattern that is created by a Poisson process. The initial part of the curve for $\hat{L}_i(d)$ displays a value of 0 as far as distance 0.08. This means that no other tree is within 0.08 of tree i, and so this is the nearest neighbor distance. Up to distance 0.14 from i, the curve remains below the expectation. The

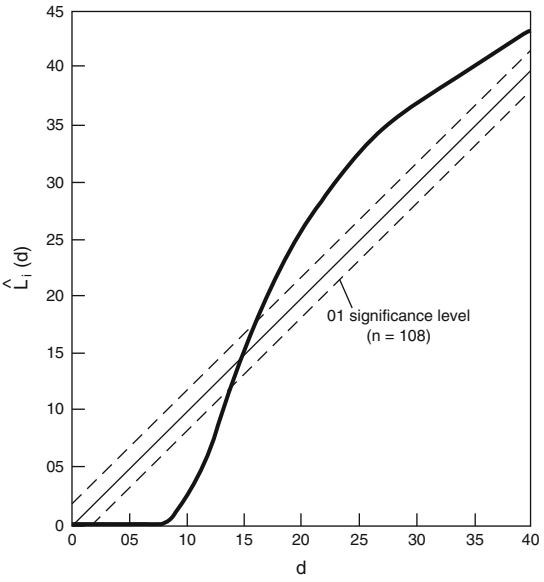


Fig. 7.1 Cumulative distribution curve (*heavy line*) of $\hat{L}_i(d)$ for hypothetical tree in a square of area 1. $L_i(d)$ is the number of points within distance d of point i corrected for the boundary effect, and scaled such that $L_i(d) = d$ when $L_i(d)$ represents a pattern produced by a Poisson process in the plane. *Dashed lines* represent 0.01 significance levels around the line representing Poisson process

fact that the curve for the observations is advancing upward at a faster rate than the theoretical curve implies a tendency for clustering or heterogeneity. It is not until 0.16 from i that one can say that the spatial distribution of pairs displays a statistically significant level (0.01) of clustering (see below). At 0.28, the curve reaches its maximum above $L_i(d)$, implying maximum clustering. In summary, the parameters representing i's relationship with all j are (1) the nearest neighbor distance, (2) the distance at which heterogeneity begins, (3) the distance at which clustering becomes statistically significant, and (4) the distance at which maximum clustering can be observed.

Statistical significance can be ascertained either by simulation or by accepting the values $\pm 1.42\sqrt{A}/(n-1)$ and $\pm 1.68\sqrt{A}/(n-1)$ as reasonable approximations of the 5% and 1% significance points, respectively. These are a modification of Ripley's (1978; 1979b) estimates for the second-order case.

In addition to the above indicators of the relationship of i to all j are the scale parameters. If we identify the $\hat{L}_i(d)$ value at certain specified distances, say 0.05, 0.10, 0.15, 0.20, for each i, we are then able to compare the spatial *situation* of each tree. One tree may display a high $\hat{L}_i(d)$ value at 0.05, implying that a number of neighbors are close by, while a second tree may have a low $\hat{L}_i(d)$ value at 0.05 but a high value at 0.20. The second tree is much less crowded by near neighbors, but is within a cluster of trees at a distance of 0.20 from it. If the chosen scale of analysis were 0.05, the first tree would be considered a member of a cluster, but the second tree would not. The distance chosen represents the scale at which one can view pattern.

To demonstrate the method, ponderosa pine tree distribution was analyzed. The locations of ≈5,000 ponderosa pine trees in the Klamath National Forest in Northern California were determined from United States Forest Service aerial photographs (nominal scale 1:24,000); trees <2.5 m apart were not resolvable, nor were small trees within the canopy of another tree. These points were digitized for automatic analysis (Franklin et al., 1985). A subarea selected for study, 120 m × 120 m, included 108 trees that visually display nonrandom characteristics: clumps and clusters of trees appear to dominate the pattern (Fig. 7.2).

7.3 Results

Figure 7.3 shows the observed and expected $L(d)$ values; $\hat{L}(d)$ represents the average distance relationships for the 108 trees in the subarea shown in Fig. 7.2. For convenience the study area was made equal to 1; thus a distance of 0.01 is equivalent to 1.2 m. All data points on Fig. 7.3 are within the 95% confidence region of the Poisson expectation. This implies that although there are clusters of points and an apparent inhibition effect, the overall pattern cannot be differentiated from one created by a Poisson spatial process.

Figure 7.4 contrasts the pattern membership characteristics of three selected trees, labeled A, B, and C in Fig. 7.2. Note that tree A appears to be a member of a small cluster of trees. Figure 7.4a shows a short nearest neighbor distance

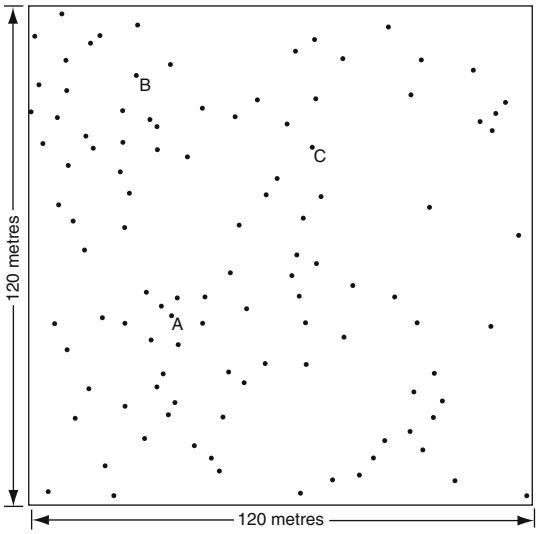


Fig. 7.2 Point pattern representation of tree locations in the study area. The letters A, B, and C mark particular individual trees, which are referred to in Fig. 7.4. North is up

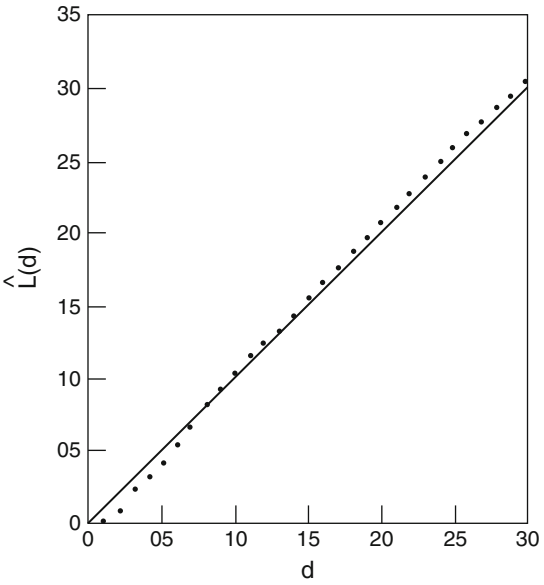


Fig. 7.3 Values for $\hat{L}(d)$ over the range $0.01 \leq d \leq 0.30$. $L(d)$ is the number of points within distance d of all points i corrected for the boundary effect, and scaled such that $L(d) = d$ when $L(d)$ represents a pattern produced by a Poisson process in the plane. $\hat{L}(d)$ may be interpreted as the average for all 108 points (from Fig. 7.2) taken together. *Solid line* shows expected values given a Poisson distribution. *Solid dots* show observed values

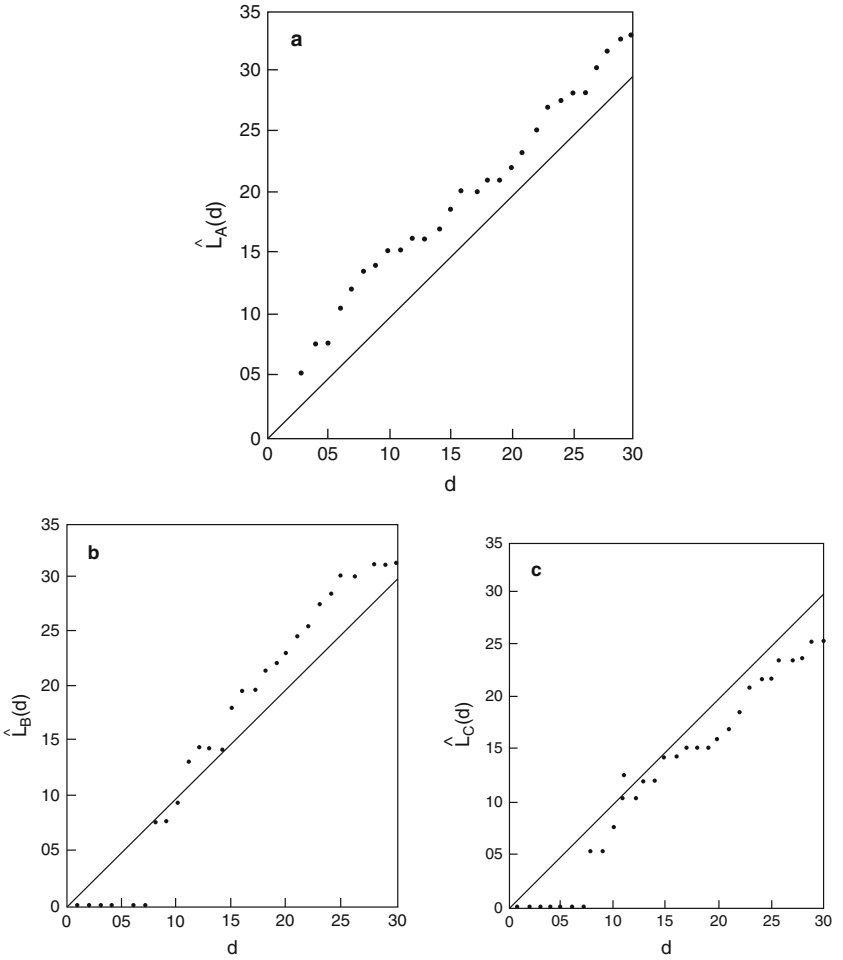


Fig. 7.4 Values for $\hat{L}_i(d)$ over the range $0.01 \leq d \leq 0.30$ when $i = A$ (a), B (b), and C (c). The locations of trees A, B, and C are shown in Fig. 7.2

(0.03 = 3.6 m), a rapid rise to clustering status at a distance of 0.03 (3.6 m), and maximum clustering at 0.09 (10.8 m). Visually, point B does not appear to be a member of a cluster, but inspection of Fig. 7.4b reveals that B is a member of a cluster at distances of 0.11 (13.2 m) and greater. The distance at which maximum clustering takes place (0.25), however, is much greater than for point A. Point C is within an area of the forest where densities are much lower than is true of either A or B. Its $\hat{L}_i(d)$ values, shown in Fig. 7.4c, reveal that it is not a member of a cluster at most scales.

Figure 7.5 shows the pattern created by the trees in our sample for scales (d) of 0.05 (6 m), 0.10 (12 m), 0.15 (18 m), and 0.20 (24 m). Of course, a much finer group

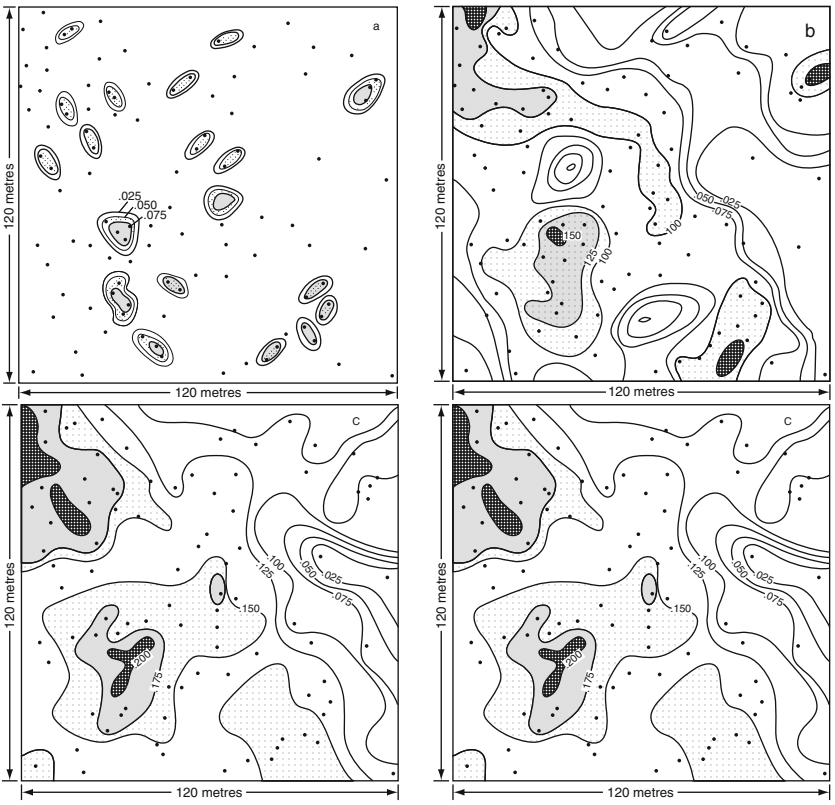


Fig. 7.5 Pattern created by assigning to each tree its $\hat{L}(d)$ value for the following values of d: (a) 0.05 (6 m), (b) 0.10 (12 m), (c) 0.15 (18 m), (d) 0.20 (24 m). The *shaded areas* contain trees that have $\hat{L}_i(d)$ values above the Poisson expectation. Intensity of shading corresponds to increase in tree density at a given scale. Isolines are in units of d

of scales could have been selected. The isolines, drawn at intervals of 0.025, indicate areas of greater or lesser tree densities. Accuracy in drawing the isolines was enhanced by the addition of control points to the empty or sparsely vegetated areas; no measurements were taken to a control point. The shaded areas contain all trees that display $\hat{L}_i(d)$ values above the expected, that is above 0.05, 0.10, 0.15, and 0.20, respectively. Intensity of shading corresponds to increases in tree density at the given scale.

Comparisons can be made among all areas of the maps or only among the areas unaffected by the border correction. For the entire area of each map in Fig. 7.5 and by casual inspection of Fig. 7.2, it is clear that the higher densities are generally in the west and the lower densities in the east. Figure 7.5 reveals, however, a number of further interesting contrasts. For example, note that at a scale of 0.05 (Fig. 7.5a), only trees within 6 m of one another are considered as members of clusters, so that large areas of Fig. 7.5a have a relatively low density of trees. When the scale is

increased to 0.10 (Fig. 7.5b), some of the clusters identified at the 0.05 level are now considered part of larger clusters or are not part of any cluster at all. By contrasting the 0.05 (6 m) and the 0.10 (12 m) levels, one can see that geographic interpretations would be greatly different due to the scale chosen. In addition, note that the relatively low-density area in the western half of Fig. 7.5a becomes part of a clustered region when the scale is increased to 0.20 (Fig. 7.5d).

The variance about the observed mean, $\bar{L}_i$, for a particular d indicates the extent of the heterogeneity within the pattern. The scale at which the variance is maximized will show the greatest contrast in pattern. This may be a reasonable choice for an investigation when no other information is available to indicate an appropriate scale. In our example, the variance reaches its first peak at 0.15 (18 m), decreases, and then increases to a maximum at 0.35 (42 m) before decreasing again. The border correction contributes greatly to the creation of the second peak.

7.4 Conclusions and Discussion

We have shown that second-order neighborhood analysis can identify different dominant patterns at different scales for mapped point data. In the example given, the overall pattern cannot be differentiated from one created by a spatial Poisson process, but close inspection of the spatial relationships of individual trees to nearby trees reveals noteworthy variations. The influence of nearest neighboring trees dominates the pattern at or below a scale of 6 m. At 12 m, clustering is seen, but it is stronger at 18 m. From 24 to 42 m the effect of the border correction appears to play a role in intensifying the clustering (for example, near the northwest border).

Second-order analysis identifies several important scales of pattern: (1) the distance to nearest neighbor, (2) the distance where heterogeneity begins, (3) the distance where clustering becomes significant, and (4) the distance where maximum clustering is observed. The technique presented here, neighborhood analysis, can be applied to selected individuals, and maps of pattern density at a given scale can be produced. A knowledge of the scale-dependent spatial setting of individuals would be useful in testing neighborhood models of population dynamics and competition (Weiner, 1984; Pacala and Silander, 1985).

This study of scale pertains specifically to points each valued ostensibly as 1. It requires only a slight modification in our model, however, to place interval scale values at each point, such as size of a tree (see Getis, 1984). In addition, if data were given for units having areal extent (nonpoint data), such as a lattice of quadrats, the analysis could be carried out if the researcher assigns the data values to points representing each sample area.

Acknowledgements This study was partly supported by a grant from the National Science Foundation (SES-8219170). The authors appreciate the comments of W.R. Tobler and the referees.

Chapter 8
A Class of Local and Global K Functions and Their Exact Statistical Methods

Atsu Okabe, Barry Boots and Toshiaki Satoh

Abstract In 1987 Getis and Franklin introduced a technique, based on second-order methods, for quantifying clustering at various scales in mapped point patterns. Subsequently, this technique has become known as local K function analysis. In this paper we develop the local and global forms of a class of K functions and cross K functions formulated on a bounded plane that includes the technique of Getis and Franklin. Exact statistical methods are formulated or discussed and computational methods are shown for the functions.

8.1 Introduction

One of the most frequently used techniques in statistical point pattern analysis is the K function method. This method, which was originally proposed by Ripley (1976, 1977, 1979b, 1981), has been extended and applied by many researchers in various fields (e.g., animal ecology as in Gaines et al., 2000; cell biology as in Prior et al., 2003; cosmology as in Stein et al., 2000; landscape ecology as in Spooner et al., 2004a; network analysis as in Okabe and Yamada, 2001; transportation as in Yamada and Thill, 2004). A general review is given by Dixon (2002). Among those, the study by Getis and Franklin (1987) is notable because it was the first to extend the original method to a K function method that focuses on the location of points with respect to a specific point. Although they did not put a special name on their method, it can be called a *local K function method*. In celebration of one of Getis' pioneering works, this chapter discusses a class of *local* and *global K* function methods.

The chapter consists of five sections including this introductory section. For ease of explanation, first, Sect. 8.2 discusses the *local* and *global cross K* functions, and

A. Okabe (✉) and T. Satoh
University of Tokyo, Tokyo, Japan

B. Boots
Wilfrid Laurier, Waterloo, ON, Canada

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*,
Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_8,

shows their exact statistical methods. Because the local space assumed in the *local cross K* function is not always natural, and moreover, the exact statistical test of the *global cross K* function requires heavy computational time, Sect. 8.3 proposes alternative *local* and *global Voronoi cross K* functions, and formulates their exact statistical methods. Section 8.4 deals with the *local* and *global auto K* functions (the original K functions), and discusses the difficulty of deriving their exact statistical methods. The chapter ends in Sect. 8.5, summarizing the results.

8.2 Local and Global Cross K Functions

Consider two sets of points $P = \{p_1, \ldots, p_n\}$ (the white circles in Fig. 8.1) and $Q = \{q_1, \ldots, q_m\}$ (the black circles in Fig. 8.1) placed in a bounded space S, the *global space* (e.g., the square in Fig. 8.1, but generally a polygon). The points P are assumed to be stochastically distributed in S, but the points Q are fixed. A typical example is that the points P represent crime spots and the points Q represent railway stations. Note that the configuration of the points Q is arbitrary; it is not necessary that the points are uniformly placed in S.

Let $D_i(t)$ be the disk centered at a point, q_i, of Q with radius t, and t_i^* be the distance between q_i and the farthest point in S (Fig. 8.1), i.e., the minimum value of t that satisfies $S \subseteq D_i(t_i^*)$. Let $K_i(t)$ be a function given by

$$K_i(t) = \text{the number of the points of } P \text{ in } D_i(t) \cap S. \tag{8.1}$$

Because $D_i(t)$ includes a local space of the global space S (i.e., $D_i(t) \subseteq S$), this function is called a *local cross K function*.

Two remarks are made on this function. First, the function $K_i(t)$ could be standardized by multiplying a constant (i.e., the density of points), but in this paper, this is not done in order to make the following mathematical derivations look simpler. Second, the space S is assumed to be bounded, and so the edge effects should be

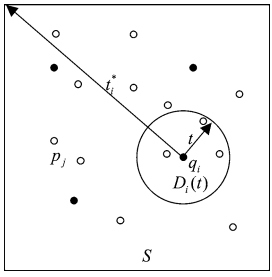


Fig. 8.1 Points of P (the *white circles*), points of Q (the *black circles*) and a disk $D_i(t)$ centered at a point of Q with radius t

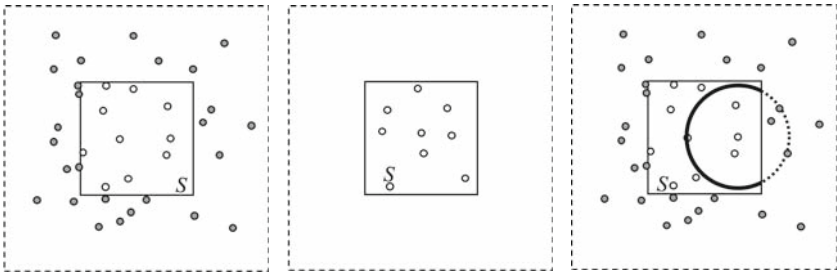


Fig. 8.2 The edge effect: (a) the first case, (b) the second case, (c) adjustment

treated. This treatment varies according to the following two cases. In the first case, other events (of the same type as those in S) can be present outside of S (Fig. 8.2a). In the second case, no other events of the type found in S are found outside of S (Fig. 8.2b). Both situations lead to edge effects but our treatment of them is different in the two cases. Most literature of the K functions assumes the first case where the space including S is unbounded (as is assumed in the Poisson point processes). A common treatment is to correct the edge effect in terms of a constant $\alpha(t)$ for t that adjusts a bounded space to an imaginary unbounded space (Getis and Franklin, 1987). For example, as in Fig. 8.2c, $\alpha(t)$ is given by the ratio of the length of the circumference of the circle included in S to the circumference of the full circle. In this paper we assume the second case (as is assumed in the binomial point processes), and exactly take the edge effect into account.

In terms of the *local cross K* functions, $K_i(t)$, $i = 1, \ldots, m$, a function, $K(t)$, is written as

$$K(t) = \sum_{i=1}^{m} K_i(t). \tag{8.2}$$

This function, which is often referred to in the literature, is called the *cross K function*. In contrast to the *local cross K* function defined by (8.1), this function is referred to as a *global cross K function*. Note again that the constants $\alpha(t)$ and $1/m$ are neglected in the term on the right-hand side of (8.2) to make the following derivations look simpler.

Having defined the *local* and *global K* functions, the remainder of this section develops exact statistical methods. Among many possible null hypotheses, the most fundamental one is that the n points of P are independently distributed in the bounded space S according to the uniform density function, $f(\mathbf{x})$, on S, i.e., $f(\mathbf{x}) = 1/|S| = 1$, $\mathbf{x} \in S$, where $|S|$ denotes the area of S, and for simplicity $|S| = 1$ is assumed without loss of generality. This hypothesis, which will be referred to as H_o, implies that the points P are independent of the configuration of

the points Q, e.g., the distribution of crime spots is not affected by the location of railway stations.

8.2.1 The Expected Value and Variance of the Local Cross K Function

For a fixed t, under the null hypothesis H_o, the value of $K_i(t)$ is a random variable following the binomial distribution with parameters n and $|D_i(t) \cap S|$. From this property, it is straightforward to obtain the expected value and variance of $K_i(t)$ as

$$E(K_i(t)) = n|D_i(t) \cap S|, \tag{8.3}$$

$$Var(K_i(t)) = n|D_i(t) \cap S|(1 - |D_i(t) \cap S|). \tag{8.4}$$

The value of $|D_i(t) \cap S|$ is explicitly written as an algebraic function of the coordinates of q_i and vertices of S for a continuous t (Okabe et al., 2000, 515–516). Therefore the computation of the values of $E(K_i(t))$ and $Var(K_i(t))$ is readily done with constant computation time, i.e., the order of computation time is $O(1)$.

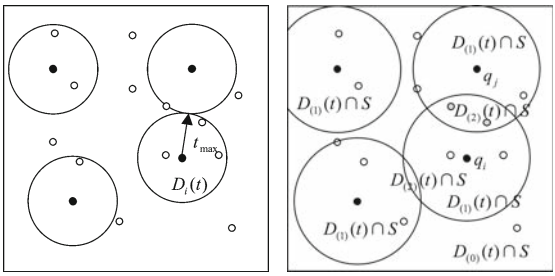
The exact test can be achieved using the binomial distribution with parameters n and $|D_i(t) \cap S|$. For a large n, the distribution can be approximated by the normal distribution with $E(K_i(t))$ given by (8.3) and $Var(K_i(t))$ given by (8.4). Therefore the exact, as well as approximate, statistical test for the null hypothesis H_o is straightforward. Note that this test exactly considers the edge effect of the second case mentioned above.

8.2.2 The Expected Value and Variance of the Global Cross K Function

One might consider that the expected value and variance of the *global cross K* function would be easily obtained from the following formulae

$$E(K(t)) = E\left(\sum_{i=1}^{m} K_i(t)\right) = \sum_{i=1}^{m} E(K_i(t)), \tag{8.5}$$

$$Var(K(t)) = Var\left(\sum_{i=1}^{m} K_i(t)\right) = \sum_{i=1}^{m} Var(K_i(t)) + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} \times Cov(K_i(t)K_j(t)), \tag{8.6}$$



(a) (b)

Fig. 8.3 Disks are not overlapped, $t \leq t_{max}$ (a), and they are overlapped, $t > t_{max}$ (b)

where each $K_i(t)$ follows the binomial distribution with parameters n and $|D_i(t) \cap S|$. However, as will be proved later, this is not true. The derivation is more complex than what it looks. To be explicit, let t_{max} be the maximum value of t that satisfies $|D_i(t) \cap D_j(t)| = 0$ for $i \neq j$, $i, j = 1, \ldots, m$, implying that disks $D_i(t)$, $i = 1, \ldots, m$ do not overlap each other except at their boundaries (Fig. 8.3a). The derivation of the expected value and variance of $K(t)$ differs according to $t \leq t_{max}$ or $t > t_{max}$.

When $t \leq t_{max}$ holds (Fig. 8.3a), $K(t)$ indicates the random number of points of P that are included in the mutually exclusive m areas $D_i(t) \cap S$, $i = 1, \ldots, m$ where the n points of P are uniformly and randomly distributed. Hence the random variable $K(t)$ follows the binomial distribution with parameters n and $\sum_{i=1}^{m} |D_i(t) \cap S|$. The expected value and variance of $K(t)$ are given by

$$E(K(t)) = n \sum_{i}^{m} |D_i(t) \cap S|, \tag{8.7}$$

$$Var(K(t)) = n \sum_{i}^{m} |D_i(t) \cap S| \left(1 - \sum_{i}^{m} |D_i(t) \cap S|\right). \tag{8.8}$$

Note that $Var(K(t))$ of (8.6) leads to a false value, $n \sum_{i}^{m} |D_i(t) \cap S|(1 - |D_i(t) \cap S|)$, even for $t \leq t_{max}$. The computational method for calculating these values is almost the same as that for the *local cross K* function mentioned above, but the order of computational time is linear to m, i.e., $O(m)$.

When $t > t_{max}$ holds (Fig. 8.3b), the calculation of the expected value and the variance becomes complex, because disks overlap and some points of P are counted twice, three times, and so forth. For instance, in Fig. 8.3b, the gray colored points in the overlapping area of the disks $D_i(t)$ and $D_j(t)$ are counted not only in $K_i(t)$ but also in $K_j(t)$; as a result, the same points are counted twice in $K(t)$. To treat this multiple count, let $D_{(k)}(t)$ be the area in which exactly k disks out of the m disks $D_i(t)$, $i = 1, \ldots, m$ overlap. Note that $D_{(0)}(t)$ indicates the area not covered with the m disks. Let $K_{(k)}(t)$ be the random number of points of P that are included in

the area $D_{(k)}(t) \cap S$ under the null hypothesis H_o. Because the points in $D_{(k)}(t)$ are counted k times, the value of $K(t)$ is given by

$$K(t) = \sum_{k=0}^{m} kK_{(k)}(t). \tag{8.9}$$

Under the null hypothesis H_o, the value of $K(t)$ is also a random variable, which follows the univariate multinomial distribution with parameters n, m, $|D_{(k)}(t) \cap S|$, $k = 1, \ldots, m$ (Johnson et al., 1992, 460–461). The probability that the *global K* function takes a specific value $K(t)$ is given by

$$\sum \frac{n!|D_{(0)}(t) \cap S|^{K_{(0)}(t)}|D_{(1)}(t) \cap S|^{K_{(1)}(t)} \cdots |D_{(m)}(t) \cap S|^{K_{(m)}(t)}}{K_{(0)}(t)!K_{(1)}(t)! \cdots K_{(m)}(t)!}, \tag{8.10}$$

where the summation is over all possible nonnegative integers $K_{(k)}(t)$, $k = 1, \ldots, m$ such that $\sum_{k=1}^{m} K_{(k)}(t) = n$. The expected value and variance of $K(t)$ are given by

$$E(K(t)) = n \sum_{k=1}^{m} k|D_{(k)(t)} \cap S|, \tag{8.11}$$

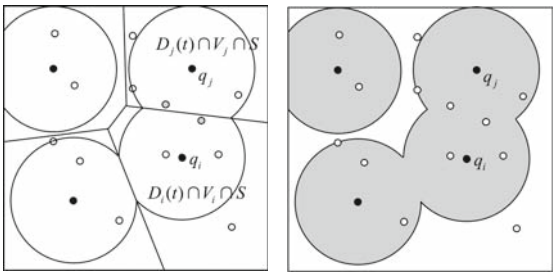
$$Var(K(t)) = n \left\{ \sum_{k=1}^{m} k^2 |D_{(k)}(t) \cap S| - \left(\sum_{k=1}^{m} k|D_{(k)}(t) \cap S| \right)^2 \right\}. \tag{8.12}$$

Note that $E(K(t))$ of (8.7) and $Var(K(t))$ of (8.8) are specific cases of that of (8.11) and that of (8.12), respectively, where $k = 1$. Also note that (8.5) and (8.6) both lead to false values.

The exact test can be achieved using the univariate multinomial distribution with parameters n, m, $|D_{(k)}(t) \cap S|$, $k = 1, \ldots, m$. For a large n, this distribution is approximated by the normal distribution with the expected value given by (8.11) and the variance given by (8.12). The computational time hinges on the geometrical computation of $|D_{(k)}(t) \cap S|$, $k = 1, \ldots, m$. Because m intersections are examined for m disks, computational time is of order $O(m^2)$.

8.3 Local and Global Voronoi Cross K Functions

The local space $D_i(t)$ of the *local cross K* function includes a local space, i.e., $D_i(t) \subset S$, but for a large t, the local space $D_i(t)$ includes the whole space, i.e., $S \subseteq D_i(t)$. This sounds somewhat peculiar since $D_i(t)$ is called a local space. Rather, it is more natural that a local space remains part of the global space S. To formulate a more natural method, this section proposes alternative *local* and *global cross K* functions whose statistical properties are nicer than those of the ordinary *local* and *global cross K* functions defined in Sect. 8.2.



(a) (b)

Fig. 8.4 Local Voronoi cross K function (a) and global Voronoi cross K function (b)

A natural way of defining local spaces is to tessellate the global space S into local spaces that are mutually exclusive and collectively exhaustive, for example, the whole area of Japan consists 47 local areas called prefectures. In the context of the railway stations Q in S, it is natural to regard the neighborhoods of the railway stations as local spaces. If people use their nearest railway station, the local spaces are given by the Voronoi diagram generated by the railway stations. Generally, let $V = \{V_1, \ldots, V_m\}$ be the Voronoi diagram generated by points Q, and $P_i = \{p_{i1}, \ldots, p_{in_i}\}$ be the set of points of P that are included in the ith Voronoi polygon V_i (Fig. 8.4a). By definition, the equations $\bigcup_{i=1}^{m} P_i = P$, $P_i \cap P_j = 0$, $i \neq j$, $i, j = 1, \ldots, m$ and $\sum_{i=1}^{m} n_i = n$ holds. For given P_i in V_i, let

$$K_{V_i}(t) = \text{the number of points of } P_i \text{ in } D_i(t) \cap V_i \cap S. \tag{8.13}$$

This function can also be regarded as a *local cross K* function. To distinguish it from the *local cross K* function defined by (8.1), the function given by (8.13) is referred to as the *local Voronoi cross K function*.

For a given t and P_i, under the null hypothesis H_o, the statistical properties of $K_{V_i}(t)$ are almost the same as those of $K_i(t)$ except for parameter values. The random variable $K_{V_i}(t)$ follows the binomial distribution with parameters n_i and $|D_i(t) \cap V_i \cap S|$. Its expected value and variance are given by

$$E(K_{V_i}(t)) = n_i |D_i(t) \cap V_i \cap S|, \tag{8.14}$$

$$Var(K_{V_i}(t)) = n_i |D_i(t) \cap V_i \cap S|(1 - |D_i(t) \cap V_i \cap S|). \tag{8.15}$$

The exact statistical test for the null hypothesis H_o can be achieved using the binomial distribution with parameters n_i and $|D_i(t) \cap V_i \cap S|$. For a large n_i, this distribution is approximated by the normal distribution with the expected value given by (8.14) and the variance given by (8.15). Therefore, the exact, as well as an approximate, statistical test is straightforwardly done.

The computation time is almost the same as that of the *local cross K* function, although a pre-processing is necessary for constructing the Voronoi diagram. The order of the computation time is $O(m \log m)$ in the worst case, but if the bucketing

method is used, the average computation time is $O(m)$ (Okabe et al., 2000). Once the Voronoi diagram is given, the value of $|D_i(t) \cap V_i \cap S|$ is explicitly written as an algebraic function of the coordinates of q_i and vertices of S for a continuous t (Okabe et al., 2000, 515–516). Therefore the computation of the values of $E(K_{V_i}(t))$ and $Var(K_{V_i}(t))$ is readily done with a constant computation time, i.e., the order of computation time is $O(1)$.

Paralleling the extension from the *local cross K* function $K_i(t)$ to the *global cross K* function $K(t)$ shown in Sect. 8.2, the local Voronoi cross K function can be extended to the *global Voronoi K function* as

$$K_V(t) = \sum_{i=1}^{m} K_{V_i}(t). \tag{8.16}$$

This function indicates the number of the points P that are included in $D_i(t) \cap V_i \cap S$, $i = 1, \ldots, m$. Because the Voronoi diagram $V = \{V_1, \ldots, V_m\}$ is a tessellation of a given space, the areas $D_i(t) \cap V_i \cap S$, $i = 1, \ldots, m$ are mutually exclusive except at their boundaries and collectively exhaustive of S; there are no overlaps among the areas $D_i(t) \cap V_i \cap S$, $i = 1, \ldots, m$ except at their boundaries. The union of these areas, i.e., $\bigcup_{i=1}^{m} D_i(t) \cap V_i \cap S$, is the buffer zone of the set of points Q, denoted by $B_Q(t)$ (the gray colored area in Fig. 8.3b). Therefore the function $K_V(t)$ implies the number of the points of P that are included in the buffer zone $B_Q(t)$.

The *global cross K* function of (8.2) and the *global Voronoi cross K* function of (8.16) deal with points of P in the global space S in a similar fashion; both deal with the points included in the area $\bigcup_{i=1}^{m} D_i(t) \cap S$. However, a distinct difference exists between them. The *global cross K* function $K(t)$ possibly counts the same points of P more than once; for example, the gray colored points in Fig. 8.4b are counted twice in $K(t)$. Stated more precisely, the points in the area $D_{(k)}(t) \cap S$, $k = 1, \ldots, m$ are counted k times in $K(t)$. On the other hand, the *global Voronoi cross K* function $K_V(t)$ counts each point of P at most once; for example, the gray colored points in Fig. 8.3b are counted once in $K_V(t)$. Stated more explicitly, the points in the area $D_{(k)}(t) \cap S$, $k = 1, \ldots, m$ are counted only once in $K_V(t)$. Both functions capture different aspects of the distribution of the points P.

The statistical properties of the *global Voronoi cross K* function $K_V(t)$ are nicer than those of the *global cross K* function $K(t)$. Unlike the case of $K(t)$, the derivation of the expected value and variance of $K_V(t)$ is much simpler. The reason is that the areas $D_i(t) \cap V_i \cap S$, $k = 1, \ldots, m$ are mutually exclusive, and hence the random point process of $K_V(t)$ is the binomial point process in which each point is placed in the buffer zone $B_Q(t)$ (the gray colored area) in S or its compliment $S \backslash B_Q(t)$ (the white area in S). Under the null hypothesis H_o, the random variable $K_V(t)$ follows the binomial distribution with parameters n and $|B_Q(t) \cap S|$. Therefore, the expected value and variance of $K_V(t)$ are simply given by

$$E(K_V(t)) = n|B_Q(t) \cap S| = n \left| \bigcup_{i=1}^{m} D_i(t) \cap V_i \cap S \right|, \tag{8.17}$$

$$Var(K_V(t)) = n|B_Q(t) \cap S|(1 - |B_Q(t) \cap S|)$$
$$= n \left| \bigcup_{i=1}^{m} D_i(t) \cap V_i \cap S \right| \left(1 - \left| \bigcup_{i=1}^{m} D_i(t) \cap V_i \cap S \right| \right). \quad (8.18)$$

The exact statistical method for testing the null hypothesis can be achieved in terms of the binomial distribution with parameters n and $|B_Q(t) \cap S|$, which is approximated by the normal distribution with the expected value of (8.17) and the variance of (8.18).

The computational method is almost the same as that of the *local Voronoi cross K* function. In pre-processing, the Voronoi diagram is constructed. The order of computational time is $O(m \log m)$ in the worst case and $O(m)$ on average. The value of $|B_Q(t) \cap S|$ is explicitly written as an algebraic function of the coordinates of q_i, $i = 1, \ldots, m$ and vertices of S for a continuous t (Okabe et al., 2000, 515–516). Therefore the computation of the values of $E(K_V(t))$ and $Var(K_V(t))$ is readily done with constant computation time, i.e., the order of computation time is $O(1)$.

8.4 Local and Global Auto K Functions

A distinct difference between the K functions to be discussed in this section and the *cross K* functions discussed in the preceding sections is that the former deals with spatial relations among points of only one set of points P (e.g., crime spots), whereas the latter deals with two different sets of points P (e.g., crime spots) and Q (railway stations). To contrast with "cross" as in the *cross K* functions, the K functions can be called *auto K functions* (where the term "auto" means among themselves, as in the spatial autocorrelation).

As referred to in the introduction, a *local auto K function* was first proposed by Getis and Franklin (1987). A slightly different definition of a *local auto K* function is given by

$$K_{A_i}(t) = \text{the number of points of } P_{-i} \text{ in } D_i(t) \cap S, \text{ where } P_{-i} = P \backslash \{p_i\}. \quad (8.19)$$

The definition looks almost the same as that of the *local cross K* function of (8.1), but in the context of testing the null hypothesis H_o, a big difference exists in that the base point p_i in the *local auto K* function is a random point, whereas the base point q_i in the *local cross K* function is a fixed point. For a fixed location of p_i, under the null hypothesis H_o, the random variable $K_{A_i}(t)$ follows the binomial distribution with parameters $n - 1$ and $|D_i(t) \cap S|$ (notice that it is n in the local cross K function). Therefore, the conditional expected value and the variance of $K_{A_i}(t)$ are given by

$$E(K_{A_i}(t)|p_i) = (n-1)|D_i(t) \cap S|, \tag{8.20}$$

$$Var(K_{A_i}(t)|p_i) = (n-1)|D_i(t) \cap S|(1 - |D_i(t) \cap S|). \tag{8.21}$$

The unconditional expected value $E(K_{A_i}(t))$ is obtained from

$$E(K_{A_i}(t)) = (n-1) \int_{p_i \in S} |D_i(t) \cap S| dp_i. \tag{8.22}$$

The integral of this equation indicates the expected area of $D_i(t) \cap S$. Stated a little more explicitly, the integral indicates the expected area of the intersection of a disk and S when the disk is randomly placed on S (recall $|S| = 1$ is assumed). Applying a formula of integral geometry (Santaló, 1976, (6.67)), (8.22) is written as

$$E(K_{A_i}(t)) = (n-1) \frac{2\pi^2 t^2}{2\pi(2\pi t^2 + 1) + 2\pi t|\partial S|}, \tag{8.23}$$

where $|\partial S|$ denotes the perimeter of S.

The unconditional variance $Var(K_{A_i}(t))$ is obtained from

$$Var(K_{A_i}(t)) = (n-1) \int_{p_i \in S} |D_i(t) \cap S| dp_i + (n-1) \int_{p_i \in S} |D_i(t) \cap S|^2 dp_i. \tag{8.24}$$

The first term is given by (8.23) but it is difficult to obtain the explicit form of the second term.

From the above examination, it is concluded that it is difficult to test the null hypothesis H_o exactly using the *local auto K* function. In theory, the observed value is compared with the exact expected value of (8.23). A statistical test should be done using Monte Carlo simulation.

The *local auto K* function can be extended to the *global auto K function* as

$$K_A(t) = \sum_{i=1}^{n} K_{A_i}(t). \tag{8.25}$$

To obtain the distribution of $K_A(t)$, recalling the case of the *local auto K* function, one might attempt to obtain the conditional distribution of $K_A(t|p_1, \ldots, p_n)$. However, this does not work because all points p_i, $i = 1, \ldots, n$ are fixed. An alternative procedure is: first, n points, $R = \{r_1, \ldots, r_n\}$, are randomly generated according to the uniform distribution over S; second, the disk centered at r_i truncated by S, i.e., $D_i(t) \cap S$, is generated for $i = 1, \ldots, n$; third, $K_{A_i}(t)$ is defined by the number of points P included in $D_i(t) \cap S$, $i = 1, \ldots, n$; fourth, $K_A(t)$ is defined by (8.25) for these $K_{A_i}(t)$, $i = 1, \ldots, n$. Then, for a given $R = \{r_1, \ldots, r_n\}$, the conditional distribution $K_A(t|r_1, \ldots, r_n)$ can be defined for the null hypothesis H_o. This conditional distribution follows the univariate multinomial distribution with parameters, n, n, $|D_{(k)}(t) \cap S|$, $k = 1, \ldots, n$ (Johnson et al., 1992, 460–461).

The probability that the conditional *global auto K* function takes a specific value $K_A(t|r_1, \ldots, r_n)$ is given by

$$\sum \frac{n!|D_{(0)}(t) \cap S|^{K_{(0)}(t)}|D_{(1)}(t) \cap S|^{K_{(1)}(t)} \cdots |D_{(n)}(t) \cap S|^{K_{(n)}(t)}}{K_{(0)}(t)!K_{(1)}(t)! \cdots K_{(n)}(t)!}, \quad (8.26)$$

where the summation is over all possible nonnegative integers $K_{(k)}(t), k = 1, \ldots, n$ such that $\sum_{k=1}^{n} K_{(k)}(t) = n$. The conditional expected value and the variance of $K_A(t)$ are given by

$$E(K_A(t)|p_i) = (n-1)\sum_{k=1}^{n} k|D_{(k)}(t) \cap S|, \quad (8.27)$$

$$Var(K_A(t)|p_i) = (n-1)\left\{\sum_{k=1}^{n} k^2|D_{(k)}(t) \cap S| - \left(\sum_{k=1}^{n} k|D_{(k)}(t) \cap S|\right)^2\right\}. \quad (8.28)$$

The unconditional expected value is obtained from

$$E(K_A(t)) = (n-1)\sum_{k=1}^{n} k \int_{r_1,\ldots,r_n \in S} |D_{(k)}(t) \cap S| dr_1 \cdots dr_n. \quad (8.29)$$

The integral term means the expected area in which exactly k disks are overlapped in S when n disks are randomly placed on S. Applying a formula of integral geometry (Santaló, 1976, (6.67)), (8.29) is written as

$$E(K_A(t)) = (n-1)\sum_{k=1}^{n} k \frac{\binom{n}{k}(2\pi^2 t^2)^k (2\pi + 2\pi t|\partial S|)^{n-k}}{(2\pi(2\pi t^2 + 1) + 2\pi t|\partial S|)^n}. \quad (8.30)$$

Like the case of the *local auto K* function, it is difficult to obtain the unconditional variance of $K_A(t)$ from (8.28).

8.5 Summary and Conclusion

This chapter has dealt with a class of K functions: the *local cross K* function, the *global cross K* function, the *local Voronoi K* function, the *global Voronoi K* function, the *local auto K* function, and the *global auto K* function, whose major statistical properties can be summarized as follows:

1. The *local cross K* function follows the binomial distribution with parameters n and $|D_i(t) \cap S|$. The exact expected value and variance are given by (8.3) and (8.4), respectively. The order of computing these values is $O(1)$.

2. The *global cross K* function follows the univariate multinomial distribution with parameters n, m, $|D_{(k)}(t) \cap S|$, $k = 1, \ldots, m$. The exact expected value and variance are given by (8.11) and (8.12), respectively. The order of computing these values is $O(m^2)$.
3. The *local Voronoi cross K* function follows the binomial distribution with parameters n_i and $|D_i(t) \cap V_i \cap S|$. The exact expected value and variance are given by (8.14) and (8.15), respectively. The order of computing these values is $O(1)$.
4. The *global Voronoi cross K* function follows the binomial distribution with parameters n and $|B_Q(t) \cap S|$. The exact expected value and variance are given by (8.17) and (8.18), respectively. The order of computing these values is $O(1)$.
5. The *local auto K* function follows a parametric binomial distribution, but its parameters are difficult to obtain. However, the exact expected value is explicitly given by (8.23), and its computational time is of order $O(1)$.
6. The *global auto K* function follows a parametric binomial distribution, but its parameters are difficult to obtain. However, the exact expected value is explicitly given by (8.30), and its computational time is of order $O(1)$.

We recognize that the Getis and Franklin technique is one from a family of similar techniques. Typically, local statistics are generated by "localizing" global statistics, which is what Getis and Franklin did. Here we have shown that it is possible (and perhaps more straightforward) to go in the other direction, i.e., creating global statistics as sums of local statistics.

We observe that exact tests are not possible for all K functions; some need to be evaluated using Monte Carlo tests. Further, the K functions differ considerably in terms of the computational effort involved (order of computing).

In terms of applications, in general, *global* and *local K* functions can be used to test different types of hypotheses, e.g., the *global K* functions can be used to test if crimes tend to occur around railway stations while the *local K* functions can be used to test if crimes tend to occur around specific stations. Further, specific types of a given K function (i.e., *local* or *global*) can be used to specify these general hypotheses more specifically, e.g., the *global cross K* function tests if the number of crimes within distance t of a typical station is significantly different from chance while the *global Voronoi cross K* function tests if the number of crimes within distance t of all stations is significantly different from chance.

Acknowledgements We express our thanks to Kei-ich Okunuki and Ikuho Yamada for their comments on earlier drafts.

Chapter 9
Spatial Point Pattern Analysis of Plants

Janet Franklin

9.1 Introduction

Plants, especially terrestrial long-lived perennials such as trees, do not usually move once established. Spatial patterns of sessile organisms can suggest or reveal ecological processes affecting the population or community in the present or the past – dispersal, establishment, competition, mortality, facilitation, growth – and as such, patterns of plants motivated early developments in spatial statistics (Pielou, 1977; Diggle, 1983). Specifically, it is intuitive to treat individual plants (or other sessile organisms) as discrete events on a plane whose locations are known and generated by point pattern processes (Ripley, 1981; Diggle, 1983; Fortin and Dale, 2005). Second-order point pattern statistics are used to measure their spatial pattern.

Arthur Getis (Getis and Franklin, 1987) introduced ecologists to the application of *local* spatial statistics, specifically *neighborhood* second-order point pattern analysis, to maps of organisms. As Wiegand and Moloney (2004) noted in their review paper, second-order global statistics based on the distribution of distances between pairs of points, especially Ripley's K-function (Ripley, 1976, 1977) derived from distances between all pairs, have been widely used in plant ecology. However, their review does not mention neighborhood analysis or local measures of spatial association (Anselin, 1995) at all. This chapter revisits the impact of the Getis and Franklin paper on the practice of spatial point pattern analysis in plant ecology, and specifically aims to determine if local statistics are being used and how.

9.2 Questions Addressed with Point Pattern Analysis of Plants

Most applications of point pattern analysis in plant ecology try to determine to what degree processes that cause patchiness, clumping or aggregation (local concentrations of events), and those that cause overdispersion, repulsion or regular patterns,

J. Franklin
School of Geographical Sciences and Urban Planning, and School of Life Sciences, Arizona State University
e-mail: Janet.Franklin@asu.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*, Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_9,

have affected the development of a population or community. "One of the most commonly observed spatial patterns in forests is the tendency for understory young stems to be clumped and for canopy old stems to be more uniformly distributed" (McDonald et al., 2003). Paraphrasing Franklin and Rey (2007), a clumped pattern of individuals or the aggregation of juveniles near adults might result from limited dispersal or from environmental heterogeneity (Palmiotto et al., 2004), while an increasingly regular pattern of older plants (Condit et al., 2000) would result from density- or distance-dependent juvenile mortality due to predation (Janzen, 1970; Connell, 1971; Connell et al., 1984), or inter-specific or intra-specific competition (Kenkel, 1988; Barberis and Tanner, 2005; Stoll and Bergius, 2005). If younger plants are found near older ones less frequently than expected by chance this would also suggest density dependent mortality (Clark and Clark, 1984). This theme is repeated often in the literature. For example, from Mast and Wolf (2004, p. 168):

> although initial spatial patterns may be determined by regeneration mechanisms, subsequent spatial distributions may result from the ability of individual trees to survive competition and dominate the patch (Oliver and Larson, 1990; Deutschman et al., 1993). As a forest ages, tree distributions within a patch may shift from a clumped distribution to a random (or regular) distribution due to self thinning and/or succession to shade-tolerant species (Cooper, 1961; Laessle, 1965; Whipple, 1980; Good and Whipple, 1982; Peet and Christensen, 1987).

Studies of processes that generate spatial pattern in plant communities focus on dispersal, establishment and mortality. Propagule dispersal can occur over short to long distances depending on the mechanisms or agents, e.g., wind, animals, gravity (Ridley, 1930; van der Pijl, 1972; Howe, 1986; Clark et al., 1999; Nathan and Muller-Landau, 2000), and competition among plants for light, water, nutrients and space is usually a very localized process.

Environmental heterogeneity occurs at multiple scales, as was nicely summarized by McDonald et al. (2003, p. 442): Many studies of mature and old-growth forests have found that establishment occurs preferentially in canopy gaps, leading to characteristic spatial clumping of new stems at the scale of a gap (e.g., Leemans, 1991; Moeur, 1993; Busing, 1996). Over time, stems that are crowded by other stems are more likely to die, and the remaining stems are more regularly dispersed (Kenkel, 1988; Moeur, 1993; Busing, 1996; He and Duncan, 2000). Many other factors may obscure the trend from a clumped understory distribution to a more regular overstory distribution, including patterns of seed dispersal (Fowler, 1986), windthrow events (Ida, 2000), and surface fire (Miller and Urban, 1999).

Thus, the inherent or endogenous biological processes (Fortin and Dale, 2005) expected to generate detectable spatial pattern in populations of sessile organisms include: dispersal limitations leading to clustering, spatial competition or inhibition resulting in less clustered patterns or repulsion between types (species, age classes), and facilitation leading to clustered patterns or attraction between types. The fundamental paradox of spatial pattern analysis is that the same pattern can result from different processes, e.g., according to the principle of equifinality. Clustering can arise from endogenous processes such as facilitation or dispersal limitations, or exogenous ones, such as spatial heterogeneity in environment, and pattern analysis

alone cannot distinguish the causes. This is emphasized in all treatments of spatial analysis including recent ones (Fortin and Dale, 2005; Perry et al., 2006).

9.3 The Impact of Getis' Paper on Plant Ecology

Getis and Franklin (1987) is cited for a number of different reasons. First, it discussed the application of second order analysis to point patterns of trees. It presented the following version of the local $L_i(t)$ statistic (using the notation $L(d)$):

$$\hat{L}_i(t) = \sqrt{A \sum_{j=1}^{n} k_{ij} / \pi(N-1)}, \tag{9.1}$$

where A is the area of the region, N is the number of points, and k_{ij} is 1 for all points j that are within distance t of point i. In contrast, the global $L(t)$ statistic is a square-root transformation of what is often referred to as Ripley's K-function (Ripley, 1976, 1977):

$$\hat{L}(t) = \sqrt{\frac{\hat{K}(t)}{\pi}} - t, \tag{9.2}$$

where

$$\hat{K}(t) = A \left(\sum_{i \neq j}^{N} \sum_{j \neq i}^{N} k_{ij} \right) / (N(N-1)). \tag{9.3}$$

Ripley (1981) attributed this transformation $L(t)$ to Besag (1977b), and it linearizes the relationship of $K(t)$ to t (distance) and stabilizes the variances. The transformation presented in Getis and Franklin for the local $L_i(t)$, shown in (9.1), makes $L(t) = t$ for a Poisson process, while a more-commonly used standardization of global Ripley's K-function, expressed here as the difference between observed and expected, makes $L(t) = 0$ for a Poisson process (Fortin and Dale, 2005) as in (9.2).

Secondly, Getis and Franklin presented a weighted boundary or edge correction, which Fortin and Dale attribute to Diggle (1983) and others (Pelissier and Goreaud, 2001) attribute to Ripley (1977). In this correction, if a circle centered on i with radius $t_{i,j}$ is completely within the study area the weight is 1, otherwise the weight is the reciprocal of the circles' circumference within the plot. Third, the paper used fixed values to approximate confidence intervals, based on Ripley (1977). Currently most use Monte Carlo simulation to evaluate significance, also discussed by Diggle (1983). Finally, but most importantly, the main purpose of the paper was to illustrate the application of a new local point pattern statistic to ecological data. "The method, second order *neighborhood* analysis, is a variation on second-order

analysis of point patterns . . . [which] while similar [to second order analysis] differs in that consideration is only given to those pairs of points having as one of its members a given point i" (Getis and Franklin, 1987, p. 473, emphasis added). Fortin and Dale (2005) stated, as other have, that global spatial statistics, summarizing the spatial pattern of the study area in a single number, are not appropriate when the assumption of stationarity is violated, and instead local spatial statistics should be used to estimate the spatial pattern for each sample location. Local spatial statistics have been developed and applied more recently than global (Anselin, 1995), and some promising methods have not yet been exploited in ecology (e.g., Okabe et al., 2000).

The 1987 paper used, as an illustration of the local statistic, a map of 108 ponderosa pine trees in a 120 m × 120 m area in northern California, USA, subset from an earlier much more extensive study (Franklin et al., 1985). In the original study we mapped 5,147 trees in six 11-ha plots from air photo interpretation, and applied several methods of global second-order analysis to these mapped point patterns, including spectral analysis which is not used very often in this context (Mugglestone and Renshaw, 1996). Ripley's K-function was calculated but the results were not presented graphically because they were so similar to the spectral analysis. When Dr. Getis and I wrote the paper illustrating his local point pattern analysis method, we originally submitted it as a technical note to *Ecology*, the premier high-impact North American journal in its field. Not only was it accepted with very little revision, the editor asked us if it could be published as a regular full-length paper. That was the first and last time such a thing has ever happened to me.

Of 98 papers citing Getis and Franklin (1987) (according to ISI Web of Science, September 2006), half (48) were related to point pattern analysis of plants (including macroalgae), overwhelmingly trees, while another 13% involved patterns of other organisms (animals) or biological variables excluding disease. Almost 15% were epidemiological studies of patterns of disease outbreaks or vectors, and the balance were either methodological papers or other applications of point patterns analysis, for example in human geography. Notably, of the 48 plant ecology studies, very few actually used the local point pattern statistic introduced in Getis and Franklin, or any type of local spatial statistic!

To examine these citation patterns in greater depth I reviewed 50 papers citing Getis and Franklin in detail, including 40 from plant ecology and 10 others. Collectively they cited it for 62 reasons (one paper could have cited it more than once for different reasons). A large number of them (38%) cited it with reference to second order point-pattern analysis using global Ripley's K-function, or even for its general definition of second order statistics (Kenkel, 1988; Moeur, 1993; Larsen and Bliss, 1998; Chen and Bradshaw, 1999; Mast and Veblen, 1999; Parish et al., 1999; Condit et al., 2000; Crook et al., 2001; Gu et al., 2001; Guerra et al., 2001; Schooley and Wiens, 2001; Mast and Wolf, 2004; Spooner et al., 2004b; Youngblood et al., 2004; Kashian et al., 2005; Munyekenye et al., 2005; Schroff et al., 2006). Strictly speaking, the paper did not present any results for (global) Ripley's K-function, and the original citations for the method are Ripley's (1976, 1977). Some refer specifically to its analysis of spatial pattern in ponderosa pines (Mast and Wolf,

2004; Youngblood et al., 2004) and, interestingly, both for its findings of clumped (Youngblood et al., 2004) and random (Wolf, 2005) spatial patterns of trees. In these cases the original study would have been a more appropriate citation (Franklin et al., 1985), as Getis and Franklin did not discuss the pattern of ponderosa pines per se. These papers and others will be considered in the next section reviewing recent applications of second order point pattern analysis in plant ecology.

Quite a few (28%) cited it with reference to the weighted edge correction (Haase, 1995; Haase et al., 1996; Pelissier, 1998; Cole and Syms, 1999; Lookingbill and Zavala, 2000; Gu et al., 2001; Schooley and Wiens, 2001; Call and Nilsen, 2003; Malkinson et al., 2003; Tirado and Pugnaire, 2003; Liang et al., 2004; LaFrankie and Saw, 2005; Shi et al., 2006), which is actually from Diggle (1977) as noted above. Ironically, more recent studies that describe and test edge correction methods incorrectly attribute the weighted edge correction to Getis and Franklin in one case (Haase, 1995), and point out errors in the formulas we published in the other case (Goreaud and Pelissier, 1999). I refer the reader to the later paper for a detailed discussion of the edge correction.

Two papers referred to the approximation for confidence intervals attributable to Ripley as noted above (Haase, 1995; Pancer-Koteja et al., 1998). One referred to it for its description of weighting Ripley's K-function by some quantitative attribute of the points (Donnegan and Rebertus, 1999), such as size or age. While Getis and Franklin discussed weighting in their conclusion, they did not implement it, although Wells and Getis (1999) did. Several papers actually cited it with reference to global Ripley's K-function, and used this global point pattern statistic in their study, but mistakenly refer to it as second order neighborhood analysis (Stamp and Lucas, 1990; Nicotra, 1998; Parker et al., 2001), presumably based on their misreading of Getis and Franklin and lack of knowledge of developments in local spatial statistics. This semantic confusion is regrettable but perhaps understandable because Ripley's K-function is based on all interpoint distances and averages the pattern observed in "neighborhoods" of a range of sizes (lag distance or scale) around each point.

It is puzzling that many of these studies would cite Getis and Franklin (1987) or other recent papers (especially Moeur, 1993; Haase, 1995) with regard to the general methods of second order point patterns statistics and their application to ecological patterns instead of the foundational work published earlier. Although the classic texts by Ripley (1981) and Diggle (1983) focus on the statistical methods rather than underlying ecological mechanisms that produce pattern, they are full of examples of applications to patterns of trees, as well as other biological (birds nests, cells) and non-biological (magnetite crystals) phenomena (see also Pielou 1977).

Thirteen studies (26%) did refer specifically to the local version of Ripley's K-function published in Getis and Franklin. A few of these studies implement it, while others develop new local spatial statistics or apply other existing ones. These studies will also be discussed below.

9.4 Recent Applications of Ripley's K-function and Related Methods in Plant Ecology

Recent studies addressing global Ripley's K-function can be divided into two groups, those that apply second-order point pattern analysis to address ecological questions about plant populations and communities, and those that present methodological developments. I identified several themes in the recent literature that I will highlight, although I make no claim that this is a comprehensive review of the literature and apologize in advance for my oversights.

Many studies, almost all of forest trees, begin with an observed or hypothesized pattern of clumped juveniles and random or regular adults. They conclude that initially clumped establishment results from dispersal limitations, e.g., juveniles clustered near adults (Forget et al., 1999; Condit et al., 2000), gap-phase regeneration, e.g., tree fall gaps (Pancer-Koteja et al., 1998; Pelissier, 1998; Forget et al., 1999; Mast and Veblen, 1999; Mast and Wolf, 2004; Wolf, 2005) or larger-scale gaps resulting from disturbances such as fire or wind (Parish et al., 1999; Wells and Getis, 1999; McDonald et al., 2003). The relatively more regular patterns of adults was frequently attributed to density-dependent (non-random) mortality (but see Wiegand et al., 2000) resulting from interspecific or intraspecific competition (Kenkel, 1988; Moeur, 1993; Haase et al., 1996; Chen and Bradshaw, 1999; Donnegan and Rebertus, 1999; Parker et al., 2001; Malkinson et al., 2003; McDonald et al., 2003; Kashian et al., 2005; Wolf, 2005), a conclusion also supported by simulation modeling in the study by Druckenbrod et al. (2005) (but see Moravie and Roberts, 2003). Competition was often invoked as an explanation of overdispersion or repulsion in studies of relatively species-poor plant communities. In a comprehensive survey of species-rich tropical forests, weakening aggregation in larger (older) trees was attributed to density-dependent mortality (Condit et al., 2000).

Some of these studies used weighted Ripley's K-function to explicitly examine pattern as a function of tree size or age (Donnegan and Rebertus 1999; Wells and Getis 1999). Others have used Ripley's K-function and related methods in a general way to describe point pattern as a function of scale, and where clumping was detected it was attributed to various causes including environmental heterogeneity (Navas and Goulard, 1991; Couteron and Kokou, 1997; Cole and Syms, 1999; Forget et al., 1999; Spooner et al., 2004b; Youngblood et al., 2004). Finally, bivariate Ripley's K-function has been used to explicitly identify attraction between species or age-classes caused by facilitation (Donnegan and Rebertus, 1999; Lookingbill and Zavala, 2000; Malkinson et al., 2003; Tirado and Pugnaire, 2003; LaFrankie and Saw, 2005), and repulsion caused by competition (Haase et al., 1996; Call and Nilsen, 2003; Malkinson et al., 2003; McDonald et al., 2003).

Although point maps of individuals were used in these studies, I should mention that measures of spatial dependence in area data, for example Moran's I (Moran, 1948), are also applied to measures of plant abundance collected in contiguous regions (e.g., Wadda and Ribbens, 1997; Almeida-Neto and Lewinsohn, 2004; Fonseca et al., 2004; Franklin and Rey, 2007) but because point pattern analysis is the focus of this paper, I will not discuss those studies in detail.

Almost all of these studies used the global K-function to explore point patterns of plants by comparing the observed to the null hypothesis of CSR, a homogeneous Poisson process. Very few fit alternative models describing clustering or inhibition (or both) such as the heterogeneous Poisson process, Poisson cluster process, Strauss process or Markov point process. This is in spite of the fact that these alternative models have been around for quite a while (e.g., Neyman, 1939) and are described in classic and recent methodological books and papers (Pielou, 1977; Getis and Boots, 1978; Ripley, 1981; Diggle, 1983; Wiegand and Moloney, 2004). We fit alternative models of heterogeneous Poisson and Poisson cluster processes in our original paper (Franklin et al., 1985), another thing for which it is frequently overlooked (Perry et al., 2006).

9.5 Methodological Developments in Spatial Point Pattern Analysis Applied to Plants

Recent developments and applications have focused on several areas including (a) refined edge-correction methods, especially for irregular boundaries (Haase, 1995; Fortin and Dale, 2005), (b) the application of the neighborhood density function (NDF) or "O-ring statistic" based on discrete ranges of distances (annuli) rather than cumulative lag distance used in K(t) (Condit et al., 2000; Wiegand and Moloney, 2004; Perry et al., 2006), and (c) the development of methods to objectively delineate subregions within a heterogeneous pattern where stationarity can be assumed (Dale and Powell, 2001; Pelissier and Goreaud, 2001). These areas were the focus of two recent review and comparison papers (Wiegand and Moloney, 2004; Perry et al., 2006) and I refer the reader there. Both papers particularly advocated the use of the NDF for testing hypotheses in plant ecology. It will be interesting to see if it becomes more widely used. Again I would like to note that Voronoi methods also generate a set of characteristics that can be associated with individual points (Okabe et al., 2000) and may also be a useful approach to explore in plant ecology.

It is relevant to mention that the methods proposed for delineating homogeneous subregions are based on local measures of pattern. For example, Pelissier and Goreaud (2001) calculate local density at some specified lag for a regular array of points, and note that this is proportional to the local K(t) presented by Getis and Franklin (1987). They then fit a surface to those values (using loess regression, although other methods could be used) and divided the study area along a contour corresponding to a natural break in the frequency distribution of local densities.

9.6 Applications of Local Point Patterns Statistics in Plant Ecology

The potential usefulness of the local second-order point pattern statistic, the neighborhood $L_i(t)$, or local spatial statistics in general, has been highlighted recently (Fortin and Dale 2005, Perry et al. 2006). Perry et al. (2006) state that "while the

global tests suggest there is spatial segregation and at what scale(s), the local tests can explicitly show where this is occurring." And yet there have been very few applications of them since Getis and Franklin (1987) was published. Apparently it is still an idea that is ahead of its time.

Camarero et al. (2005) mapped the values of the local statistic to show where tree seedlings of a relict pine population were aggregated or repulsed at selected scales. This was one line of evidence supporting their conclusion that frequent short-distance dispersal events induced the primary spatial clustering of seedlings in safe sites, while wind turbulence caused rare medium-distance dispersal events, resulting in clustering at multiple scales. As part of a detailed study of subalpine forest succession based on forests reconstructed over time using dendrochronological (tree ring) data, Donnegan and Rebertus (1999) calculated a weighted bivariate neighborhood $L_i(t)$ from the number of spruce and fir neighbors surrounding a target adult limber pine to account for shading and other interactions between the species. They then used this index of clumping in a logistic regression model of pine survival and found that mortality was highest when pines were surrounded by many spruces and firs at lag distance of 2 m. Potvin et al. (2003), studying habitat selection by deer, used high values of the local K-function (at 0.5–2 km distance) based on observed locations of deer to map areas of animal concentrations, and found those patterns to be consistent with those derived from habitat selection indices and kernel estimators.

Dale and Powell (2001) presented a new method of second order point pattern analysis based on circumcircles, and included a spatially explicit mapping of positive and negative residual scores to locate events that are members of clusters vs. gaps. In one example this method was used to show that regions of high spruce seedling density and high tree density did not coincide, suggesting that local conditions for germination were as important to establishment as high density seed source.

Although they did not use local $L_i(t)$, Shi and Zhang (2003) applied local spatial statistics or "local indicators of spatial association" (LISA) (Anselin, 1995; Ord and Getis, 1995) to forests. These are area pattern, rather than point pattern statistics, applied to sample data, usually measurements of a continuous variable for a point or area. In this study several LISAs including local Moran's I, local Geary's C and the local G_i^* statistic, were derived from size measurements of individual trees and compared to traditional forestry competition indices used in models of tree growth. They performed quite well as predictors and were also useful for identifying clusters of trees of similar size.

Wells and Getis (1999) also used a local statistic G_i^* (Ord and Getis, 1995) applied to measurements of tree age made for individuals, to identify the locations of clusters of old or young trees. In their study Torrey pine trees (*Pinus torreyana*), a rare, endemic California pine species, were mapped in three 1-ha plots. They located clusters of younger trees and found greater clustering in young stands, consistent with establishment of cohorts following episodic disturbance (fire). This local area statistic was used in lieu of the local point pattern statistic, neighborhood $L_i(t)$, and without discussion of the earlier work, although Wulder and Boots (1998) identify second order neighborhood analysis (Getis and Franklin 1987) as an early

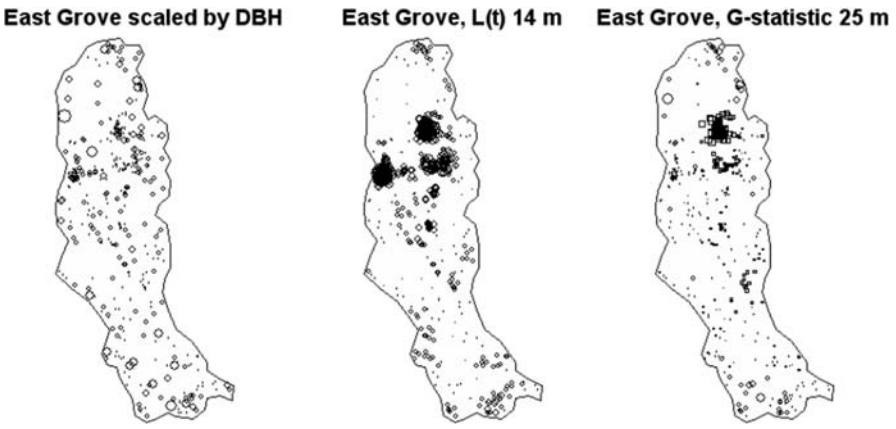


Fig. 9.1 Locations of 440 Torrey pine trees (*Pinus torreyana*) in the East Grove area of Torrey Pines State Reserve, La Jolla, CA, USA (E. Santos and J. Franklin, unpublished data). *Map on left* shows tree locations scaled by size (DBH, trunk diameter at 1.3 m height), and *center map* shown the tree locations scaled by the local value of L(t) at lag of 14 m (see Fig. 9.2). Negative values shown as *squares*, positive values as *circles*. *Map on right* shown tree locations scaled by the values of local Gi* (see text) where neighborhood contiguity is based on a lag distance of 25 m (maximum nearest neighbor distance used to avoid islands). These analyses were carried out using the spatstat package in the R statistical environment (R Development Core Team, 2004)

formulation if the Gi* statistic – the Getis model (circa 1984) – which it is. Recently, my student and I initiated a project to map the entire mainland population of Torrey pines (over 5,000 trees) and measure the size and condition of each tree for conservation monitoring purposes. For illustration, I show the distributions of all 440 trees in the 5.69-ha East Grove area (encompassing one of Wells and Getis' sites), of local $L_i(t)$, and of local G_i^* (9.1). Global $L(t)$ indicates significant clumping peaking at about 14-m lag distance, and Moran's I indicated positive spatial association of tree size at roughly the same scale (9.2). Figure 9.1 shows that regions of high tree density mainly comprise clumps of small trees, consistent with the previous observations of Wells and Getis (1999).

9.7 Conclusion

Judging from its citation patterns, the paper by Getis and Franklin (1987) continues to influence the practice of spatial point pattern analysis in plant ecology. However, most practitioners continue to apply global analyses, while local spatial statistics are mainly advocated by specialists in methodological papers. Therefore, the legacy of this paper is, in part, accidental. Although written to introduce local spatial statistics to ecologists, it is most often cited with reference to global point pattern statistics, perhaps because of its clear summary of classic work in this area, and perhaps also because of the visibility of the journal in which it appeared.

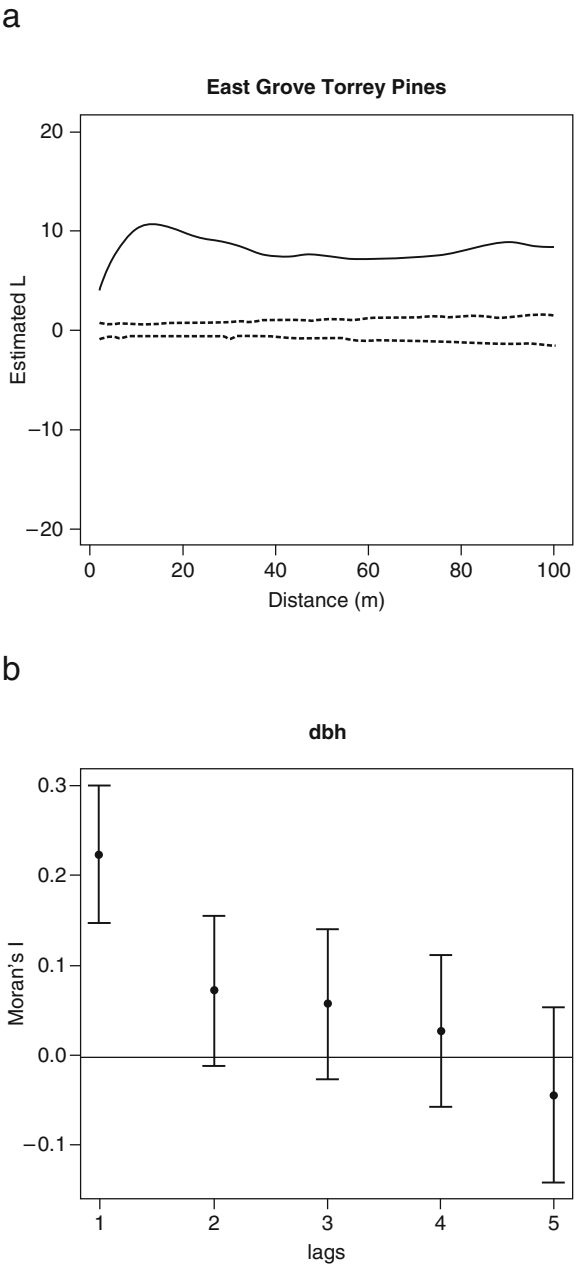


Fig. 9.2 Global $L(t)$ for the trees shown in Fig. 9.1 at lags of 2–100 m showing significant clumping at all scales and a peak in $L(t)$ at 10–18 m; (b) Moran's I as a measure of spatial autocorrelation of tree size (DBH, see Fig. 9.1 caption) where neighborhood contiguity is based on a lag distance of 10 m, indicating significant positive spatial association of tree size at lag 1 (10 m). These analyses were carried out using the splancs package (Rowlingson and Diggle, 1993) in the R statistical environment (R Development Core Team, 2004)

Those specialists considering methodology have focused their recent efforts on improved edge corrections and automated methods for delineating homogeneous subareas. They have advocated the application of the neighborhood density function to explore pattern as a function of discrete ranges of distances, and I would also advocate fitting alternative models of clustering or inhibition (certainly not a new idea). Perhaps citation patterns, like sausage- and law-making, should not be examined in such detail, but the citing of Getis and Franklin for reasons other than the local spatial statistic they introduced, in a majority of cases, is perhaps indicative of a gap that still exists between theory and practice in spatial analysis.

Where local spatial statistics have been applied to point patterns in forestry and ecology they have proven very useful in identifying individuals that are members of clusters or that fall within gaps, and in locating groups of individuals that share some characteristic (such as similar size). They have been related empirically to, e.g., regeneration patterns, growth characteristics, competition and mortality. Methodologically, measures of local context have been used delineate areas over which the assumption of stationarity is valid. There is considerable future opportunity for both exploratory spatial data analysis and hypothesis testing in spatial ecology using global and local methods.

Acknowledgements I thank E. Santos for the use of unpublished data from her graduate research, and the Torrey Pines Association and D.S. Smith, California State Parks, for supporting her study. My NSF grant on spatial inference and prediction from species data (0452389) supported the writing of this chapter, and I thank my lab reading group on spatial ecology for their feedback in Fall 2006. This chapter was greatly improved by the comments of S.J. Rey and B. Boots, and I thank S.J. Rey and L. Anselin for providing me the privilege of contributing to this book.

Part III
Local Statistics

Chapter 10
The Analysis of Spatial Association by Use of Distance Statistics

Arthur Getis and J. Keith Ord

This Chapter was originally published in:
Getis, A., Ord, K.(1992) The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis 24:189-206. Reprinted with permission of Blackwell Publishing, Oxford

Abstract Introduced in this paper is a family of statistics, G, that can be used as a measure of spatial association in a number of circumstances. The basic statistic is derived, its properties are identified, and its advantages explained. Several of the G statistics make it possible to evaluate the spatial association of a variable within a specified distance of a single point. A comparison is made between a general G statistic and Moran's I for similar hypothetical and empirical conditions. The empirical work includes studies of sudden infant death syndrome by county in North Carolina and dwelling unit prices in metropolitan San Diego by zip-code districts. Results indicate that G statistics should be used in conjunction with I in order to identify characteristics of patterns not revealed by the I statistic alone and, specifically, the G_i and G_i^* statistics enable us to detect local "pockets" of dependence that may not show up when using global statistics.

10.1 Introduction

The importance of examining spatial series for spatial correlation and autocorrelation is undeniable. Both Anselin and Griffith (1988) and Arbia (1989) have shown that failure to take necessary steps to account for or avoid spatial autocorrelation can lead to serious errors in model interpretation. In spatial modeling, researchers must not only account for dependence structure and spatial heteroskedasticity, they must also assess the effects of spatial scale. In the last twenty years a number of instruments for testing for and measuring spatial autocorrelation have appeared.

A. Getis (✉) and J.K. Ord
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

To geographers, the best-known statistics arc Moran's I and, to a lesser extent, Geary's c (Cliff and Ord, 1973). To geologists and remote sensing analysts, the semi-variance is most popular (Davis, 1986). To spatial econometricians, estimating spatial autocorrelation coefficients of regression equations is the usual approach (Anselin, 1988).

A common feature of these procedures is that they are applied globally, that is, to the complete region under study. However, it is often desirable to examine pattern at a more local scale, particularly if the process is spatially nonstationary. Foster and Gorr (1986) provide an adaptive filtering method for smoothing parameter estimates, and Cressie and Head (1989) present a modeling procedure. The ideas presented in this paper are complementary to these approaches in that we also focus upon local effects, but from the viewpoint of testing rather than smoothing.

This paper introduces a family of measures of spatial association called G statistics. These statistics have a number of attributes that make them attractive for measuring association in a spatially distributed variable. When used in conjunction with a statistic such as Moran's I, they deepen the knowledge of the processes that give rise to spatial association, in that they enable us to detect local "pockets" of dependence that may not show up when using global statistics. In this paper, we first derive the statistics $G_i(d)$ and $G(d)$, then outline their attributes. Next, the $G(d)$ statistic is compared with Moran's I. Finally, there is a discussion of empirical examples. The examples are taken from two different geographic scales of analysis and two different sets of data. They include sudden infant death syndrome (SIDS) by county in North Carolina, and house prices by zip-code district in the San Diego metropolitan area.

10.2 The $G_i(d)$ Statistic

This statistic measures the degree of association that results from the concentration of weighted points (or area represented by a weighted point) and all other weighted points included within a radius of distance d from the original weighted point. We are given an area subdivided into n regions, $i = 1, 2, \ldots, n$, where each region is identified with a point whose Cartesian coordinates are known. Each i has associated with it a value x (a weight) taken from a variable X. The variable has a natural origin and is positive. The $G_i(d)$ statistic developed below allows for tests of hypotheses about the spatial concentration of the sum of x values associated with the j points within d of the ith point.

The statistic is

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j}, \qquad j \text{ not equal to } i, \tag{10.1}$$

where $\{w_{ij}\}$ is a symmetric one/zero spatial weight matrix with ones for all links defined as being within distance d of a given i; all other links are zero including the link of point i to itself. The numerator is the sum of all x_j within d of i but not including x_i. The denominator is the sum of all x_j not including x_i.

Adopting standard arguments (cf. Cliff and Ord, 1973, pp. 32–33), we may fix the value x_i for the ith point and consider the set of $(n-1)!$ random permutations of the remaining x values at the j points. Under the null hypothesis of spatial independence, these permutations are equally likely. That is, let X_j be the random variable describing the value assigned to point j, then

$$P(X_j = x_r) = \frac{1}{(n-1)}, \qquad r \neq i,$$

and $E(X_j) = \sum_{r \neq i} x_r/(n-1)$. Thus

$$\begin{aligned} E(G_i) &= \textstyle\sum_{j \neq i} w_{ij}(d) E(X_j) / \sum_{j \neq i} EX_j \\ &= W_i/(n-1), \end{aligned} \tag{10.2}$$

where $W_i = \sum_j w_{ij}(d)$.

Similarly,

$$E(G_i^2) = \frac{1}{(\sum_j x_j)^2} \Big[\sum_j w_{ij}^2(d) E(X_j^2) + \sum \sum_{j \neq k} w_{ij}(d) w_{ik}(d) E(X_j X_k) \Big].$$

Since

$$E(X_j^2) = \sum_{r \neq i} x_r^2/(n-1)$$

and

$$\begin{aligned} E(X_j X_k) &= \textstyle\sum \sum_{r \neq s \neq i} x_r x_s/(n-1)(n-2) \\ &= \{(\textstyle\sum_{r \neq i} x_r)^2 - \sum_{r \neq i} x_r^2\}/(n-1)(n-2). \end{aligned}$$

Recalling that the weights are binary

$$\sum \sum_{j \neq k} w_{ij} w_{ik} = W_i^2 - W_i$$

and so

$$E(G_i^2) = \frac{1}{(\sum_j x_j)^2} \left\{ \frac{W_i \sum_j x_j^2}{(n-1)} + \frac{W_i(W_i-1)}{(n-1)(n-2)} [(\sum_j x_j)^2 - \sum_j x_j^2] \right\}.$$

Thus

$$\begin{aligned} Var(G_i) &= E(G_i^2) - E^2(G_i) \\ &= \frac{1}{(\sum_j x_j)^2} \left[\frac{W_i(n-1-W_i) \sum_j x_j^2}{(n-1)(n-2)} \right] \\ &\quad + \frac{W_i(W_i-1)}{(n-1)(n-2)} - \frac{W_i^2}{(n-1)^2}. \end{aligned}$$

Table 10.1 Characteristics of G_i statistics

	j not equal to i	j may equal i
Statistic	$G_i(d)$	$G_i^*(d)$
Expression	$\frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}$	$\frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}$
	$W_i = \sum_j w_{ij}(d)$	$W_i^* = \sum_j w_{ij}(d)$
Definitions	$Y_{i1} = \frac{\sum_j x_j}{(n-1)}$	$Y_{i1}^* = \frac{\sum_j x_j}{n}$
	$Y_{i2} = \frac{\sum_j x_j^2}{(n-1)} - Y_{i1}^2$	$Y_{i2}^* = \frac{\sum_j x_j^2}{n} - (Y_{i1}^*)^2$
Expectation	$W_i/(n-1)$	W_i^*/n
Variance	$\frac{W_i(n-1-W_i)Y_{i2}}{(n-1)^2(n-2)Y_{i1}^2}$	$\frac{W_i^*(n-W_i^*)Y_{i2*}}{n^2(n-1)(Y_{i1}^*)^2}$

If we set $\frac{\sum_j x_j}{(n-1)} = Y_{i1}$ and $\frac{\sum_j x_j^2}{(n-1)} - Y_{i1}^2 = Y_{i2}$, then

$$Var(G_i) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)}\left(\frac{Y_{i2}}{Y_{i1}^2}\right). \tag{10.3}$$

As expected, $Var(G_i) = 0$ when $W_i = 0$ (no neighbors within d), or when $W_i = n-1$ (all $n-1$ observations are within d), or when $Y_{i2} = 0$ (all $n-1$ observations are equal).

Note that W_i, Y_{i1}, and Y_{i2} depend on i. Since G_i is a weighted sum of the variable X_j, and the denominator of G_i is invariant under random permutations of $\{x_j, j \neq i\}$, it follows, provided $W_i/(n-1)$ is bounded away from 0 and from 1, that the permutations distribution of G_i under H_o approaches normality as $n \to \infty$; cf. Hoeffding (1951) and Cliff and Ord (1973, p. 36). When d, and thus W_i is small, normality is lost, and when d is large enough to encompass the whole study area, and thus $(n-1-W_i)$ is small, normality is also lost. It is important to note that the conditions must be satisfied separately for each point if its G, is to be assessed via the normal approximation.

Table 10.1 shows the characteristic equations for $G_i(d)$ and the related statistic, $G_i^*(d)$, which measures association in cases where the j equal to i term is included in the statistic. This implies that any concentration of the x values includes the x at i. Note that the distribution of $G_i^*(d)$ is evaluated under the null hypothesis that all n! random permutations are equally likely.

10.3 Attributes of G_i Statistics

It is important to note that G_i is scale-invariant ($Y_i = bX_i$ yields the same scores as X_i) but not location-invariant ($Y_i = a + X_i$ gives different results than X_i). The statistic is intended for use only for those variables that possess a natural origin. Like all other such statistics, transformations like $Y_i = \log X_i$, will change the results.

$G_i(d)$ measures the concentration or lack of concentration of the sum of values associated with variable X in the region under study. $G_i(d)$ is a proportion of the sum of all x_j values that are within d of i. If, for example, high-value $x_j s$ are within d of point i, then $G_i(d)$ is high. Whether the $G_i(d)$ value is statistically significant depends on the statistics distribution.

Earlier work on a form of the $G_i(d)$ statistic is in Getis (1984), Getis and Franklin (1987), and Getis (1991). Their work is based on the second-order approach to map pattern analysis developed by Ripley (1977).

In typical circumstances, the null hypothesis is that the set of x values within d of location i is a random sample drawn without replacement from the set of all x values. The estimated $G_i(d)$ is computed from (10.1) using the observed x_j values. Assuming that $G_i(d)$ is approximately normally distributed, when

$$Z_i = \{G_i(d) - E[G_i(d)]\}/\sqrt{VarG_i(d)} \tag{10.4}$$

is positively or negatively greater than some specified level of significance, then we say that positive or negative spatial association obtains. A large positive Z_i implies that large values of x_j (values above the mean x_j) are within d of point i. A large negative Z_i means that small values of x_j are within d of point i.

A special feature of this statistic is that the pattern of data points is neutralized when the expectation is that all x values are the same. This is illustrated for the case when data point densities are high in the vicinity of point i, and d is just large enough to contain the area of the clustered points. Theoretical $G_i(d)$ values are high because W_i is high. However, only if the observed x, values in the vicinity of point i differ systematically from the mean is there the opportunity to identify significant spatial concentration of the sum of $x_j s$. That is, as data points become more clustered in the vicinity of point i, the expectation of $G_i(d)$ rises, neutralizing the effect of the dense cluster of j values.

In addition to its above meaning, the value of d can be interpreted as a distance that incorporates specified cells in a lattice. It is to be expected that neighboring G_i will be correlated if d includes neighbors. To examine this issue, consider a regular lattice. When n is large, the denominator of each G_i is almost constant so it follows that corr (G_i, G_j) proportion of neighbors that i and j have in common.

Example 1. Consider the rook's case. Cell i has no common neighbors with its four immediate neighbors, but two with its immediate diagonal neighbors. The numbers of common neighbors are as illustrated below:

$$\begin{array}{ccccc} & 0 & 1 & 0 & \\ 0 & 2 & 0 & 2 & 0 \\ 1 & 0 & i & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 \\ & 0 & 1 & 0 & \end{array}$$

All the other cells have no common neighbors with i. Thus, the G-indices for the four diagonal neighbors have correlations of about 0.5 with G_i, four others have correlations of about 0.25 and the rest are virtually uncorrelated.

For more highly connected lattices (such as the queen's case) the array of nonzero correlations stretches further, but the maximum correlation between any pair of G-indices remains about 0.5.

Example 2.

$$\begin{array}{cccccccccc} m & m & m & m & m & m & m & m & m & m \\ m & A & A & A & m & m & B & B & B & m \\ m & A & A & A & m & m & B & B & B & m \\ m & A & A & A & m & m & B & B & B & m \\ m & m & m & m & m & m & m & m & m & m \end{array}$$

Set $A + B = 2m$, therefore $\bar{x} = m$; $n = 50$;
$A \geq 0$;
$B \geq 0$;
put $A = m(l + c)$, $B = m(l - c)$, $0 \leq c \leq 1$

Using this example, the G_i and G_i^* statistics are compared in the following table.

G_i and $G_i{}^*$ values (queen's case; non-edge cells)

Cell	G_i	$Z(G_i)$	G_i^*	$Z(G_i^*)$
A, surrounded by As	$\frac{8+8c}{49-c}$	$5.30^{\#}$	$\frac{9+9c}{50}$	5.47
A, adjacent to ms	$\frac{8+3c}{49-c}$	$2.06^{\#}$	$\frac{9+4c}{50}$	2.43
Central m, adjacent to As	$\frac{8+3c}{49}$	$1.89^{\#}$	$\frac{9+3c}{50}$	1.82
Other m, adjacent to As	$\frac{8+2c}{49}$	$1.26^{\#}$	$\frac{9+3c}{50}$	1.21

Values for B_s are the same, with negative signs attached
*These values are lower bounds as $c \to 1$; they vary only slightly with c

We note that G_i, and G_i^* are similar in this case; if the central A was replaced by a B, $Z(G_i)$ would be unchanged, whereas $Z(G_i^*)$ drops to 4.25. Thus, G_i and G_i^* typically convey much the same information.

Example 3. Consider a large regular lattice for which we seek the distribution under H_o for Gi^* with W_i neighbors. Let $p =$ proportion of $A_s =$ proportion of B_s and $1 - 2p =$ proportion of ms.

Let (k_1, k_2, k_3) denote the number of A_s, B_s, and ms, respectively so that $k_1 + k_2 + k_3 = n$. For large lattices, in this case, the joint distribution is approximately tri(multi-)nomial with index W and parameters $(p, p, 1 - 2p)$. Since $G_i^* = [W_i + (k_1 - k_2)c]/n$ clearly $E(G_i^*) = W_i/n$ as expected and $V(G_i^*) = 2pWi/n$, reflecting the large sample approximation. The distribution is symmetric and the

standardized fourth moment is

$$3 + \frac{1 - 6p}{2pW_i}.$$

This is close to 3 provided pW_i is not too small.

Since we are using G_i, and G_i^* primarily in a diagnostic mode, we suggest that $W \geq 8$ at least (that is, the queen's case), although further work is clearly necessary to establish cut-off values for the statistics.

10.4 A General *G* Statistic

Following from these arguments, a general statistic, $G(d)$ can be developed. The statistic is general in the sense that it is based on all pairs of values (x_i, x_j) such that i and j are within distance d of each other. No particular location i is fixed in this case. The statistic is

$$G_i(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d)x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \qquad j \text{ not equal to } i. \tag{10.5}$$

The G-statistic is a member of the class of linear permutation statistics, first introduced by Pitman (1937). Such statistics were first considered in a spatial context by Mantel (1967) and Cliff and Ord (1973), and developed as a general cross-product statistic by Hubert (1977, 1979) and Hubert et al. (1981).

For (10.5),

$$W = \sum_{i=1}\sum_{j=1} w_{ij}(d), \qquad j \text{ not equal to } i$$

so that

$$E[G(d)] = W/[n(n-1)]. \tag{10.6}$$

The variance of G follows from Cliff and Ord (1973, pp. 70–71):

$$E(G^2) = \frac{1}{(m_1^2 - m_2)^2 n^{(4)}}[B_0 m_2^2 + B_1 m_4 + B_2 m_1^2 m_2 + B_3 m_1 m_3 + B_4 m_1^4],$$

where $m_j = \sum_{j=1} {x_i}^j, j = 1, 2, 3, 4$ and $n^{(r)} = n(n-1)(n-2)\cdots(n-r+1)$. The coefficients, B, are

$$\begin{aligned}
B_0 &= (n^2 - 3n + 3)S_1 - nS_2 + 3W^2, \\
B_1 &= -\,[(n^2 - n)S_1 - 2nS_2 + 6W^2], \\
B_2 &= -\,[2nS_1 - (n+3)S_2 + 6W^2], \\
B_3 &= 4(n-1)S_1 - 2(n+1)S_2 + 8W^2, \quad \text{and} \\
B_4 &= S_1 - S_2 + W^2,
\end{aligned}$$

where $S_1 = 1/2 \sum_i \sum_j (w_{ij} + w_{ji})^2$, j not equal to i and $S_2 = 1/2 \sum_i (w_{i.} + w_{.i})^2$; $w_{i.} = \sum_j w_{ij}$, j not equal to i; thus

$$Var(G) = E(G^2) - \{W/[n(n-1)]\}^2 \tag{10.7}$$

10.5 The $G(d)$ Statistic and Moran's I Compared

The $G(d)$ statistic measures overall concentration or lack of concentration of all pairs of (x_i, x_j) such that i and j are within d of each other. Following (10.5), one finds $G(d)$ by taking the sum of the multiples of each x_i with all $x_j s$ within d of all i as a proportion of the sum of all $x_i x_j$. Moran's I, on the other hand, is often used to measure the correlation of each x_i with all $x_j s$ within d of i and, therefore, is based on the degree of covariance within d of all x_i. Consider K_1, K_2 as constants invariant under random permutations. Then using summation shorthand we have

$$G(d) = K_1 \sum \sum w_{ij} x_i x_j$$

and

$$\begin{aligned} I(d) &= K_2 \sum \sum w_{ij}(x_i - \bar{x})(x_j - \bar{x}) \\ &= (K_2/K_1)G(d) - K_2 \bar{x} \sum (w_{i.} + w_{.i}) x_i + K_2 \bar{x}^2 W, \end{aligned}$$

where $w_{i.} = \sum_j w_{ij}$ and $w_{.i} = \sum_j w_{ji}$.

Since both $G(d)$ and $I(d)$ can measure the association among the same set of weighted points or areas represented by points, they may be compared. They will differ when the weighted sums $\sum w_{i.} x_i$ and $\sum w_{.i} x_i$ differ from $W\bar{x}$, that is, when the patterns of weights are unequal. The basic hypothesis is of a random pattern in each case. We may compare the performance of the two measures by using their equivalent Z values of the approximate normal distribution.

Example 4.

Let us use the lattice of Example 2. As before,
Set $A + B = 2m$, therefore $\bar{x} = m$; $n = 50$;
$A \geq 0$;
$B \geq 0$;
put $A = m(1 + c), B = m(1 - c), 0 \leq c \leq 1$.
In addition, put

$a = A - m$;
$B - 2m - A = m - a$;
$B - m = a$;
$m \geq a$;
j not equal to i.

For the rook's case, $W = \sum\sum w_{ij} = 170$.

$$I = \frac{n\sum\sum w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W\sum(x_i - \bar{x})^2} = \frac{50 \cdot 24a^2 \cdot 2}{170 \cdot 18a^2} = 0.784$$

for all choices of a, m.

$$\begin{aligned} Var(I) &= 0.010897 \\ Z(I) &= 7.7088 \text{ whenever } A > B. \\ G &= \frac{\sum\sum w_{ij}x_ix_j}{\sum\sum x_ix_j} = \frac{24A^2+24B^2+24Am+24Bm+74m^2}{2500m^2-9A^2-9B^2-32m^2} \\ &= \frac{170+48c^2}{2450-18c^2} \end{aligned}$$

When $c = 0, A = B = m$, and G is a minimum.

$$\begin{aligned} G_{min} &= 170/2450 = 0.0694. \\ Var(G_{min}) &= 0.0000 \text{ from } (10.7). \end{aligned}$$

When $c = I, A = 2m, B = 0$, and G is a maximum.

$$\begin{aligned} G_{max} &= 218/2432 = 0.0896. \\ Var(G_{max}) &= 0.000011855. \\ Z(G_{max}) &= 5.87 \text{ for any m.} \end{aligned}$$

G depends on the relative absolute magnitudes of the sample values. Note that I is positive for any A and B, while G values approach a maximum when the ratio of A to B or B to A becomes large.

Example 5.

$$\begin{array}{cccccccccc} m & m & m & m & m & m & m & m & m & m \\ m & m & m & m & m & m & m & m & m & m \\ m & m & A & m & m & m & m & B & m & m \\ m & m & m & m & m & m & m & m & m & m \\ m & m & m & m & m & m & m & m & m & m \end{array}$$

$A, B, \bar{x}, n, W$ as in Examples 2 and 4.

$$\begin{aligned} I &= 0, \text{ for any possible } A,\ B,\ or\ m. \\ Z(I) &= 0.1920 \text{ since } E(I) = -1/(n-1), \text{ whenever } A > B. \\ G_{min} &= G_{max} = 0.0694, \text{ for any possible } A,\ B,\ or\ m. \\ Var(G_{min}) &= 0, \text{ but } Var(G_{max}) = 0.00000059. \\ Z(G_{max}) &= 0.0739. \end{aligned}$$

Neither statistic can differentiate between a random pattern and one with little spatial variation. Contributions to $G(d)$ are large only when the product $x_i x_j$ is large, whereas contributions to $I(d)$ are large when $(x_i - m)(x_j - m)$ is large. It should be noted that the distribution is nowhere near normal in this case.

Example 6.

$$\begin{array}{cccccccccc} m & m & m & m & m & m & m & m & m & m \\ m & A & B & A & m & m & B & A & B & m \\ m & B & A & B & m & m & A & B & A & m \\ m & A & B & A & m & m & B & A & B & m \\ m & m & m & m & m & m & m & m & m & m \end{array}$$

$A, B, \bar{x}, n, W$ as in the above examples.

$$\begin{aligned} I &= -0.7843 \\ Var(Z) &= 0.010897 \\ Z(I) &= -7.3177 \\ \text{When A} &= \text{2m and } B = 0, \\ G &= 0.0502 \\ Var(G) &= 0.00001189 \\ Z(G) &= -5.5760 \end{aligned}$$

The juxtaposition of high values next to lows provides the high negative covariance needed for the strong negative spatial autocorrelation $Z(I)$, but it is the multiplicative effect of high values near lows that has the negative effect on $Z(G)$.

Table 10.2 gives some idea of the values of $Z(G)$ and $Z(I)$ under various circumstances. The differences result from each statistics structure. As shown in the examples above, if high values within d of other high values dominate the pattern, then the summation of the products of neighboring values is high, with resulting high positive $Z(G)$ values. If low values within d of low values dominate, then the sum of the product of the xs is low resulting in strong negative $Z(G)$ values. In the Moran's case, both when high values are within d of other high values and low-values are within d of other low values, positive covariance is high, with resulting high $Z(I)$ values.

10.6 General Discussion

Any test for spatial association should use both types of statistics. Sums of products and covariances are two different aspects of pattern. Both reflect the dependence structure in spatial patterns. The $I(d)$ statistic has its peculiar weakness in not being able to discriminate between patterns that have high values dominant within d or low values dominant. Both statistics have difficulty discerning a random pattern from

Table 10.2 Standard normal variates for $G(d)$ and $I(d)$ under varying circumstances for a specified d value

Situation	$Z(G)$	$Z(I)$
HH	$++$	$++$
HM	$+$	$+$
MM	0	0
Random	0	0
HL	$-$	$--$
ML	$-$ #	$-$
LL	$--$	$++$

Key: HH pattern of high values of xs within d of other high x values
M moderate values
L low values
Random no discernible pattern of xs
$++$ strong positive association (high positive Z scores)
$+$ moderate positive association
0 no association
this combination tends to be more negative than HL

one in which there is little deviation from the mean. If a study requires that $I(d)$ or $G(d)$ values be traced over time, there are advantages to using both statistics to explore the processes thought to be responsible for changes in association among regions. If data values increase or decrease at the same rate, that is, if they increase or decrease in proportion to their already existing size, Moran's I changes while $G(d)$ remains the same. On the other hand, if all x values increase or decrease by the same amount, $G(d)$ changes but $I(d)$ remains the same. It must be remembered that $G(d)$ is based on a variable that is positive and has a natural origin. Thus, for example, it is inappropriate to use $G(d)$ to study residuals from regression. Also, for both $I(d)$ and $G(d)$ one must recognize that transformations of the variable X result in different values for the test statistic. As has been mentioned above, conditions may arise when d is so small or large that tests based on the normal approximation are inappropriate.

10.7 Empirical Examples

The following examples of the use of G statistics were selected based on size and type of spatial units, size of the x values, and subject matter. The first is a problem concerning the rate of SIDS by county in North Carolina, and the second is a study of the mean price of housing units sold by zip-code district in the San Diego metropolitan region. In both cases the data arc explained, hypotheses made clear, and $G(d)$ and $I(d)$ values calculated for comparable circumstances.

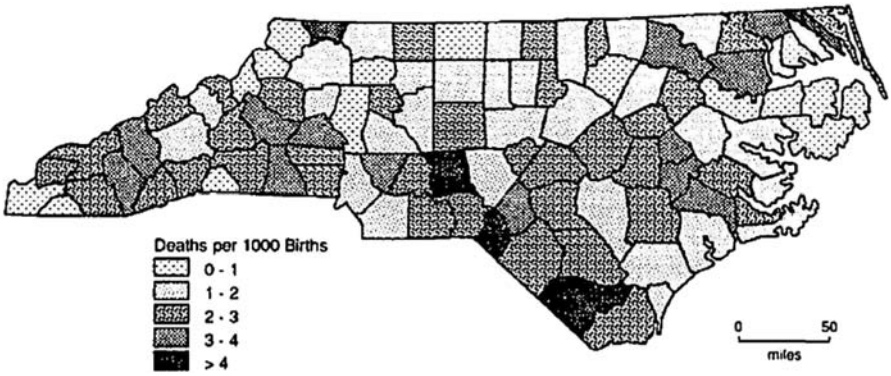


Fig. 10.1 Sudden infant death rates for counties of North Carolina, 1979–1981

10.7.1 Sudden Infant Death Syndrome by County in North Carolina

SIDS is the sudden death of an infant 1-year old or less that is unexpected and inexplicable after a postmortem examination (Cressie and Chan, 1989). The data presented by Cressie and Chan were collected from a variety of sources cited in the article. Among other data, the authors give the number of SIDs by county for the period 1979–1984, the number of births for the same period, and the coordinates of the counties. We use as our data the number of SIDs as a proportion of births multiplied by 1,000 (see Fig. 10.1). Since no viral or other causes have been given for SIDS, one should not expect any spatial association in the data. To some extent, high or low rates may be dependent on the health care infants receive. The rates may correlate with variables such as income or the availability of physicians' services. In this study we shall not expect any spatial association.

Table 10.3 gives the values for the standard normal variate of I and G for various distances.

Results using the G statistic verify the hypothesis that there is no discernible association among counties with regard to SIDS rates. The values of $Z(G)$ are less than one. In addition, there seems to be no smooth pattern of Z values as d increases. The $Z(I)$ results are somewhat contradictory, however. Although none are statistically significant at the 0.05 level, $Z(I)$ values from 30–50 miles, about the distance from the center of each county to the center of its contiguous neighboring counties, are well over one. This represents a tendency toward positive spatial autocorrelation at those distances. Taking the two results together, one should be cautious before concluding that a spatial association exists for SIDS among counties in North Carolina. Perhaps more light can be shed on the issue by using the $G_i(d)$ and $G_i^*(d)$ statistics.

Table 10.3 Spatial association among counties: SIDS rates by county in North Carolina, 1979–1984

d (miles)	$Z(G)$	$Z(I)$
10	0.82	−0.55
20	0.29	0.99
30	−0.12	1.68
33[a]	0.40	1.84
40	−0.04	1.32
50	0.60	1.20
60	−0.36	0.48
70	−0.28	−0.45
80	−0.19	−0.13
90	0.11	−0.19
100	0.30	0.18

[a]At all distances of this length or longer each district is connected to at least one other county

Table 10.4 Highest positive and negative standard normal variates by county for $G_i^*(d)$ and $G_i(d)$: SIDS rates in North Carolina, 1979–1984 ($d = 33$ miles)

County	$ZG_i^*(d)$	County	$ZG_i^*(d)$
	Highest Positive		
Richmond	+ 3.34	Richmond	+ 3.62
Robeson	+ 3.12	Robeson	+ 3.09
Scotland	+ 2.78	Hoke	+ 1.78
Hoke	+ 2.12	Northampton	+ 1.44
Cleveland	+ 1.78	Moore	+ 1.39
	Highest Negative		
Washington	−2.63	Washington	−2.18
Dare	−1.84	Davie	−1.92
Davie	−1.76	Dare	−1.70
Cherokee	−1.55	Bertie	−1.64
Tyrrell	−1.53	Stokes	−1.58

Table 10.4 and Fig. 10.2 give the results of an analysis based on the $G_i(d)$ and $G_i^*(d)$ statistics for a d of 33 miles. This represents the distance to the furthest first-nearest neighbor county of any county.

The $G_i^*(d)$ statistic identifies five of the one hundred counties of North Carolina as significantly positively or negatively associated with their neighboring counties (at the 0.05 level). Four of these, clustered in the central south portion of the state, display values greater than +1.96, while one county, Washington near Albemarle Sound, has a Z value of less than 1.96 (see Fig. 10.2). Taking into account values greater than +1.15 (the 87.5 percentile), it is clear that several small clusters in addition to the main cluster are widely dispersed in the southern part of the state. The main cluster of values less than 1.15 (the 12.5 percentile) is in the eastern part of the state. It is interesting to note that many of the counties in this cluster are in the sparsely populated swamp lands surrounding the Albemarle and Pamlico Sounds. If

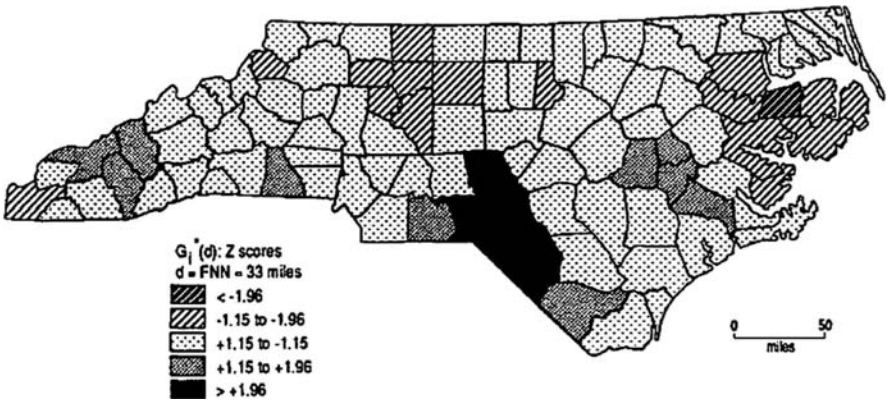


Fig. 10.2 $Z[G_i^*(d = \text{furthest nearest neighbor} = 33\,\text{miles})]$ for SIDS rates of counties of North Carolina, 1979–1984

overall error is fixed at 0.05 and a Bonferroni correction is applied, the cutoff value for each county is raised to about 3.50. However, such a figure is unduly conservative given the small numbers of neighbors.

In this case it becomes clear that an overall measure of association such as $G(d)$ or $I(d)$ can be misleading because it prompts one to dismiss the possibility of significant spatial clustering. The $G_i(d)$ statistics, however, are able to identify the tendency for positive spatial clustering and the location of pockets of high and low spatial association. It remains for the social scientist or epidemiologist to explain the subtle patterns shown in Fig. 10.2.

10.7.2 Dwelling Unit Prices in San Diego County by Zip-Code Area, September 1989

Data published in the Los Angeles Times on October 29, 1989, give the adjusted average price by zip code for all new and old dwelling units sold by builders, real estate agents, and homeowners during the month of September 1989 in San Diego County (see appendix). The data are supplied by TRW Real Estate Information Services. One outlier was identified: Rancho Santa Fe, a wealthy suburb of the city of San Diego, had prices of sold dwelling units that were nearly three times higher than the next highest district (La Jolla). Since neither statistic is robust enough to be only marginally affected by such an observation, Rancho Santa Fe was not considered in the analysis.

Although the city of San Diego has a large and active downtown, San Diego County is not a monocentric region. One would not expect housing prices to trend upward from the city center to the suburbs in a uniform way. One would expect, however, that since the data are for reasonably small sections of the metropolitan area, that there would be distinct spatial autocorrelation tendencies (see Fig. 10.3).

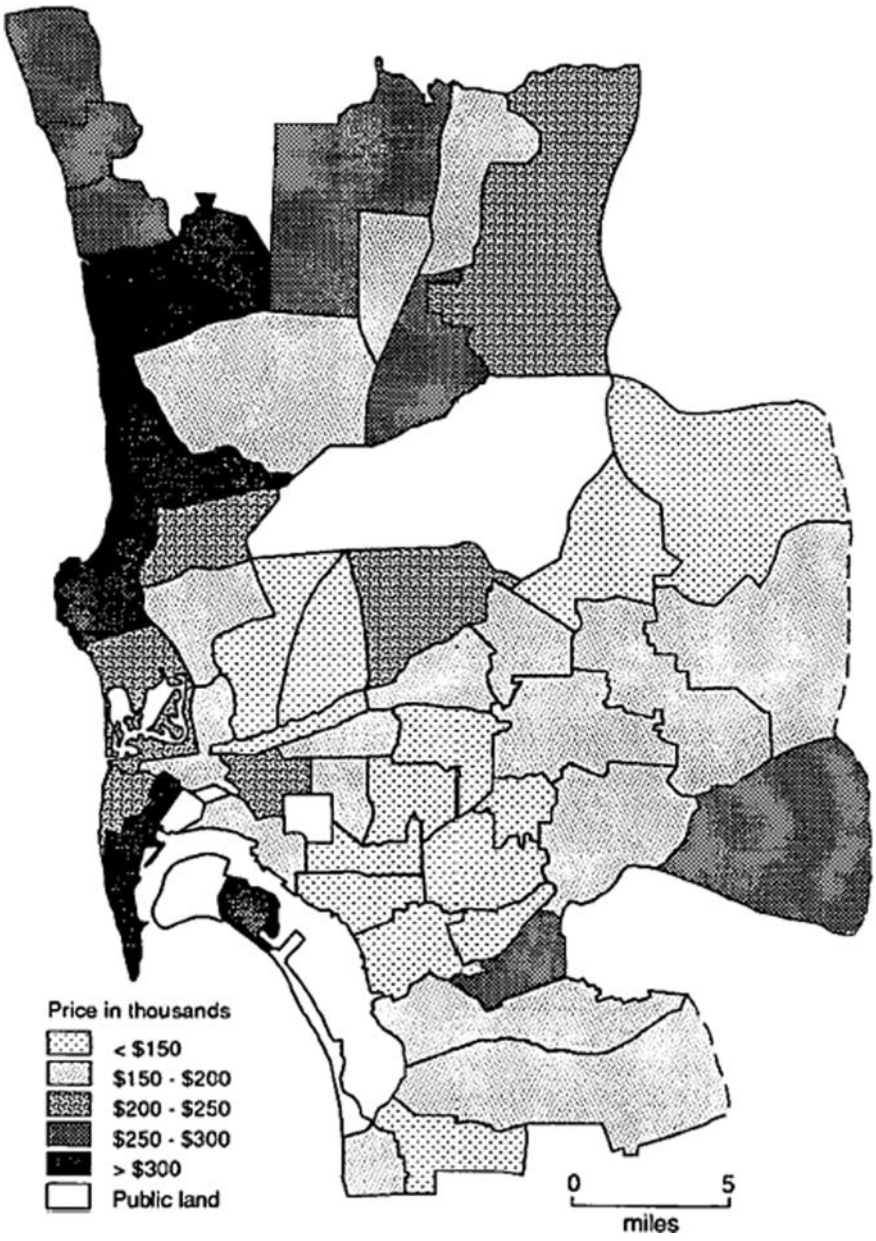


Fig. 10.3 San Diego house prices, September 1989

High positive I values are expected. $G(d)$ values are dependent on the tendencies for high values or low values to group, if the low cost areas dominate, the $G(d)$ value is negative. In this case, $G(d)$ is a refinement of the knowledge gained from I.

Table 10.5 shows that there are strong positive values for $Z(I)$ for distances of 4 miles and greater. $Z(G)$ also shows highly significant values at 4 miles and beyond, but here the association is negative, that is, low values near low values are much more influential than are the high values near high values. Moran's I clearly indicates that there is significant spatial autocorrelation, but, without knowledge of $G(d)$, one might conclude that at this scale of analysis, in general, high income districts are significantly associated with one another.

By looking at the results of the $G_i(d)$ statistics analysis for d equal to five, the individual district pattern is unmistakable. The $Z(G_i^*(5))$ values shown in Table 10.6 and Fig. 10.4 provide evidence that two coastal districts are positively

Table 10.5 Spatial association among zip code districts: dwelling unit prices in San Diego county, September 1989

d (miles)	$Z(G)$	$Z(I)$
2	−0.67	0.33
4	−2.36	2.36
5[a]	−2.32	4.13
6	−2.47	4.16
8	−2.80	3.51
10	−2.66	3.57
12	−2.20	3.53
14	−2.34	3.92
16	−2.54	4.27
18	−2.30	3.57
20	−2.25	2.92

[a] At all distances of this length or longer each district is connected to at least one other district

Table 10.6 Highest positive and negative standard normal variates by zip code district for $G_i^*(d)$ and $G_i(d)$: dwelling unit prices in San Diego county, September 1989 ($d = 5$ miles)

Neighborhood	$ZG_i^*(d)$	Neighborhood	$ZG_i^*(d)$
	Highest positive		
Cardiff	+2.27	Cardiff	+2.08
Solana Beach	+2.02	Solana Beach	+1.81
Point Loma	+1.93	Mini Mesa	+1.56
La Jolla	+1.89	Ocean Beach	+1.37
Del Mar	+1.55	R. Penasquitos	+1.33
	Highest negative		
East San Diego	−3.22	East San Diego	−2.99
East San Diego	−2.74	East San Diego	−2.54
East San Diego	−2.64	North Park	−2.48
North Park	−2.56	East San Diego	−2.48
Mission Valley	−2.38	College	−2.19

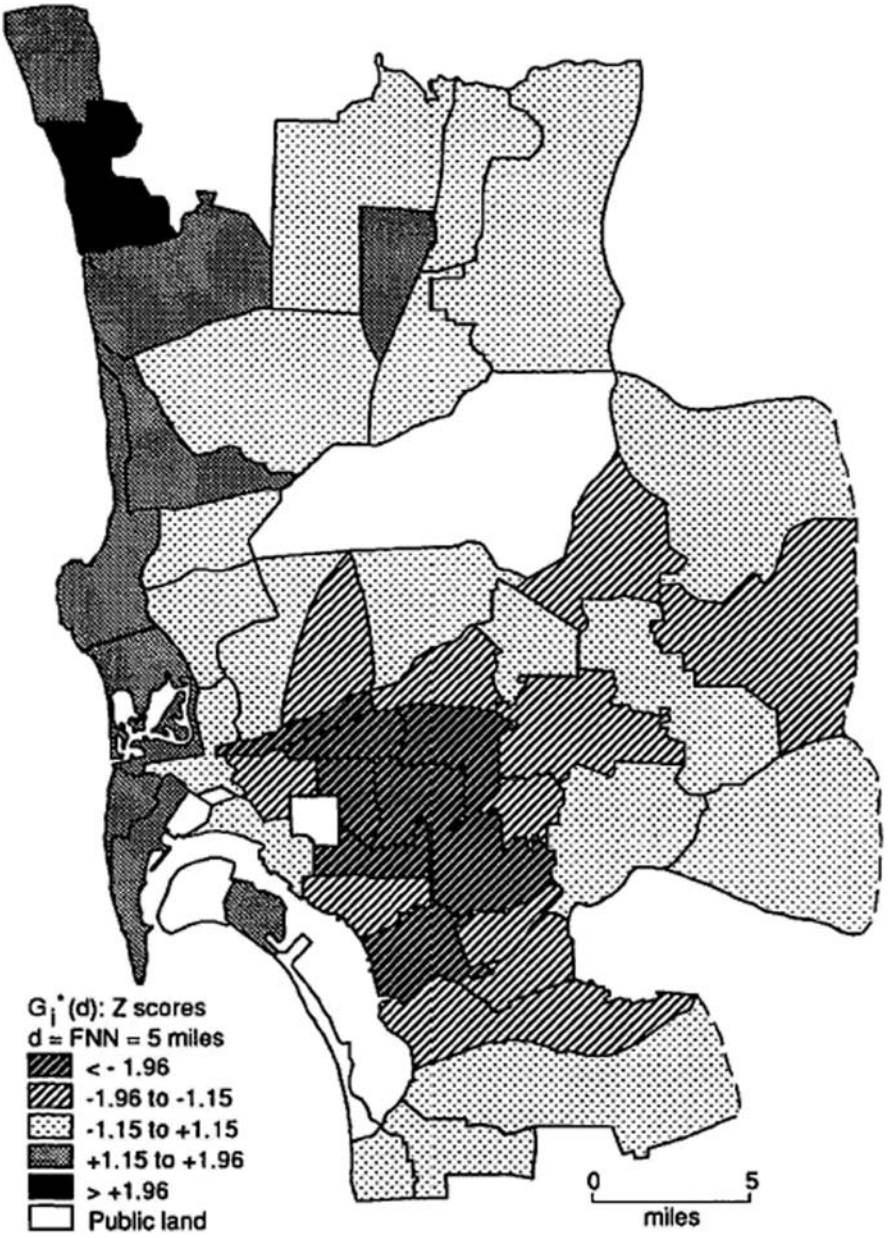


Fig. 10.4 $Z[G_i^*(d = \text{furthest nearest neighbor} = 5\,\text{miles})]$ for house prices of San Diego county zip code districts, September 1989

associated at the 0.05 level of significance while eight central and south central districts are negatively associated at the 0.05 level. There is a strong tendency for the negative values to be higher. It is for this reason that the $Z(G)$ values given above are so decidedly negative. The districts with high values along the coast have fewer near neighbors with similar values than do the central city lower value districts. The cluster of districts with negative $Z(G^*)$ values dominates the pattern. The adjusted Bonferroni cutoff is about 3.27, but again is overly conservative.

10.8 Conclusions

The G statistics provide researchers with a straightforward way to assess the degree of spatial association at various levels of spatial refinement in an entire sample or in relation to a single observation, when used in conjunction with Moran's I or some other measure of spatial autocorrelation, they enable us to deepen our understanding of spatial series. One of the G statistics' useful features, that of neutralizing the spatial distribution of the data points, allows for the development of hypotheses where the pattern of data points will not bias results.

When G statistics are contrasted with Moran's I, it becomes clear that the two statistics measure different things. Fortunately, both statistics are evaluated using normal theory so that a set of standard normal variates taken from tests using each type of statistic are easily compared and evaluated.

Appendix

San Diego county average house prices for September 1989 by zip-code district

	Zip code	Principal neighborhood name	Coordinates (miles) x	y	Price (in thousands)
01	92024	Encinitas	1	39	264
02	92007	Cardiff	2	36	260
03	92075	Solana Beach	3	34	261
04	92014	Del Mar	5	32	309
05	92127	Lake Hodges	10	34	265
06	92129	R. Penasquitos	12	32	194
07	92128	R. Bernardo	15	35	191
08	92064	Poway	17	32	236
09	92131	Scripps Ranch	13	29	270
10	92126	Mira Mesa	8	28	162
11	92037	Lajolla	3	22	398
12	92122	University City	6	23	201
13	92117	Clairemont	6	20	192
14	92109	Beaches	4	18	249

continued

continued

San Diego county average house prices for September 1989 by zip-code district

	Zip code	Principal neighborhood name	Coordinates (miles) x	y	Price (in thousands)
15	92110	Bay Park	6	15	152
16	92111	Kearny Mesa	8	19	138
17	92123	Mission Village	10	19	131
18	92124	Tierrasanta	13	20	221
19	92120	Del Cerro	14	18	187
20	92119	San Carlos	17	19	182
21	92071	Santee	20	22	124
22	92040	Lakeside	23	24	147
23	92021	El Cajon	24	19	151
24	92020	El Cajon	22	17	150
25	92041	La Mesa	18	16	169
26	92115	College	14	16	138
27	92116	Kensington	11	16	192
28	92108	Mission Valley	9	16	89
29	92103	Hillcrest	8	14	225
30	92104	North Park	11	14	152
31	92105	East San Diego	13	14	111
32	92045	Lemon Grove	17	13	137
33	92077	Spring Valley	20	13	150
34	92035	Jamul	24	12	291
35	92002	Bonita	17	8	297
36	92139	Paradise Hills	16	9	117
37	92050	National City	13	8	99
38	92113	Logan Heights	11	10	84
39	92102	East San Diego	12	12	88
40	92101	Downtown	8	12	175
41	92107	Ocean Beach	3	14	229
42	92106	Point Loma	3	12	338
43	92118	Coronado	7	10	374
44	92010	Chula Vista	15	6	165
45	92011	Chula Vista	17	4	184
46	92032	Imperial Beach	11	1	164
47	92154	Otay Mesa	15	2	126
48	92114	East San Diego	15	11	126

Source of Data: Los Angels Times, October 29, 1989, page K15.

Chapter 11
Constructing the Spatial Weights Matrix Using a Local Statistic

Arthur Getis and Jared Aldstadt

This Chapter was originally published in:
Getis, A., Aldstadt, J. (2002) Constructing the Spatial Weights Matrix Using a Local Statistic. Geographical Analysis 34 (2)130-140. Reprinted with permission of Blackwell Publishing, Oxford

Abstract Spatial weights matrices are necessary elements in most regression models where a representation of spatial structure is needed. We construct a spatial weights matrix, $\mathbf{W}$, based on the principle that spatial structure should be considered in a two-part framework, those units that evoke a distance effect, and those that do not. Our two-variable local statistics model (LSM) is based on the G_i^* local statistic. The local statistic concept depends on the designation of a critical distance, d_c, defined as the distance beyond which no discernible increase in clustering of high or low values exists. In a series of simulation experiments LSM is compared to well-known spatial weights matrix specifications – two different contiguity configurations, three different inverse distance formulations, and three semi-variance models. The simulation experiments are carried out on a random spatial pattern and two types of spatial clustering patterns. The LSM performed best according to the Akaike Information Criterion, a spatial autoregressive coefficient evaluation, and Moran's I tests on residuals. The flexibility inherent in the LSM allows for its favorable performance when compared to the rigidity of the global models.

11.1 Introduction

One or more spatial weights matrices are key elements in most regression models where a representation of spatial structure is needed. In this paper we outline and test an approach for constructing a spatial weights matrix, $\mathbf{W}$. Our method is based

A. Getis (✉)
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

J. Aldstadt
Department of Geography, University at Buffalo, Buffalo, NY, USA
e-mail: geojared@buffalo.edu

on the principle that spatial structure should be considered in a two-part framework, those units that reflect a distance effect and those that do not.

We report on the results of a series of simulation experiments on well-known spatial weights matrix specifications – two different contiguity configurations, three different inverse distance formulations, and three semi-variance models. These are compared to a two-variable local statistics model (LSM) which is based on the G_i^* local statistic (Getis and Ord, 1992; Ord and Getis, 1995). The G_i^* statistic is based on the spatial association between observations to a distance d from i. Values of G_i^* are given in standard normal variates. The local statistic concept depends on the designation of a *critical distance*, d_c, defined as the distance beyond which no discernible increase in clustering of high or low values exists. This definition implies that any continuity in spatial association over distance ends at the critical distance. The simulation experiments are carried out on a variety of possible raster spatial distribution patterns including: random and two types of clustering. The appropriateness of the various $\mathbf{W}$ specifications are evaluated by a series of goodness-of-fit regression tests.

11.2 Previous Attempts to Create a Spatial Weights Matrix

The spatial weights matrix is an integral part of spatial modeling. It is defined as the formal expression of spatial dependence between observations (Anselin, 1988). It is curious to note that while most spatial analysts recognize that $\mathbf{W}$ is supposed to be a theoretical conceptualization of the structure of spatial dependence, these same analysts more often than not use in their work a $\mathbf{W}$ which is at best empirically convenient. In many instances, $\mathbf{W}$ has no obvious relationship whatsoever to dependence structure. Thus, models employing such structures are misspecified. This is not to say that analysts have not struggled with the problem of a proper dependence representation in the $\mathbf{W}$ matrix. A bevy of schemes have been created to attempt to fashion the needed theoretical conceptualization. Typical of the well-known schemes are:

1. Spatially contiguous neighbors
2. Inverse distances raised to some power
3. Lengths of shared borders divided by the perimeter
4. Bandwidth as the nth nearest neighbor distance
5. Ranked distances
6. Constrained weights for an observation equal to some constant
7. All centroids within distance d
8. n Nearest neighbors, and so on

Some of the newer schemes are:

1. Bandwidth distance decay (Fotheringham et al., 1996)
2. Gaussian distance decline (LeSage, 2004)
3. "Tri-cube" distance decline function (McMillen and McDonald, 2004)

Another approach, in the spirit of Kooijman (see below), that by Griffith (1996), is designed to find a $\mathbf{W}$ that "extracts" or filters the spatial effects from the data $\mathbf{y}$. A

comparable approach by Getis (Getis, 1995b; Getis and Griffith, 2002) is designed to find that part of a variable that is spatially autocorrelated. It may be that one of the above choices leads to good, parsimonious results but the pall of misspecification hanging over the chosen model may still remain. In this paper, we propose a form for $\mathbf{W}$ that is based on the distance beyond which there is a specified change in the nature of spatial association. In the next section, we discuss the nature of $\mathbf{W}$. This is followed by a description of our model in Sect. 11.4. In Sect. 11.5, we demonstrate its operation using simulated data representing typical patterns in a 30-by-30 raster setting. The results are compared to many different $\mathbf{W}$ specifications in Sect. 11.6. Finally, in Sect. 11.7, we summarize our results and consider future strategies.

11.3 On the Nature of W

As early as the 1960s, researchers such as Dacey (1965) were aware that by calculating join-count statistics for the purpose of identifying spatial autocorrelation, results would vary with one's definition of a neighbor. Using raster data, popular were rook's case and queen's case definitions of neighbors. When data are in a vector structure, models of $\mathbf{W}$ usually were constructed in the form of contiguity matrices, that is, matrices that take as neighbors those regions having a side in common. Contiguous neighbors are elements of $\mathbf{W}$ equal to one while all other elements are given the value 0. Oftentimes, the contiguity $\mathbf{W}$ matrix is row-standardized. By definition, the ith observation is not considered a neighbor of itself.

Research on $\mathbf{W}$ has been reviewed by Griffith (1996, p. 80), who concludes that five rules of thumb aid in the specification of weights matrices:

1. "It is better to posit some reasonable geographic weights matrix than to assume independence." This implies that one should search for or theorize about an appropriate $\mathbf{W}$ and that better results are obtained when distance is taken into account.
2. "It is best to use surface partitioning that falls somewhere between a regular square and a regular hexagonal tessellation." Griffith suggests that for planar data, a specification between four and six neighbors is better than something either above six or below four. Of course, the configuration of the planar tessellations will play a role here (Boots and Tiefelsdorf, 2000).
3. "A relatively large number of spatial units should be employed, $n > 60$." Following from the law of large numbers, most spatial research, especially due to unequal size spatial units, would require fairly large samples.
4. "Low-order spatial models should be given preference over higher-order ones." Following from the scientific principle of parsimony, it is always wise to choose less complicated models when the opportunity presents itself.
5. "In general, it is better to apply a somewhat under-specified (fewer neighbors) rather than an over-specified (extra neighbors) weights matrix." Florax and Rey (1995) found this result by identifying the power of tests. Overspecification reduces power. They recognize that "Uncertainty with respect to proper specification has long been recognized as a fundamental problem in applied spatial econometric modeling" (p. 132).

Kooijman (1976) proposed to choose $\mathbf{W}$ in order to maximize Moran's coefficient. Reinforcing this view is Openshaw (1977) who selected that configuration of $\mathbf{W}$ which results in the optimal performance of the spatial model. Boots and Dufournaud (1994) create a binary contiguity/noncontiguity matrix by means of a linear programming technique that maximizes and minimizes spatial autocorrelation. We subscribe to these approaches with one major caveat, that is, that the proposed spatial structure be theoretically defensible. Bartels (1979) agrees that these approaches would be better justified if appropriate tests could be constructed to assure that dependency structure is taken into account. He concludes, however, that since such tests are unavailable, binary $\mathbf{W}$ is defensible. The Hammersley–Clifford (Bennett, 1979) approach to spatial Markov models allows for near neighbor properties of $\mathbf{W}$, but special assumptions of the local Markov conditions must be invoked. In our view, a realistic spatial dependency structure should not be sacrificed for mathematical convenience.

In recent research, Tiefelsdorf et al. (1999) caution that a row-standardized $\mathbf{W}$ gives too much weight to observations with few spatial links, like those on the edge of the study region. Conversely, they point out that a globally standardized $\mathbf{W}$ places too much emphasis on observations with a large number of spatial links. Most researchers, however, have found that row-standardization is helpful in two ways: weighting observations (but not spatial links) equally and interpreting autoregressive parameters and Moran statistics. With regard to an autoregressive spatial process, Tiefelsdorf (2000, pp. 43–45) provides a formal interpretation of the role of the spatial autocorrelation coefficient.

In a recent study by Florax and de Graaff (2004), it is suggested that an indicator be used to evaluate whether a $\mathbf{W}$ is misspecified because of matrix sparseness (proportion of a matrix that is zeroes). This suggestion corresponds to a path we have chosen for our work.

It is in the nature of the variables being adjusted for spatial effects that is the key to an appropriate $\mathbf{W}$. Variables showing a good deal of local spatial heterogeneity at the scale of analysis chosen would probably be more appropriately modeled by few links in $\mathbf{W}$, while a homogeneous or spatial trending variable would better be modeled by a $\mathbf{W}$ with many links. This reasoning is borne out by the concept of the *range* in geostatistics. Since W is defined as a model of spatial dependence, it would seem plausible to include in the matrix the complete and, as far as possible, accurate representation of the dependence structure of the variable(s) in question. This implies that the scale characteristics of data are crucial elements in the creation of $\mathbf{W}$. As spatial units become large, spatial dependence between units tends to fall. In an intrinsically stationary setting, larger units tend to have values of variables close to the mean for the region as a whole.

11.4 The Local Statistics Model

For the *local statistics model* (LSM), we take advantage of the G_i^* local statistic (Ord and Getis, 1995). A positive G_i^* indicates that there is clustering of high values around i; a negative number indicates low values. These G_i^* values are scrutinized

cumulatively, rather than by distance bands, around each observation as distance increases from it. When these values fail to rise *absolutely* with distance, the cluster diameter is reached, implying that any continuity in spatial association or dependence over distance ends at that distance. We have called this the critical distance, d_c. This is an empirically derived value. No statistical test is associated with it. The individual cell values of $\mathbf{W}$ are determined by the following:

$$
\begin{aligned}
&\text{When } d_c > d_{NN1},\\
&\qquad w_{ij} = \begin{cases} \dfrac{|G_i^*(d_c) - G_i^*(d_{ij})|}{|G_i^*(d_c) - G_i^*(0)|}, & \text{for all } j \text{ where } d_{ij} \le d_c;\\ 0, & \text{otherwise.} \end{cases}\\
&\text{When } d_c = d_{NN1},\\
&\qquad w_{ij} = \begin{cases} 1, & \text{for all } j \text{ where } d_{ij} = d_c;\\ 0, & \text{otherwise.} \end{cases}\\
&\text{When } d_c = 0,\\
&\qquad w_{ij} = 0, \quad \text{for all } j,
\end{aligned}
\tag{11.1}
$$

where d_{NN1} is the first nearest neighbor distance for observation i. $G_i^*(d_c)$ is the G_i^* score at the critical distance, and $G_i^*(0)$ is the G_i^* score for the ith observation only. Thus, $G_i^*(0)$ represents a base from which other measures of G_i^* are compared.

This procedure is based on positive association between nearby values, whether or not the values themselves are low or high. The result is that all values in $\mathbf{W}$ are greater than or equal to 0. The variable under study, $\mathbf{y}$, is not restricted to a natural origin nor to any particular measurement scale.

Equation (11.1) shows that each weight is a function of the trend in G_i^* as distance increases from i. From this, it is clear that spatial correlation is 0 at and beyond d_c. The correlation values are entered into the appropriate cell of the $\mathbf{W}$ matrix. As is true of other models, we enter a zero in the ii cells. On the other hand, if d_c is 0, using this reasoning, a zero would be placed in the appropriate row and column of the $\mathbf{W}$ matrix. Zero rows and columns in $\mathbf{W}$, without compensation for those of the N observations so affected, destroys any statistical interpretation of $\mathbf{y}$. This problem leads to our *local statistics model*:

$$\mathbf{y} = \alpha + \rho\mathbf{W}\mathbf{y} + \beta\mathbf{x} + \epsilon \tag{11.2}$$

In this setup, it is conceivable for rows of $\mathbf{W}$ to be completely filled with zeroes indicating that there is no autocorrelation surrounding an observation. To compensate for the zero-row effect, we create a dummy variable, $\mathbf{x}$, that takes on the value one for all observations having no dependence structure and zero otherwise. Thus, (11.2), has two spatial parameters, ρ and β, where each parameter equates the effect of a different aspect of the spatial structure: ρ represents the dependence structure of the variable $\mathbf{y}$, while β equates the effect on $\mathbf{y}$ of those observations that are not correlated with any of their neighbors (the nondependence structure). It is conceivable

for the $\mathbf{x}$ vector to contain all zeroes, although this is not likely in practice. In this special case, we would not have the $\beta\mathbf{x}$ term in (11.2). The parameters are estimated using maximum-likelihood techniques. This formulation is not limited to a univariate approach. As in spatial lag models, one could have regressor variables in addition to the dummy variable of (11.2). Technically, there is the question of matrix singularity. In our approach, the matrix $(\mathbf{I} - \rho\mathbf{W})$ is invertible and thus fulfills the non-singularity requirement of a spatial autoregressive equation.

11.5 Experiments with LSM

11.5.1 Data Sets

We artificially created three types of 30-by-30 raster data sets (900 observations). Each type is simulated 25 times for 75 experiments. The data sets represent a wide variety of spatial patterns. Their construction is described in Table 11.1. The first type, a *random normal*, represents a situation in which there is complete spatial independence among the values placed in cells. The second type displays a pattern of *two clusters*, and the third type is made up of *six clusters*. All patterns contain as their data standard normal deviates. The 50 cluster patterns represent a wide variety of spatial structures usually found in research based on georeferenced variables. The LSM is designed to be used as a $\mathbf{W}$ specification for any model where clustered data obtains. Figure 11.1 shows one realization of the random normal pattern type, Fig. 11.2 displays one realization of the two cluster pattern type, and Fig. 11.3

Table 11.1 Data set descriptions

Data set	Description
Random	Random placement of values sampled from the normal distribution with mean 0, and standard deviation 1; 25 simulations
Two-clusters	1 cluster of high values at (10, 10) with radius 8 and 1 cluster of low values at (20, 20) with radius 8 – values from the normal distribution with mean 0, and standard deviation 1; 25 simulations. The highest values from the random generation were placed randomly in the high value cluster, while the lowest were placed randomly in the cluster of low values. The remaining values, those in the middle of the distribution, were placed randomly outside the clusters
Six-clusters	6 randomly placed clusters, 3 of high values and 3 of low values with radii 2, 4, and 6 respectively; values are sampled from the normal distribution with mean 0, and standard deviation 1; 25 simulations. As in the two-cluster case the highest values were placed randomly, but this time in the three high value clusters. The low values were placed randomly in the three low value clusters. The remaining middle values were placed randomly outside the clusters

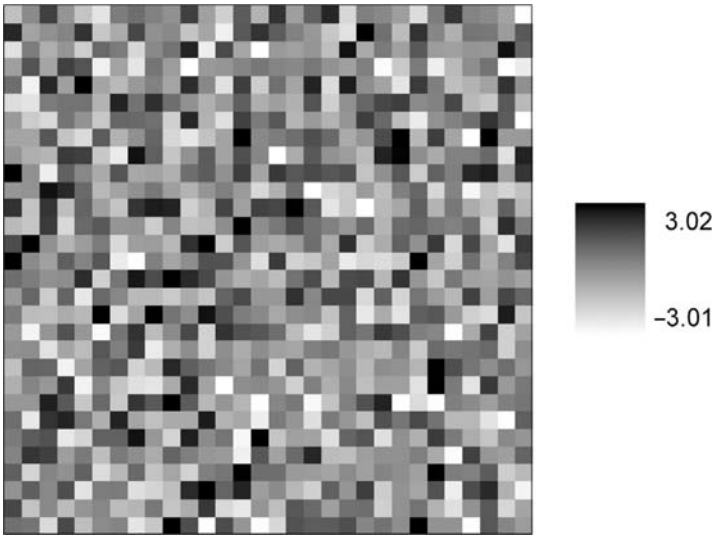


Fig. 11.1 Random data set. Shading values are in random normal deviates

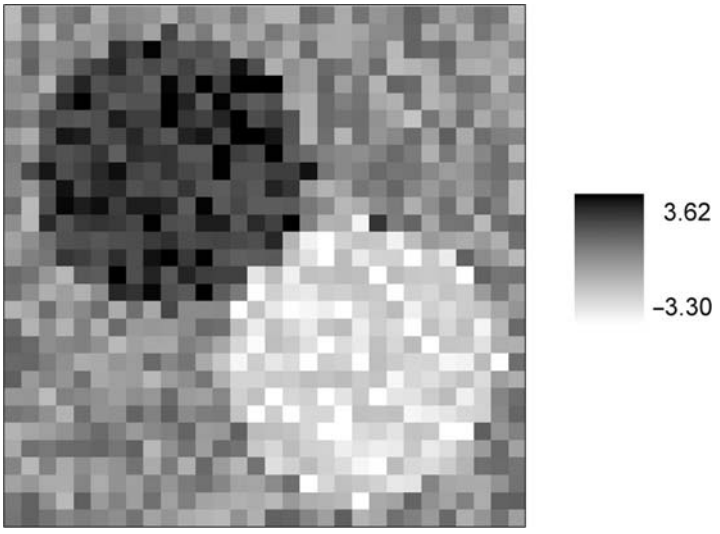


Fig. 11.2 Two cluster data set. Shading values are in random normal deviates

displays one realization of the six cluster pattern. Figure 11.4 shows the spatial distribution of the critical distances for the data sets shown in Figs. 11.1, 11.2, and 11.3. Note that the longest d_c are within the clusters. This is indicative of the spatial extent of the autocorrelation.

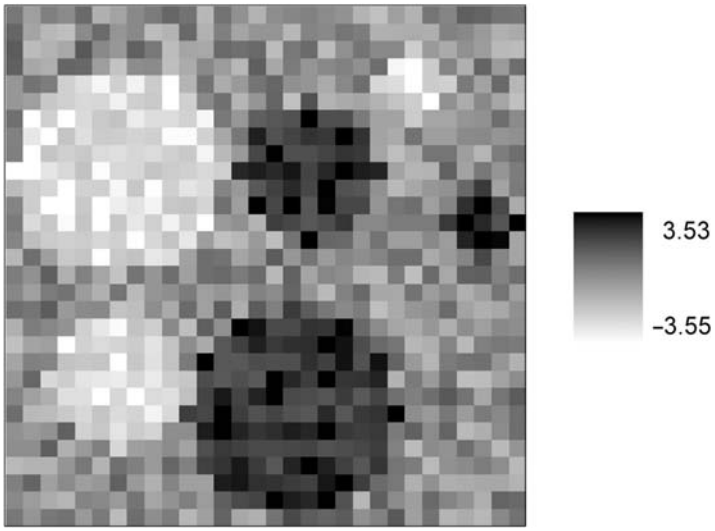


Fig. 11.3 Six cluster data set. Shading values are in random normal deviates

11.5.2 Forms of W

Of the eight different experimental forms of **W** to which the LSM is to be compared, the first five are called *geometric* and the final three *geostatistical*. By geometric we mean that the matrices are mainly a function of the configuration of cells and/or the distances separating them. The final three **W** can be compared more directly with LSM since their form is a function of the values within the cells and thus are empirically derived, as is LSM. These are the geostatistics models described below (in the section titled "Geostatistical **W**"). In all cases the **W** are row-standardized.

Geometric W

1. Rook Contiguity
   The four neighbors of each cell in the cardinal directions are given the value 1, all others 0. This is the most popular formulation of **W**.
2. Queen Contiguity
   The eight neighbors of each cell in all directions are given the value 1, all others 0.
3. Inverse Distance $(1/d)$
   Taking the distance between near neighbors as 1, reciprocals of all pairs of distances are calculated and entered into **W**.
4. Inverse Distance $(1/d^2)$
   Same as in 3, except that distances are squared. This formulation of **W** is probably the most popular of all the distance-based **W**.

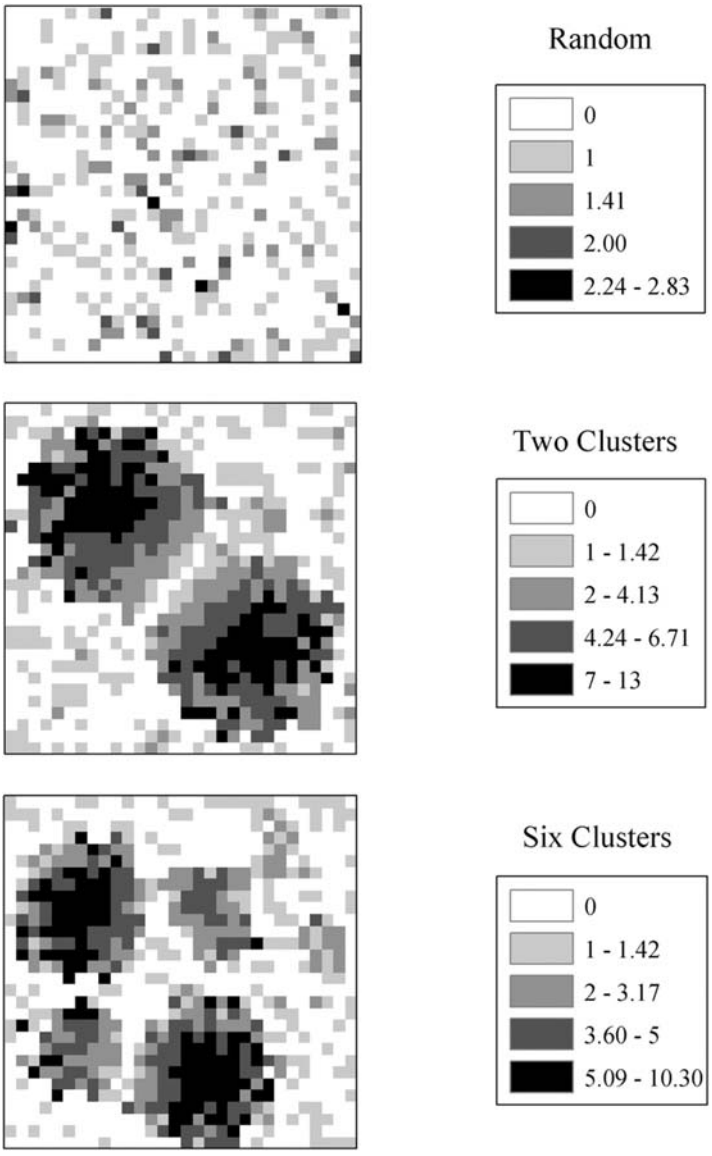


Fig. 11.4 d_c's calculated for data sets in Figs. 11.1, 11.2, and 11.3. Distances are based on one unit separating centers of rook's case neighbors

5. Inverse Distance $(1/d^5)$
   This $\mathbf{W}$, is similar to the two previous ones, except in this case the emphasis is on the near neighbors of each cell. The higher the exponent, the greater the influence of neighboring cells as opposed to greater distance between cells. This formulation, in many respects, is comparable to contiguity $\mathbf{W}$.

Geostatistical W

The variogram, $\gamma(d)$, describes the distance characteristics of a set of georeferenced values. If the assumption of intrinsic stationarity holds, the function that describes the distance characteristics of the data set is consistent over the entire set of data at all distances d from all sites. The variogram is a *global* function. We may estimate the variogram by

$$\gamma(d) = \frac{1}{2N} \sum \{y(\mathbf{u}) - y(\mathbf{v})\}^2, \tag{11.3}$$

where the sum is taken over all locations at distance d apart, and N denotes the number of pairs in the distance band which d represents. Given the assumption of intrinsic stationarity, the variance is σ^2 and the autocorrelation is

$$\rho(d) = 1 - \frac{\gamma(d)}{\sigma^2}. \tag{11.4}$$

There are many variogram models. The ones chosen are often the best fit to an empirical distribution of $\gamma(d)$. The form of the distance relationship of observed values usually defines the models. The empirical distribution of the variogram is fit by a curve representing the theoretical variogram. From our selected variogram model we create $\mathbf{W}$ by placing values in the cells representing the degree of correlation be tween each uv pair of observations. These values will vary between one and zero according to (11.4). For distances beyond the range, the correlation is zero. In our work we selected three popular semivariogram models. Where appropriate, for each data set we create a $\mathbf{W}$ based on the distance separating pairs of points. It is inappropriate to use variogram models where required assumptions do not hold. For the three models that we used, only the two clustering data sets can be considered as intrinsically stationary:

1. Spherical variogram

$$\gamma(d) = \begin{cases} \sigma^2 \left(\dfrac{3d}{2r} - \dfrac{d^3}{2r^3} \right), & \text{for } d < r, \\ \sigma^2, & \text{otherwise,} \end{cases} \tag{11.5}$$

   where r is the range (when $\gamma(d) = \sigma^2$).
2. Gaussian variogram

$$\gamma(d) = \sigma^2(1 - e^{-3d^2/r^2}). \tag{11.6}$$

3. Exponential variogram

$$\gamma(d) = \sigma^2(1 - e^{-3d/r}). \tag{11.7}$$

11.6 Results

Tables 11.2–11.4 give the results for each set of simulations. We chose as our criteria for evaluation: the Akiake Information Criterion (AIC), the autocorrelation coefficient, and a measure (Moran's I) of the residuals of regressions where the weight matrix is the independent variable. All of the comparisons are based on a spatial lag model. Of course, the geometric and geostatistical examples do not contain the $\beta\mathbf{x}$ term of (11.2).

11.6.1 Evaluation Criteria

1. AIC

The AIC uses the likelihood function in conjunction with the number of independent variables (unknown parameters) to discriminate between models. The lower the AIC value, the better the fit. This measure was chosen for two reasons. First, it is based on the likelihood function and corresponds to other goodness of fit measures, such as the Schwarz criterion. Second, it is heavily influenced by the number of independent variables, penalizing formulations with more independent variables than those with fewer independent variables. For the LSM approach, two independent variables are needed to describe the spatial structure. None of the other approaches requires more than one variable. Thus, the AIC provides us with a goodness-of-fit test that is particularly demanding for the LSM approach.

2. The Autocorrelation Coefficient ρ

The autocorrelation coefficient gives an interpretation for the possible association between $\mathbf{Wy}$ and $\mathbf{y}$. For example, if $\rho = 1$ the implication is that $\mathbf{y}$ can be described by $\mathbf{Wy}$, meaning that $\mathbf{W}$ does a good job of expressing the spatial relationships embedded in $\mathbf{y}$. On the other hand, a ρ value near 0 implies that $\mathbf{W}$ has little to do with the spatial structure of $\mathbf{y}$.

3. Residuals

If the $\mathbf{W}$ matrix completely accounts for all of the variation in $\mathbf{y}$, the residuals of a regression having $\mathbf{Wy}$ as the independent variable will be spatially random. In our experiments, we use Moran's I as our measure of spatial pattern. Moran's I is computed using the same $\mathbf{W}$ matrix that is used to estimate the corresponding model.

11.6.2 The Tests

1. AIC

Table 11.2 shows the AIC values for the six-clusters, two-clusters, and random cases. The mean AIC values are considerably lower for LSM as opposed to the geometric and the geostatistics models for the cluster cases. Even though the AIC values

show greater variation for the LSM model, the highest AIC for LSM is lower than the mean for all other models and cluster tests (16 tests). The geostatistics models perform better than do the geometric models. The worst fit is for the inverse distance model among the six-clusters and two-clusters tests. The random data-pattern type gives evidence that the LSM model reflects whatever clustering that might be present in a random pattern of data. Some might argue that the LSM value of w_{ij} will indicate clustering at least some of the time in a random pattern. Given that all of the other matrices hover around the AIC = 1,930 level for the random patterns, the fact that LSM has an AIC value of 1,841 indicates that some clustering exists within these patterns. Also, it might be well to think of 1,930 as a base level on which to evaluate all other results since an AIC of 1,930 represents unequivocal randomness. If this is the case, then the mean AIC value of 718 for LSM in the two-clusters cases and 936 in the six-clusters cases represents a 63% and 52%, respectively, improvement over a null model (means used as predictors). Note that no AIC values are calculated for the geostatistical models for the random patterns. These models are not defined on randomness; thus it is inappropriate to use them in this regard. All of these results give strong evidence for the strength and efficacy of an LSM model.

2. The Autocorrelation Coefficient ρ

Table 11.3 clearly shows the strength of LSM as representing a truly autocorrelated model. Although several of the other models are highly spatially autocorrelated, none reach the level of the LSM model. A curious result is noted in the simulations for the random patterns for the LSM. One might ask how there can be autocorrelation in a random pattern. Actually, there is considerable local spatial autocorrelation in such patterns. The LSM model picks up the positive correlation between near values that are high or low and trending in a high or low direction.

3. Residuals

After applying the various $\mathbf{W}$ models, residuals were subjected to a test using Moran's I so as to identify any remaining autocorrelation in the pattern (see Table 11.4). In all cases, the mean value of the standard normal variate of Moran's statistic $[Z(I)]$ should be close to 0, and the spread should be normal around this mean. In both of the clustering cases the LSM model outperforms the other models. Note that the positive residuals nicely balance the negative residuals in the six-clusters cases, and they are well within a normal curve. In the two-clusters cases the balance is not in evidence, but the Z values and the standard deviation would make it difficult to reject the existence of normally distributed residuals. Therefore, again the LSM model outperforms all of the others.

The extremely high values for $Z(I)$ for some of the models indicates that their $\mathbf{W}$ matrices do a poor job of describing the cluster patterns. It is interesting to note that as the power of d increases from 1 to 5 in the distance decay models, they appear to perform better. Higher powers give greater weight to near neighbors than to those further away. Note how the rook's case model and the $1/d^5$ model give similar results. The negative $Z(I)$ values result from the nature of the clusters

Table 11.2 AIC results

Data set		LSM	$Rook$	$Queen$	$1/d$	$1/d^2$	$1/d^5$	$Spherical$	Gauss	Exponential
6-Clusters	Mean	935.81	1,299.65	1,179.37	1,454.21	1,222.77	1,224.69	1,146.77	1,136.77	1,137.94
$N = 25$	Max	1,038.83	1,375.19	1,237.78	1,557.30	1,301.11	1,290.30	1,206.56	1,196.23	1,197.16
	Min	787.37	1,187.88	1,072.01	1,342.09	1,153.60	1,114.91	1,047.18	1,043.89	1,045.83
	SD	65.68	54.83	47.82	53.46	41.53	52.43	47.72	45.39	45.90
2-Clusters	Mean	717.64	1,094.30	985.83	1,132.62	949.41	1,013.00	904.94	921.98	913.21
$N = 25$	Max	853.55	1,168.33	1,066.05	1,220.17	1,028.09	1,088.37	976.50	991.31	984.80
	Min	620.48	1,002.08	909.12	1,044.45	888.62	929.50	849.85	870.20	860.02
	SD	50.87	42.85	43.43	45.38	42.11	42.38	41.26	39.93	40.26
Random	Mean	1,841.36	1,930.39	1,930.04	1,929.78	1,929.94	1,930.33			
$N = 25$	Max	1,899.72	1,998.73	1,996.50	1,998.46	1,998.17	1,998.64			
	Min	1,769.62	1,867.02	1,867.25	1,867.13	1,867.25	1,867.09			
	SD	37.49	34.95	34.55	34.30	34.34	34.84			

Table 11.3 Estimated autocorrelation coefficient values

Data set		LSM	Rook	Queen	$1/d$	$1/d^2$	$1/d^5$	Spherical	Gauss	Exponential
6-Clusters	Mean	1.00	0.75	0.84	0.99	0.99	0.82	0.89	0.91	0.91
$N = 25$	Max	1.00	0.79	0.86	0.99	0.99	0.84	0.91	0.93	0.93
	Min	1.00	0.72	0.83	0.99	0.99	0.80	0.87	0.89	0.89
	SD	0.00	0.02	0.01	0.00	0.00	0.01	0.01	0.01	0.01
2-Clusters	Mean	1.00	0.80	0.86	0.99	0.99	0.85	0.96	0.97	0.97
$N = 25$	Max	1.00	0.81	0.88	1.00	0.99	0.86	0.97	0.98	0.98
	Min	1.00	0.78	0.85	0.99	0.99	0.83	0.94	0.96	0.96
	SD	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01
Random	Mean	0.80	0.01	−0.02	−0.18	−0.04	0.00			
N25	Max	0.86	0.08	0.12	0.40	0.21	0.09			
	Min	0.75	−0.11	−0.18	−0.86	−0.42	−0.15			
	SD	0.03	0.04	0.06	0.32	0.15	0.05			

Table 11.4 Moran's Z(I) of residuals

Data set		LSM	Rook	Queen	$1/d$	$1/d^2$	$1/d^5$	Spherical	Gauss	Exponential
6-Clusters	Mean	−0.115	−6.733	−3.292	22.950	10.077	−5.160	−1.738	−0.668	−0.745
$N = 25$	Max	1.937	−5.970	−2.642	28.782	13.483	−4.129	−0.489	0.676	0.954
	Min	−1.633	−7.757	−4.153	18.584	7.833	−6.333	−2.580	−1.321	-1.635
	SD	0.849	0.474	0.403	2.830	1.647	0.501	0.519	0.557	0.649
2-Clusters	Mean	1.233	−7.340	−4.517	25.854	7.551	−5.801	1.281	2.960	3.004
$N = 25$	Max	2.709	−6.577	−3.961	27.898	9.551	−5.065	2.611	4.521	4.350
	Min	0.176	−8.414	−5.340	23.275	5.824	−6.800	−0.012	1.640	1.706
	SD	0.619	0.456	0.391	1.168	0.779	0.444	0.738	0.860	0.747
Random	Mean	0.398	0.044	0.057	0.302	0.124	0.050			
N25	Max	0.596	0.079	0.090	0.418	0.200	0.093			
	Min	0.224	−0.003	0.011	0.091	0.010	0.015			
	SD	0.081	0.017	0.018	0.077	0.041	0.017			

themselves. Recall, the clusters were constructed as random spatial distributions of high (low) values. Thus, the negative values represent the negative autocorrelation characteristic within the patterns.

11.7 Interpretation and Conclusions

We highlight some of the characteristics of the tested models in light of the evaluation criteria. As mentioned above, the LSM performed best according to the AIC, ρ, and residuals criteria. In general, the geostatistics models were next with good scores on all three criteria. Of the geostatistics models, the Gaussian appeared to be slightly more evocative of the data than the other two.

This may be a function of the greater complexity, and thus better descriptive characteristics, of this model than of the other two. In quality, the queen's contiguity formulation, with its eight neighbors, appeared to be next, but further behind the LSM and the geostatistics models. The rigidity of the queen's case robs it of the flexibility inherent in the LSM and the geostatistics models. As expected, the rook's case is among the least effective, again because of its inflexibility and because only four neighbors for each cell are brought to bear on $\mathbf{W}$. The distance decline functions, surprisingly, do poorly, about as effective as the rook's case with regard to the AIC and residuals criteria, but $1/d$ and $1/d^2$ respond well as measures of autocorrelation (the ρ criterion). Interestingly, $1/d^5$, a model that puts a great deal of emphasis on near neighbors performs similarly to the rook model.

The spatial structure represented by LSM is made up of two parts, those observations that reflect a distance effect and those that do not. This is a distinct strength of the LSM. Apparently, the heterogeneity embodied in most spatial distributions can be effectively captured by this two variable approach.

More of the observed spatial structure is embodied within the LSM formulation than in the other models. It must be remembered, however, that the LSM is empirically based, and any explanation of the usefulness of its structure should allude to the fact that what is being modeled are the spatial relations within the already existing data. Any theoretical notion about its form should be defended by a discussion concerning not only its cluster structure but by the model's dummy variable that represents no apparent spatial dependence between nearby cells.

One might argue that the comparative success of the LSM over the geostatistical and geometric models is unfair. LSM is locally adaptive; that is, it is based on a series of local measures giving it great flexibility. The geostatistical and geometric models are global measures based on a limited set of parameters. This has implications for the use of the AIC as a measure of fit, implying that the LSM has an advantage because of its greater number of what could be called degrees of freedom.[1] Our view is that since the LSM model outperforms the others, and that the

[1] This point was made to us in correspondence by Michael Tiefelsdorf, the editor of this article.

others are the "usual" models used in spatial autoregressive research, it is helpful to know that with a locally based model, much better results obtain. We suggest that this type of empirical approach be used as a substitute for the rigidity of the global models.

Further work in this area should be directed at lowering the AIC scores. That is, in the LSM case, we use a particular definition of clustering. Simulations might indicate that a somewhat different definition gives us a better model fit. In addition, other local statistics were not applied on the supposition that the fundamental additive quality of the G_i^* measure best represents the clustering inherent in the spatial association between nearby units while the others represent other attributes of patterns such as covariance and difference. In further work, where we theorize differently about the form of spatial autocorrelation, we will use other local statistics for the creation of $\mathbf{W}$. In addition, note that in Fig. 11.4, the d_c tends to be high near the edges of clusters. This implies that the d_c is sensitive to the values of cells contained in the clusters. Currently, we are preparing a procedure that takes cluster boundaries into account. Finally, the spatial filtering work mentioned earlier appears to represent another promising approach to the problem of $\mathbf{W}$ specification.

Acknowledgements The authors greatly appreciate the comments of Michael Tiefelsdorf and three anonymous reviewers. The paper has been considerably strengthened due to their suggestions.

Chapter 12
Spatial Autocorrelation: A Statistician's Reflections*

J. Keith Ord

Abstract Improvements in both technology and statistical understanding have led to considerable advances in spatial model building over the past 40 years, yet major challenges remain both in model specification and in ensuring that the underlying statistical assumptions are validated. The basic concept in such modeling efforts is that of spatial dependence, often made operational by some measure of spatial autocorrelation. Such measures depend upon the specification or estimation of a set of weights that describe spatial relationships. We examine how the identification of weights has evolved and briefly describe recent developments.

After a brief examination of some of the key assumptions commonly made in spatial modeling, we consider the selection of tests of spatial dependence and their application to irregular sub-regions. We then move on to a consideration of local tests and estimation procedures and identify ways in which local procedures may be useful, particularly for large data sets. We conclude with a brief review of a recently developed method for modeling anisotropic spatial processes.

12.1 Introduction

In 1966, notwithstanding my academic status as ABD, I took a faculty position with the Economics Department at the University of Bristol, England. Although I would not recommend such a step nowadays, the academic world was a kinder, gentler place 40 years ago and the step proved singularly worthwhile. In addition to interactions with my new-found colleagues in the Economics Department, I met up with Andrew Cliff, who was working with Peter Haggett. My previous exposure to geography had been limited to pathetic attempts to draw maps, combined with futile

*This chapter is based upon a presentation made at the annual conference of the American Association of Geographers in Chicago in March 2006.

J. K. Ord
Georgetown University Washington, DC, USA
e-mail: ordk@georgetown.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*, Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_12,

attempts to remember the names of rivers and capital cities, so the whole notion of a quantitative approach was completely novel and most intriguing. The Bristol Geography Department attracted a number of prominent visitors from across the Atlantic and Art Getis was among the first. Art's primary interest at that time lay in spatial point processes, which duly led to his widely acclaimed monograph (Getis and Boots, 1978). At that time, my knowledge of point processes and their underlying distributions was essentially non-spatial, so our several discussions focused primarily on statistical distributions for counts data (the topic of my by-then completed dissertation). We interacted periodically in the intervening years, but we did not engage in any joint research projects until the beginning of the nineties when Art contacted me about local measures of spatial autocorrelation that he had developed. We consider that topic in Sect. 12.4 of this chapter, but in order to place the discussion in context, we first travel back in time to Bristol in the second half of the 1960s.[1]

Andrew Cliff was working on statistical measures of spatial dependence. As time passed we began to work on these issues together, which led to our first joint paper (Cliff and Ord, 1969), presented at the annual conference of the British Section of the Regional Science Association in London. We shared our session with [now Sir and Nobel Laureate] Clive Granger who made a number of very perceptive comments about spatial analysis and the underlying assumptions. We return to his observations, which remain highly relevant today, in Sect. 12.1.2 below.

The basic question we addressed in that paper was how to test for spatial autocorrelation in possibly irregular spatial configurations. An intrinsic element in our approach was the specification of a weighting matrix, whose form could be quite general but was used to specify the type of dependence we might reasonably expect to see in spatial data if, indeed, the observations were not randomly assigned across the spatial units (or areas). Once the general concept was formulated, the question arose of how to determine the weights. Indeed, over the years, the refrain "how do you choose the weights?" has been heard at many a conference session. Accordingly, it seemed appropriate in this retrospective study to explore some of the answers to that question that have emerged over the years. The question is likely to continue to stimulate further research for some time to come.

At the outset, we should delineate the scope of the chapter. The focus is exclusively upon purely spatial processes and, even within that framework we restrict attention to models that are appropriate for areal units within an overall (study region). Thus, we refer to the units for which data are recorded as areas within the region unless specific applications lead elsewhere (e.g., cells in a regular grid). In particular, we do not consider observations located at points within a continuum, for which kriging methods are more appropriate (cf. Cressie, 1993, Chap. 3).

At the time that Andrew and I began our work only binary weights describing physical adjacencies had been used, but we decided to generalize the notion and

[1] It is sometimes said that if you remember the 1960s, you weren't there. I believe my recollections to be reasonably accurate but I hope the reader will forgive any transgressions.

to specify weights that depended upon both the length of the common boundary between spatial units and the distances between their geographical centroids. The determination of these quantities is now straightforward in an era of digitized maps, but back then it involved Andrew in many hours of work poring over maps of Eire and making the requisite measurements! We re-examine the issue of choosing the weights and the form of the test statistic in Sect. 12.2. The application of spatial methods to irregularly shaped regions is then briefly discussed in Sect. 12.3. As computing power has increased, local tests have become feasible and we consider some of the issues underlying their application in Sect. 12.4.

Increased computing power has also led to a shift in emphasis from the specification of the weights for testing purposes towards model specification and estimation. Indeed, tests are rarely an end in themselves but they are a useful tool for model checking and development. Thus, tests for spatial autocorrelation are useful in testing regression residuals, to determine whether or not a model has captured the essence of the data. In turn, a spatial model will often incorporate spatial lags, wherein specification of the weighting matrix is even more critical than for testing purposes. Conversely, if an initial test does not indicate any particular patterns of spatial dependence, a spatial model is unlikely to provide much in the way of insight. These thoughts lead naturally to issues of model specification and estimation. Again, improvements in computing power and better computational algorithms have led to the development of local models, which we consider in Sect. 12.5.

A feature of local models is that they allow for directional symmetry, but more descriptively than in a formal inferential framework. Thus, in Sect. 12.6, we return to global models, but review some recent work by Deng (2008) that allows for asymmetric relationships. The chapter concludes with a few brief comments about future developments.

12.1.1 The Laws of Geography

Tobler (1970) referred to "the first law of geography: everything is related to everything else, but near things are more related than distant things." This notion underpins the emphasis upon models that rely upon physical adjacency and in turn provides a basis for specifying the weights in hypothesis testing. The space in question need not be purely physical but may depend upon other attributes. For example, the cities of Detroit and Buffalo may be perceived as more similar than the physically closer cities of Frederick and Annapolis (see a map of Maryland for details). Whether we consider a physical or a more general space, model development may proceed from this basic law and it will often provide guidance for model specification.

Before we get too carried away with model development, we should keep in mind George Box's first law of statistics: all models are wrong but some are useful. We might adapt Box's law to generate the second law of geography:

All maps are wrong but some are useful.

Examples are not hard to find. The schematic maps of subway systems are extremely useful to riders, yet such representations clearly fail any examination based upon distance-based criteria. The famous map created by Charles Joseph Minard, which describes Napoleon's 1812 Moscow campaign (cf. Tufte, 2001) is an even more graphic example of the need to decide upon the metric to be used before judging the utility of a map. We will focus upon quantitative measures but should never lose sight of the power of graphical displays.

12.1.2 Granger's Comments

In his 1969 paper, Clive Granger raised a number of issues that should be a standard lexicon for anyone considering an analysis of spatial data:

- *Is the process isotropic?* Can we ignore direction in formulating a spatial model? Does a suburban community have the same impact upon the city center as the city does upon the suburb? The answer is that directional invariance may well hold in some physical situations but is unlikely in economic contexts. When we have data for multiple time periods, this asymmetry can be addressed; see, for example the spatial econometric framework developed by Anselin (1988, and later work) and a number of the papers in Anselin et al. (2004a). For purely spatial data we have hitherto been forced to assume a symmetric relationship; we return to this question in Sect. 12.6.
- *Is the process spatially stationary?* Is it only the relative distance between two locations that matters? Granger is doubtful and reasonably so. He considers time series data at different locations and uses such data to examine spatial dependence via the time series spectrum. In general, it may be reasonable to assert spatial stationarity as a basis for testing the residuals from a model, but such a property is unlikely for original processes of interest.
- *Did we observe the sample or the population?* This question plagues both non-spatial and spatial econometrics. Typically, we assume some kind of hyper-population and make inferences on that basis. But to what entity are the inferences to be made? If we are looking at spatial data through time, it is reasonable to make forecasts for future activity, based upon the usual time series assumptions. When the data are purely spatial, the nature of the inferences needs careful consideration. If indeed we have observed the population (of all milch cows in Eire or any other phenomenon) our inferences may be restricted to statements about patterns, based upon permutations tests rather than some more far-reaching framework.
- *Did we observe one population or many populations?* For simple random sampling the notion of repeated drawings from an urn (a single population) may be reasonable. Once we leave behind the notion of independent and identically distributed observations, we must consider models that allow for drawings from distinct populations, or some suitable joint distribution. The framework for inference needs careful consideration.

Although these comments were made nearly 40 years ago, they remain as challenges to the aspiring spatial modeler, and should be considered prior to launching into any investigation. We use these questions as part of the framework for our ruminations.

12.2 Tests for Spatial Dependence

We start with tests for spatial dependence and a consideration of tests for data on a regular grid. Such tests are important in their own right given the widespread availability of imaging data, and also serve to simplify the discussion as the interactions among cells are easier to define.

The two classical tests use patterns of joins known as the *rook's case* (horizontal and vertical linkages on the grid) and the *queen's case* (rook's case plus diagonal links). In each case, all links are usually given equal weight, although edge and corner cells may be treated differently; for example, the row sums of weights may be scaled to equal sums, thereby giving greater weight to links between edge and corner cells with other cells. Given a grid with n cells ($n = RC$, where R and C refer to the numbers of rows and columns respectively) the total number of edge and corner cells is of order $n^{\frac{1}{2}}$ so that edge effects have much more impact upon the distributions of test statistics than in time series, where there are only two endpoints. Pinske (2004) provides a comprehensive discussion of the conditions under which Moran-type tests are asymptotically normally distributed, so we do not pursue that topic further in this chapter.

Florax and de Graaff (2004) performed a meta-analysis of the many simulation studies relating to the performance of the more popular tests for spatial dependence. Among their findings are two items relating to weighting matrices:

1. Increased density (e.g., a higher proportion of ones in a binary matrix) has a negative impact upon the power of a test.
2. Higher connectedness (e.g., average number of links per cell) has a positive effect upon power.

As the authors note, these results are somewhat unexpected and warrant further investigation. Folk-lore on this topic suggests relative sparseness of the weights is a benefit, consistent with (1), but this perspective has probably developed mostly from analyses with small numbers of areas. The scale of the data generating process is clearly also important. An interesting question is whether any analytical studies can complement these empirical findings and we now consider this question. As we are looking for qualitative insights, we will focus upon asymptotic effects and ignore edge effects (e.g., a regular grid could be mapped onto a torus).

Consider the possible hypotheses of spatial dependence on a regular grid. Three obvious possibilities come to mind:

1. The rook's case
2. The queen's case
3. Isotropic dependence (i.e., direction invariant but distance-dependent)

True Pattern				Test pattern		
a	1-a	a		b	1-b	b
1-a	X	1-a		1-b	X	1-b
a	1-a	a		b	1-b	b

Fig. 12.1 True and test pattern for an interior cell on a regular grid

The third pattern comes closest to our usual expectations, yet in testing we typically assume the first or second case. Figure 12.1 illustrates the situation. If the true pattern corresponds to the rook's case, we have the left-hand pattern in Fig. 12.1 with $a = 0$. Likewise, when the correct version is the queen's case, we would have $a = 1/2$. If the dependence decays with the square of the distance between cell midpoints, the value would be $a = 1/3$ since the squared diagonal distance is twice that of the horizontal and vertical distances. Other choices of decay rate clearly produce other values for a. However, the discussion in Florax and de Graaff (2004) suggests that there is quite a strong argument for a choice that is intermediate between the rook's and queen's cases, at least for smaller study regions.

If we now turn to the right-hand panel of Fig. 12.1, we can postulate various test patterns, corresponding to choices of b, and see how these perform relative to the "true" patterns given in the left-hand panel. We may consider the ARE (Asymptotic Relative Efficiency) of the various tests, following the procedure laid out in Cliff and Ord (1981, pp. 163–170) but originally explored in Cliff and Ord (1973, Chap. 7). If test T is the most powerful test available for a specified pattern, but some other (inefficient) test U is employed with $ARE = 100A_0$ where $A_0 < 1$ we say that U would need n/A_0 observations to be as locally powerful as T with n observations. The ARE is an asymptotic comparison of local alternatives, but typically provides a reasonable benchmark for test comparisons. In the present case, the ARE for test U (based upon an assumed pattern with weights determined by b) relative to the best test T (with weights specified using a) is

$$ARE = \frac{100[(1-a)(1-b)+ab]^2}{[(1-a)^2+a^2][(1-b)^2+b^2]}. \tag{12.1}$$

Table 12.1 Asymptotic relative efficiencies for different patterns of weights

		True pattern	
		Rook	Queen
Test pattern	Rook	100	50
	Queen	50	100
	Best ($b = 0.293$)	85	85
	$b = 1/3$	80	90

Thus, we may take each of the three cases listed above as the true pattern and evaluate the test performance when one of the three patterns is used to define the test statistic. The results are given in Table 12.1.

The results are quite striking. If we mis-specify the pattern as the rook's case, when it is really the queen's case, the ARE drops to 50%. The same applies when the roles are reversed. By contrast, if we put $b = 0.293$ we are guaranteed an ARE of at least 85% for any pattern at or between the rook's and queen's cases. Such a value for b may be rather unappealing but the value $b = 1/3$, which corresponds to the "quadratic-isotropic" case discussed earlier, is almost as good with an ARE of 80% or more. Further, this pattern has the intuitive appeal that it corresponds to combining the (binary) weights for the rook's and queen's cases. Accordingly, we recommend that this test, which we designate the RQ test, be used in preference to either of the standard procedures. The details for the RQ test are given in the appendix.

More generally, the calculation of the ARE offers a quick guide to test selection and is very easy to calculate even for completely general weights. Let W represent the true weighting matrix (used in test T) and W_2 be the corresponding matrix for test U. The ARE for T relative to U is

$$ARE = \frac{tr(W^T W + W^2)tr(W_2^T W_2 + W_2^2)}{[tr(W^T W_2 + W_2 W)]^2}. \tag{12.2}$$

This ARE cannot be less than 1.0 (or 100%).

12.3 Irregular Regions and Spatial Stationarity

Irregularly shaped regions may arise by administrative fiat (e.g., city boundaries), as natural features (e.g., zones of vegetation) or as essentially random locations (e.g., oil drillings). In the third case, spatial stationarity may be a viable assumption, but in the first two it seems inherently unlikely. In principle, if an underlying spatial process can be specified, at least in terms of the mean, variance and covariance structures, we could generate the random variables for each region by aggregation (cf. Granger, 1969, p. 14). In practice, as Granger observed, spatial stationarity is implausible for economic variables and the aggregation process is intractable.

If we think of a region as being a large number of small equal-sized grid cells packed into the irregular space, we can make limited progress for testing purposes. Under the null hypothesis of spatial independence, each cell might be assumed to have area δA and to follow a distribution with mean $\mu\delta A$ and variance $\sigma^2\delta A$. The overall region (with area A) then has a distribution with mean μA and variance $\sigma^2 A$. A test for independence might reasonably proceed on the assumption that the autocorrelation is local, so that weights are proportional to the length of common boundary. Once the model has been specified, tests should take account of the implied heteroscedasticity. For example, we might proceed using appropriately adjusted variables, such as

$$z_j = \frac{x_j - mA_j}{s\sqrt{A_j}}$$

for region j, where (m, s^2) denote the weighted sample mean and variance:

$$m = \frac{\sum x_i}{\sum A_i} \quad \text{and} \quad s^2 = \sum \frac{(x_j - mA_i)^2}{A_i}.$$

Kelejian and Robinson (2004) explore this issue in depth.

12.4 Local Statistics

As time has progressed, two related factors have served to shift the emphasis in much of spatial modeling. First, technological developments have led to huge spatial data sets, such as those obtained from medical or satellite imaging. Second, computational speeds have increased so that analyses are now possible that were previously infeasible. In turn, these developments have produced more efficient numerical methods (discussed very briefly in the next section) and an emphasis on more local analyses, such as a search for "hot spots." It was in this context that Art contacted me regarding a local statistic he had developed to identify local activity. This was the G-statistic, which we now define.

The G-statistic (Getis and Ord, 1992; Ord and Getis, 1995, 2001): The value for site k is

$$G_k = \sum_{j \in N(k)} w_{jk} z_j, \tag{12.3}$$

where $N(k)$ denotes the set of "neighbors" for site k, in the sense that all members of the set have non-zero weights assigned. Thus, the set may include site k itself, when the notation G^* is used. G^* is a more natural statistic to use in the search for hot-spots, whereas G is better for looking for local spatial dependence (e.g., among regression residuals). The original definition in Getis and Ord (1992) used a scaling factor for ease of interpretation and assumed non-negative observations, but the formal test is the same when version (12.3) is used. We refer to G below, but the comments apply in similar fashion to G^*. The discussion in Ord and

Getis (2001) covers tests for local spatial dependence within a background of global spatial autocorrelation, but the basic ideas are the same.

The crucial question that Art raised related to the level of significance to use in testing these local coefficients. The problem relates to the fact that we are carrying out a large number of tests (equal to the sample size). If the objective is to control the overall probability of a Type I error (i.e., concluding that spatial dependence exists when there is none at all) the Bonferroni limit suggested in Getis and Ord (1992) is appropriate. However, if we are operating in a more exploratory mode, this approach is too extreme. A reasonable alternative is to use a fairly stringent conventional level (such as $\alpha = 0.01$) and to combine this with the recognition that n tests will produce an expected number of $n\alpha$ rejections just by chance. Better yet, a normal probability plot of the z-scores can reveal where the true exceptions arise.

An alternative to these statistics is the local indicator developed by Anselin (1995):

The Local Indicator of Spatial Association – LISA (Anselin, 1995)

$$I_k = z_k \sum_{j \in N(k)} w_{jk} z_j. \tag{12.4}$$

The LISA statistic looks similar to G, but includes the extra term, z_k which clearly changes the results. The LISA statistics possess the property that $\sum_k I_k$ yields the global Moran statistic and so they are often referred to as *local Moran statistics*. It is instructive to compare the performance of the two measures in a qualitative fashion; Table 12.2 illustrates the nature of the behavioral differences between the two statistics.

When the value of z_k is close to zero, the G-statistic will signal extreme neighbors, whereas LISA will not. However, LISA is able to distinguish cases where z_k is similar to its neighbors from those where it is strongly in the opposite direction. Thus, the two measures are not competitors but are useful local measures that identify different patterns in the data. A reasonable analogy is the distinction between the *influence measure* and *Cook's D* in regression analysis.

Table 12.2 Relative magnitudes of the G and LISA coefficients: major differences are in bold and moderate differences are in italics

Value of z_k	Sum of neighbors	G_k	I_k
Large & positive	Large & positive	Large & positive	Large & positive
Large & positive	Small	Small	Moderate
Large & positive	Large & negative	Large & negative	Large & negative
Large & negative	Large & positive	**Large & positive**	**Large & negative**
Large & negative	Small	Small	Moderate
Large & negative	Large & negative	**Large & negative**	**Large & positive**
Small	Large & positive	*Large & positive*	*Moderate*
Small	Small	Small	Small
Small	Large & negative	*Large & negative*	*Moderate*

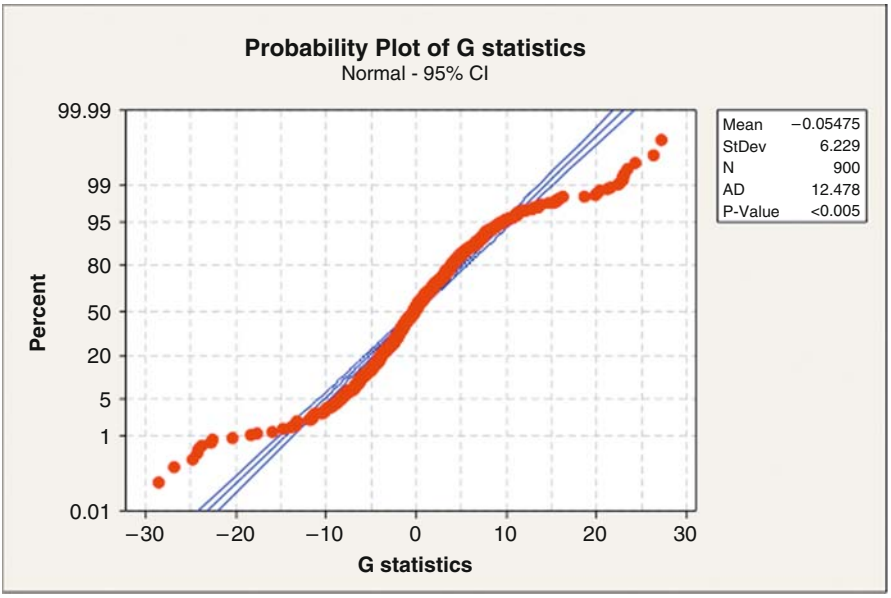


Fig. 12.2 Normal probability plot for G statistics with two clusters of extreme values (Minitab plot)

In the spirit of a more exploratory approach, we may use the local coefficients in plots of various kinds. Anselin (1996) introduced the concept of *Moran scatter plots*, which consist of plotting I_k against z_k. Interpretations of the plot can be determined from the entries in Table 12.2. An alternate plot and set of interpretations are readily generated by plotting G_k against z_k. Further plots may be obtained using probability plots for each statistic (Fig. 12.2). Interactive graphics packages allow identification of the sites associated with specific points making pattern recognition more straightforward.

By way of example, we generated sample G-statistics corresponding to a 30×30 grid. The observations were randomly drawn from a standard normal distribution, save that two well-separated 5×5 blocks were identified, one receiving an additional +2 in each cell and the other -2. The G statistics were computed using the RQ format discussed in Sect. 12.2. The overly heavy tails of the empirical distribution are clearly apparent and the individual extreme points fall within the artificial clusters. A real data set would not produce such clean results, but the general idea is evident. Similar plots and insights may be obtained from the LISA statistic. In either case, it is also feasible to generate plots for subsets of the data, on either the same or separate charts. An interesting application of both local and global test statistics is provided by Trevelyan et al. (2005), who examine the spatial spread of an epidemic over time.

12.5 Local Estimation

Model building for spatial processes is even trickier than hypothesis testing! Whereas an incorrect choice of weights causes a loss of power when testing, the wrong set of weights leads to inconsistent estimators in (auto)regression modeling. Notwithstanding the more serious consequences, we may recall Box's rule in Sect. 12.1.1 and expect to obtain useful results if we are able to formulate a plausible set of weights. In turn, this raises the question of whether we can estimate the weights, subject to reasonable constraints to produce identifiable results.

12.5.1 Estimation with Pre-specified Weights

A standard model for spatial interaction is the joint dependence scheme with independent normally distributed errors:

$$y_i = \alpha + \beta \sum_j w_{ij} y_j + \varepsilon_i; \quad \varepsilon_i \sim IIN(0, \sigma^2), \quad i = 1, \ldots, n. \tag{12.5}$$

The errors do not have to be identically distributed, but that assumption is commonly made. The log-likelihood function depends upon the determinant $|I - \beta W|$, where the matrix $W = \{w_{ij}\}$ contains the set of weights specified in the model in (12.5), and it may be written as

$$\begin{aligned} l(\alpha, \beta, \sigma | y) = \text{const} - n \ln(\sigma) + \ln |I - \beta W| - \frac{1}{2}(y - \alpha 1)^{'}(I - \beta W^{'}) \\ \times (I - \beta W)(y - \alpha 1). \end{aligned} \tag{12.6}$$

Here 1 and y are vectors. Maximum likelihood estimation requires repeated evaluation of the determinant, which is of order n. Ord (1975) provided a numerical procedure for evaluating this determinant using its eigenvalues. Since that time highly efficient numerical techniques have evolved for dealing with extremely large regular lattices, see for example Griffith (2000, and related work). Anselin et al. (2004b) summarize recent computational developments.

These methods are useful for fitting global models, but do not provide insights into more local variations. Also, they assume a pre-specified weighting matrix. Accordingly, we now explore local approaches to estimation that enable us to relax these constraints. First, we stay with pre-specified weights but consider local estimation. For example, when the weights are non-zero only for relatively near neighbors, it becomes possible to partition the complete study area into K non-overlapping sub-regions that are statistically unrelated, as illustrated in Fig. 12.3.

Each black-colored sub-region may be considered independently, conditionally upon the white-colored areas. Each sub-region may be evaluated using its own log-likelihood, which may be written as $l_r(\alpha_r, \beta_r, \sigma_r | y_r)$ for each of the K subsets.

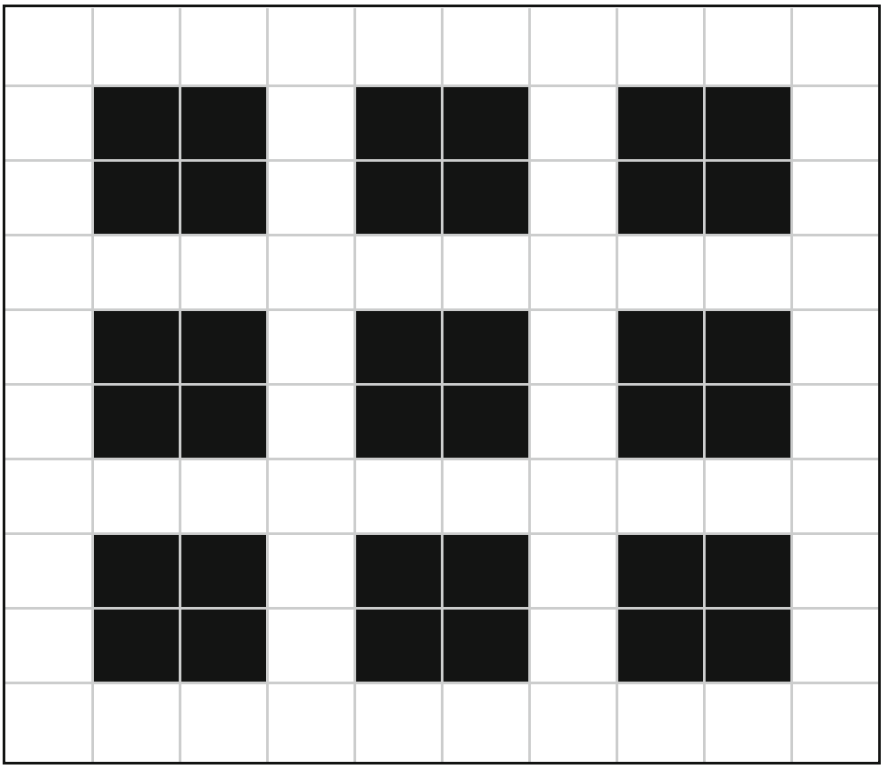


Fig. 12.3 Example of a partitioned study area

Standard procedures may be used to estimate the parameters for each sub-region with correspondingly much smaller determinants. Alternatively, we may combine the K log-likelihoods to produce overall estimators, retaining the advantage of the smaller determinants but with some loss of efficiency. The simplest form of these estimators, with black and white sub-regions each consisting of one cell were introduced by Besag (1977c).

Given the numerical advances noted earlier, it may be asked why we should go down this road? If we retain the single spatial parameter model given in (12.5), there is indeed no real benefit. However, this path opens the way to consideration of multi-parameter schemes for which existing numerical recipes are inadequate for very large data sets. For example, the smaller scale of the individual sub-regions would allow much more general models to be fitted, perhaps combined with the computational power provided by Markov Chain Monte Carlo; see Smith and Roberts (1993) for an overview.

In addition to providing separate estimates for each sub-region, we may also form likelihood ratio tests to determine whether all the regions have the same parametric structure. The overall test of

$$H_0 : (\alpha_r, \beta_r, \sigma_r) = (\alpha, \beta, \sigma)\ \forall r,\ \text{vs. } H_A : (\alpha_r, \beta_r, \sigma_r) \neq (\alpha, \beta, \sigma) \text{ for some } r \tag{12.7}$$

may be written in the general form (after substituting in the ML estimates for each parameter):

$$\chi^2 = 2 \sum_r \{l_r(\alpha_r, \beta_r, \sigma_r) - l_r(\alpha, \beta, \sigma)\}. \tag{12.8}$$

The asymptotic distribution of the test statistic in (12.8) is chi-square with $3(K-1)$ degrees of freedom. Further, the individual terms in the test statistic in (12.8) will be independent and identically distributed under H_0, each being close to chi-square with three degrees of freedom. Thus, unusual sub-regions may be identified from chi-square probability plots. At a more heuristic level, the individual coefficients may be used to develop approximate standardized scores for each of the sets of slope, autocorrelation and variance estimates, and these values used to explore differences among regions.

A key question is how large should the sub-regions be? Ultimately this question will need to be answered empirically and will depend upon the size of the study region. However, for satellite image data, blocks of 20×20 or larger would not seem unreasonable. Further extensions are clearly possible, such as the use of a checkerboard pattern of sub-regions and the evaluation of black sub-regions conditioned on the white sub-regions and vice-versa.

Another possibility is the "rolling" selection of sub-regions, such as taking columns 1–20 then 11–30, 21–40 and so on. That approach leads naturally to the more general framework of locally weighted maximum likelihood estimation, where separate estimates are obtained for each area in the study region. Details are provided by McMillen and McDonald (2004) and LeSage (2004).

12.5.2 Direct Estimation of the Weights

When we move from estimation for fixed (sub)sets of the data to rolling selections or local weighted schemes for individual cells, we also change the inferential framework. In the first case we can make the usual kinds of inference from the likelihood function. However, in the second case we may implicitly use many more parameters than we have observations, so formal inference is infeasible without imposing a considerable number of constraints upon the estimators. This comment is not intended as a criticism, but only as an observation. We need to be clear about the purposes of the analysis. If we are interested in making formal inferences, we must stay within a framework for which a likelihood function can be specified. If we are interested in description or hypothesis generation, the local methods are invaluable.

The first approach to direct estimation of the weights was due to Kooijman (1976), who estimated the weighting matrix by maximizing the value of the global Moran statistics. The optimization is most readily achieved using linear programming, when non-negativity constraints and other conditions (such as assigning a maximum amount of weight to each unit) may be imposed. Several other approaches

have been suggested over the years; see Getis and Aldstadt (2004) for a review. In turn, these authors develop a local estimation procedure for the weights using the G^* statistic. Their simulation studies illustrate the effectiveness of this approach in identifying local patterns.

12.6 Directional Dependence for Purely Spatial Models

By their nature, purely spatial models typically do not allow for directional dependence. For example, we may be well aware of population movements from a city center to its suburbs, but hitherto there has been no way to incorporate such effects into a purely spatial model like (12.5). When appropriate data are available, space–time models provide one way out of this dilemma, since we can incorporate time lags on the right-hand side, such as

$$y_i(t) = \alpha + \beta \sum_j w_{ij} y_j(t-1) + \varepsilon_i(t); \quad \varepsilon_i(t) \sim IIN(0, \sigma^2), \quad i = 1, \ldots, n. \tag{12.9}$$

Regular regression methods are then available if we absorb the slope coefficient into the weights and relax the non-negativity and summation conditions on the weights. Even if those conditions are retained, linear or quadratic programming procedures enable the models to be fitted.

Turning back to purely spatial models, Deng (2008) has developed an ingenious way of allowing for directional dependence. He formulates an *anisotropic spatial lag model* as

$$y_i = \alpha + \sum_j f(g_i|\theta) w_{ij} y_j + x_i'\beta + \varepsilon_i; \quad \varepsilon_i \sim IIN(0, \sigma^2), \quad i = 1, \ldots, n \tag{12.10}$$

Model (12.5) has been extended to include regression effects (β is now a vector) and a general anisotropic function $f(g|\theta)$; the weights may be specified in the usual way. The regression terms are a standard extension and could have been included earlier; they are useful at this stage for expository purposes. In particular, the function $f(g|\theta)$ may be represented as linear in variables g and parameters θ so that familiar estimation procedures are available. The key element in Deng's model is that the new functional form can allow for directional dependence. This framework is best discussed in the context of an example, taken from Deng's paper.

Deng examines the Boston housing prices data set originally presented in Harrison and Rubinfeld (1978) and considered in a spatial context by Pace and Gilley (1997). Housing prices are adversely affected by high crime rates and the direct effects may be captured through standard regression terms, as in (12.10). However, if area A has a high crime rate and adjacent area B has a low crime rate, a reasonable hypothesis is that area A will have a negative impact on B's price level that is stronger than any positive impact that B's low crime rate has on A's prices. To test for such directional effects, Deng considers dummy variables such as

$$D_{qij} = \begin{cases} 1, & \text{if } x_{qj} > x_{qi}; \\ 0, & \text{otherwise.} \end{cases} \tag{12.11}$$

In the current example, q denotes the regression variable of interest (crime rate) and the indicator would take the value 1 for high crime area A in the regression equation for B. Conversely, low crime area B would have indicator value 0 in the equation for area A. The detailed analysis in Deng (2008) shows this approach to be an effective way of formulating and then testing for directional dependence. Deng's work is an important conceptual breakthrough and paves the way for further research on anisotropic models.

12.7 Conclusions and Directions for Further Research

Statistical model building for spatial processes needs to keep in mind the three integrated elements shown in Fig. 12.4. Going counter-clockwise, existing maps can inform the process of data collection which in turn lead to model specification. Once a theoretical model has been developed, we may proceed clockwise, collecting the data to test the hypotheses and then summarizing the results in a map. Substantive studies will typically involve iterative developments or multiple circuits around the loop and the use of maps as diagnostic devices as well as final summaries.

Improved techniques for preliminary data analysis, both non-spatial and spatial methods, provide greater insights into the complexity of spatial processes, and these tools should be used both for initial model development and for testing (e.g., in the examination of regression residuals). Further, the availability of ever-greater computer power means that it becomes possible to explore more complex models, both in terms of local and global models and in terms of general semi-parametric models using Markov Chain Monte Carlo methods.

Finally, the ubiquitous nature of computer power means that many users of geographical information systems will do their own analysis using the tools to hand. It is imperative that those interested in spatial modeling continue to incorporate state of the art methods into the main GIS programs, as illustrated by Anselin et al. (2004c).

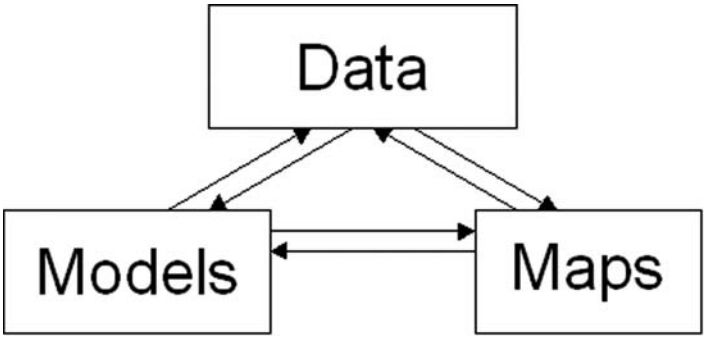


Fig. 12.4 Components of spatial model building

Appendix

Let $(w_{ij} : i, j = 1, K, n, i \neq j)$ denote the set of weights defining links between all pairs of cells in the grid. Following Cliff and Ord (1981, p. 17) we may define the generalized Moran statistic as

$$I = \frac{n \sum_{(2)} w_{ij} z_{ij}}{S_0 \sum_{i=1}^{n} z_i^2}, \quad \text{where } z_i = x_i - \bar{x}. \tag{12.12}$$

In this expression $S_0 = \sum_{(2)} w_{ij}$ and $\sum_{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n}$ with $i \neq j$. We assume that the random variables are independent and identically distributed. It follows that the first two moments of the statistic, under the null hypothesis of spatial independence are (Cliff and Ord, 1981, pp. 42–45):

$$E(I) = \frac{-1}{(n-1)}, \quad \text{and}$$

$$E(I^2) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{(n-1)(n+1) S_0^2}, \quad \text{where} \tag{12.13}$$

$$S_1 = \frac{1}{2} \sum_{(2)} (w_{ij} + w_{ji})^2 \quad \text{and} \quad S_2 = \sum_{i=1}^{n} (w_{i*} + w_{*i})^2,$$

$$w_{i*} = \sum_{j=1}^{n} w_{ij} \quad \text{and} \quad w_{*i} = \sum_{j=1}^{n} w_{ji}.$$

The values of the various coefficients for the RQ statistic on a regular RxC grid, with horizontal and vertical links having weight $= 2$ and diagonal link weight $= 1$ are

$$\begin{aligned} S_0 &= [12RC - 8R - 8C + 4], \\ S_1 &= [20RC - 12R - 12C + 4], \quad \text{and} \\ S_2 &= 16[36RC - 40R - 40C + 41]. \end{aligned} \tag{12.14}$$

The same coefficients are used for a test based upon random permutations, but the second moment becomes

$$E(I^2) = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2] - b_2[(n^2 - n)S_1 - 2nS_2 + 6S_0^2]}{(n-1)(n-2)(n-3)S_0^2}, \tag{12.15}$$

where $b_2 = \dfrac{n \sum z_i^4}{[\sum z_i^2]^2}$.

Chapter 13
Health Surveillance Around Prespecified Locations Using Case-Control Data

Peter A. Rogerson

Abstract There are several approaches one may use to model or test for potential risk around point sources of interest. These approaches have been developed almost universally to (a) fit model parameters to estimate the nature and significance of decline in risk as one moves away from the point source, or (b) assess the significance of a test statistic based upon the null hypothesis of no raised incidence around the source. In this paper, I assume that the data on the locations of cases and controls often used for these questions may be arranged in temporal order (for example, data might consist of the date of diagnosis for both case and control diseases). I then illustrate how conventional modeling approaches may be adapted to use the dataset observation by observation, to detect as quickly as possible a change from one set of model parameters to another.

13.1 Introduction

There is often interest in determining whether the number of observed health events in the vicinity of putative sources is greater than could be expected by chance alone. Both nonparametric tests (Stone, 1988) and parametric approaches making use of point process models (Diggle, 1990; Diggle and Rowlingson, 1994; Lawson, 1993) have been suggested. Diggle and Rowlingson (1994) have suggested a likelihood approach for testing retrospectively the null hypothesis of no raised incidence around a prespecified location, when data on the locations of cases and controls are available. Using their notation, assume that data are available on the locations of n cases and m controls. The intensity of disease, $\lambda(x)$, is modeled as

$$\lambda(x) = \rho\lambda_0(x)f(x - x_0; \Theta),$$

where $\lambda_0(x)$ represents background intensity due to the population at risk, and ρ is a scaling parameter related to the number of cases and controls. Furthermore, risk at

P. A. Rogerson
Departments of Geography and Biostatistics, University at Buffalo, Buffalo, NY, USA
e-mail: rogerson@buffalo.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*,
Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_13,

location x is presumed to vary with location according to the function $f(x - x_0; \Theta)$, where x_0 is the prespecified location, and where Θ is a set of parameters. They suggest the function:

$$f(x - x_0, \theta) = 1 + \theta_1 e^{-\theta_2 d^2}, \tag{13.1}$$

where d^2 is used to indicate the squared distance between locations x and x_0. The parameter θ_1 estimates the excess risk at the source, and the parameter θ_2 represents exponential decline in risk as one travels away from the source. Although we will use this function for convenience, the approach for prospective monitoring outlined here is general and other specifications for f that may be deemed more appropriate could also be adopted.

Conditional on the locations, the probability that an event at x is a case is

$$p(x) = \frac{\rho f(x - x_0; \theta)}{1 + \rho f(x - x_0; \theta)}.$$

The likelihood of the observed sample of cases and controls is given by

$$L(\rho, \theta) = n \ln \rho + \sum_{i=1}^{n} \ln f(x_i - x_0; \theta) - \sum_{i=1}^{n+m} \ln\{1 + \rho f(x_i - x_0; \theta)\}. \tag{13.2}$$

When the null hypothesis of no raised incidence around the prespecified point is true, and when (13.1) is used to model the relationship between location and risk, $\theta_1 = \theta_2 = 0$, and the likelihood equation in (13.2) reduces to

$$L_0(\rho) = n \ln \rho - (n + m) \ln(1 + \rho). \tag{13.3}$$

Under the null hypothesis, (13.3) is maximized at $\hat{\rho} = n/m$, and thus:

$$L_0(\hat{\rho}) = n \ln(n/m) - (n + m) \ln \left(\frac{n + m}{m} \right).$$

A formal test of the null hypothesis is carried out by (a) finding the parameter estimates that maximize (13.2), and then comparing the quantity $D = 2\{L(\hat{\rho}, \hat{\theta}) - L_0(\hat{\rho})\}$ with the critical value of a χ^2 distribution having degrees of freedom equal to the number of parameters in θ.

13.2 Prospective Monitoring

The test described above is appropriate when carrying out a single test to determine whether there is significant excess risk around a source, and whether risk declines significantly as one travels away from the source. However, it may also be of interest

to monitor data around point sources, to detect as quickly as possible any change in risk that may occur. It is therefore of interest to develop an appropriate methodology for the *prospective* detection of spatial variation in risk. In this section, we assume that case-control data are characterized not only by their locations, but also by a time variable (e.g., time of diagnosis). We make use of cumulative sum methods (Hawkins and Olwell, 1998) to detect changes in model parameters as quickly as possible.

These methods are optimal for detecting step changes in the parameters. They are based on the score statistic, z, which in turn is derived from the observations, x:

$$z_t = \ln\left(\frac{f(x_t|\theta^{(1)})}{f(x_t|\theta^{(0)})}\right) = \ln f(x_t|\theta^{(1)}) - \ln f(x_t|\theta^{(0)}), \tag{13.4}$$

where f designates the likelihood function, and $\theta^{(0)}$ and $\theta^{(1)}$ refer to the vector of parameters before and after the change, respectively.

These scores are then used to formulate the cumulative sum, S_t:

$$S_t = \max(0, S_{t-1} + z_t).$$

A change from $\theta^{(0)}$ to $\theta^{(1)}$ is detected when the cumulative sum, S_t, exceeds some predefined threshold, h. There is an inverse relation between the threshold and the rate of false alarms; higher values for h will lead to fewer false alarms, but also to longer times of detection when a true change has occurred.

Suppose for monitoring case-control data, we adopt Diggle's function (13.1), and set $\theta^{(0)} = \{\rho, \theta_1^{(0)}, \theta_2^{(0)}\} = \{\rho, 0, 0\}$, implying no raised incidence around the prespecified source prior to the change. Let $\theta^{(1)} = \{\rho, \theta_1^{(1)}, \theta_2^{(1)}\}$ be the parameters after the change. The likelihood of observing a control at a distance r from the source under the new regime is, using (13.2),

$$L(\theta^{(1)}) = -\ln\{1 + \rho(1 + \theta_1^{(1)} e^{-\theta_2^{(1)} r^2})\}. \tag{13.5}$$

The likelihood of observing a control under the initial regime implies, using (13.3) and the adoption of $\theta^{(0)} = \{\rho, 0, 0\}$,

$$L(\theta^{(0)}) = \ln(\rho) - \ln(1 + \rho). \tag{13.6}$$

Similarly, observation of a case under the new regime using (13.2) has the likelihood

$$L(\theta^{(1)}) = \ln\rho + \ln(1 + \theta_1^{(1)} e^{-\theta_2^{(1)} r^2}) - \ln\{1 + \rho(1 + \theta_1^{(1)} e^{-\theta_2^{(1)} r^2})\} \tag{13.7}$$

and observation of a case under the initial regime, using (13.3) and the adoption of $\theta^{(0)}\} = \{\rho, 0, 0\}$ leads to the following likelihood

$$L(\theta^{(0)}) = \ln(\rho) - \ln(1 + \rho). \tag{13.8}$$

These four expressions in (13.5) through (13.8) can then be used to form the likelihood ratio, and hence the score statistic for new observations as they become available.

13.3 Illustration

To illustrate the use of monitoring in this context, we adopt a circular study area of radius one, surrounding the hypothetical putative source. Observations were assigned case status with probability one-half and control status with probability one-half, and consequently we take the value of ρ under both the null and alternative hypotheses to be equal to one. Population is assumed to be distributed uniformly, and therefore cases and controls are generated under the null hypothesis by choosing locations at random within the study area.

To simulate the distances that cases and controls lie from the source, the following approach, based on the cumulative distribution function of distances, was used:

1. Take a randomly chosen number, u, from a uniform distribution on the interval (0,1)
2. Set it equal to the cumulative distribution function associated with the distribution of distances
3. Solve for the random variable representing distance from source

Thus the uniform random number, u, is set equal to the probability that an observation lies within a distance r (less than or equal to one, which is the radius of the study area) of the source. For both cases and controls under the null hypothesis,

$$u = F(r) = \frac{\int_0^r 2\pi x \, dx}{\int_0^{R=1} 2\pi x \, dx} = \frac{r^2/2}{1/2} = r^2.$$

Hence we can simply take as the simulated distance the square root of a random number chosen from a uniform distribution on the interval (0,1), since $r = \sqrt{u}$. The result represents the distance from the source at the center of the study area (the direction, and hence precise location is not important, since we are only interested in the distance from the source).

Under the alternative hypothesis, controls are generated in the same way, but cases are now chosen in such a way that they are more likely to occur near the putative source. Using the function suggested by Diggle (13.1),

$$u = F(r) = \frac{2\pi \int_0^r (1 + \theta_1 e^{-\theta_2 x^2}) x \, dx}{2\pi \int_0^{R=1} (1 + \theta_1 e^{-\theta_2 x^2}) x \, dx}$$

leads to

$$u = F(r) = \frac{\theta_1 + \theta_2 r^2 - \theta_1 e^{-\theta_2 r^2}}{\theta_1 + \theta_2 R^2 - \theta_1 e^{-\theta_2 R^2}} = \frac{\theta_1 + \theta_2 r^2 - \theta_1 e^{-\theta_2 r^2}}{\theta_1 + \theta_2 - \theta_1 e^{-\theta_2}}, \tag{13.9}$$

where the latter term on the right-hand side results from our choice of $R = 1$. For any choice of the random number u on the interval (0,1), we wish to solve for r. Since it is not possible to solve for r directly as a function of u, the solution for r is achieved by numerical root-finding methods (see appendix). Alternatively, it is possible to find an approximation for r in terms of u. First rearrange (13.9) as

$$u(\theta_1 + \theta_2 - \theta_1 e^{-\theta_2}) - \theta_1 = \theta_2 r^2 - \theta_1 e^{-\theta_2 r^2}.$$

Designating the left-hand side by y and solving this for r^2 in Maple 9.5 yields

$$r^2 = \frac{y + LambertW(0, \theta_1 e^{-1})}{\theta_2} \tag{13.10}$$

For a given argument x, the $LambertW$ function (Corless et al., 1996) returns the (possibly multiple) values $W(x)$ satisfying $W(x)e^{W(x)} = x$. For example, when $x = -0.1$, $W(x) = -3.577$ and $W(x) = -0.1118$ represent solutions. In (13.10), the first argument of "0" refers to a particular branch (in fact, the main branch) of the multivalued function, and the second term is the argument of the function.

There are alternative approaches to the numerical evaluation of the $LambertW$ function (Chapeau-Blondeau and Monir, 2002). There are also various series expansions for the $LambertW$ function (Corless et al., 1997); an evaluation of them reveals that different series are most accurate across different ranges of the argument. In particular, for various values of the argument z, $LambertW(0, z)$ may be approximated via the following series expansions:

$$\begin{aligned} LambertW(0, z) &\approx z - z^2 + z^3/3 - \cdots \quad z \le 0.3 \\ &\approx \frac{2z}{z+e} + \frac{z(z-e)^2}{2(z+e)^3} + \cdots \quad 0.3 < z \le 4 \\ &\approx v + \frac{vp}{1+v} + \frac{vp^2}{2(1+v)^3} + \cdots \quad z > 4, \end{aligned} \tag{13.11}$$

where $v = \ln z$ and $p = -\ln(\ln z))$. When $z < 0.3$, combining (13.10) and the first few terms of (13.11) yields

$$r^2 \approx \frac{y + \theta_1 e^{-y} - (\theta_1 e^{-y})^2}{\theta_2} \approx \frac{y + \theta_1 e^{-y}}{\theta_2}. \tag{13.12}$$

Using just the first term on the RHS of (13.11) for the middle range of z yields

$$r^2 \approx \frac{y + 2q/(q+1)}{\theta_2},$$

where $q = \theta_1 e^{-y-1}$. Finally, when z is large, using the first two terms of the last approximation in (13.11) leads to

$$r^2 \approx \frac{y + w - (w \ln w)/(1+w)}{\theta_2},$$

where $w = \ln \theta_1 - y$.

For example, suppose we wish to simulate for the scenario where $\theta_1 = 2$, $\theta_2 = 4$. Suppose the random number is $u = 0.5$. Then $y = 0.982$. Then $r^2 = 0.771$ or 0.768, using the two- and three-term numerators of (13.12), respectively.

After distances have been simulated, the scores (z_t) are computed. Observation of a control leads to a score found by subtracting (13.6) from (13.5):

$$\begin{aligned} z_t &= -\ln(1 + (1 + \theta_1^{(1)} e^{-\theta_2^{(1)} r_1^2})) - \ln(1) + \ln(2) \\ &= -\ln(2 + \theta_1^{(1)} e^{-\theta_2^{(1)} r_1^2}) + \ln(2) \end{aligned} \tag{13.13}$$

Observation of a case leads to a score found by subtracting (13.8) from (13.7),

$$\begin{aligned} z_t &= \ln(1) + \ln(1 + e^{-\theta_2^{(1)} r_1^2}) - \ln(1 + 1 + e^{-\theta_2^{(1)} r_1^2}) - \ln(1) + \ln(2) \\ &= -\ln(1 + e^{-\theta_2^{(1)} r_1^2}) - \ln(2 + e^{-\theta_2^{(1)} r_1^2}) + \ln(2) \end{aligned} \tag{13.14}$$

These scores are then used in the cumulative sum.

We now illustrate the procedure and results for the choices $\theta^{(0)} = \{1, 0, 0\}$ and $\theta^{(1)} = \{1, 2, 4\}$. The null hypothesis is simulated by first assigning case/control status (using $1/2$ as the probability an observation is a case) and then choosing distances $r = \sqrt{u}$ for both cases and controls. These distances are then used to determine scores [(13.13) and (13.14)], and the cumulative sum is run until it reaches a threshold, h. Suppose that we desire an average run length (ARL) of 250 observations between false alarms – i.e., declarations of change when in fact none has occurred. Experimentation with different threshold values revealed that a value of $h = 2.0$ is consistent with an in-control ARL of approximately 250. Next, the alternative hypothesis was simulated by using distances determined from (13.9) and the numerical method outlined in the appendix, using parameters equal to those chosen for $\theta^{(1)} = \{2, 4\}$. The ARL under this alternative hypothesis was approximately 103; this is the number of observations that would be required on average to detect the change in risk.

For more pronounced increases in risk near the source, detection occurs more quickly, as would be expected. For example, with $\theta^{(1)} = \{6, 5\}$, a threshold of $h = 2.7$ leads to an ARL of approximately 250 (making this instance comparable with the previous one); when the alternative hypothesis is simulated using the chosen values of $\theta^{(1)}$, the average time to detection declines to approximately 60 observations.

It will of course not usually be possible to specify correctly the magnitude of the shift. For example, suppose that a shift to $\theta^{(1)} = \{2, 4\}$ is hypothesized, but the

actual shift is to $\{2, 4\}$. A threshold of $h = 1.59$ leads to an ARL of 250 under the null hypothesis, and now the average time to detect a change is approximately 110 – slightly longer than the 103 found earlier when a correct estimate of the shift is adopted (using $\theta^{(1)} = \{2, 4\}$). Similarly if a shift to $\theta^{(1)} = \{3, 5\}$ is hypothesized, a threshold of $h = 2.14$ leads to an ARL of 250 under the null hypothesis of no change, and the average time to detect the shift to $\{2, 4\}$ is approximately 107. At least in these examples, therefore, mis-estimation of the magnitude of the shift does not affect significantly the time to detection.

13.4 Summary

Cumulative sum methods may be used together with case-control data to detect quickly changes in risk that occur in the neighborhood of a putative source. The methods are based upon scores that represent the difference in log-likelihoods of the observations before and after the change.

One limitation of the approach outlined here is that it is necessary to prespecify the parameters before and after the change. Although the choice of $\theta_1^{(0)} = \theta_2^{(0)} = 0$ is natural, it will often be more difficult to specify $\theta_1^{(1)}$ and $\theta_2^{(1)}$. In general the cumulative sum procedure will be efficient at detecting quickly changes from $\theta^{(0)}$ to $\theta^{(1)}$, and will be relatively less efficient at detecting changes to parameter values other than those specified.

Similarly, we have simplified the problem by assuming that ρ does not change. This is analogous to cumulative sum procedures that are optimized for detection of changes in the mean, where it is assumed that the variance does not change. Any change in ρ will of course render the cumulative sum scheme less effective.

Finally, we have focused here solely on the spatial aspects of the process. Although we have been interested here in the unfolding of spatial patterns over time, we have essentially ignored the temporal aspect of the process. In any analysis of disease, attention must of course also be given to how the intensity of the disease varies over time.

Appendix

To find r in terms of u, we first write

$$g(r) = u - \frac{\theta_1 + \theta_2 r^2 - \theta_1 e^{-\theta_2 r^2}}{\theta_1 + \theta_2 R^2 - \theta_1 e^{-\theta_2 R^2}} = 0.$$

Taking the derivative, we find

$$g'(r) = \frac{-2\theta_2 r(1 + \theta_1 e^{-\theta_2 r^2})}{\theta_1 + \theta_2 R^2 - \theta_1 e^{-\theta_2 R^2}}.$$

An initial guess for r, say r_0, is made. This initial guess is updated to derive an improved estimate of r, say r_1, via

$$r_1 = r_0 - \frac{g(r_0)}{g'(r_0)}.$$

This is used to iterate until convergence has been achieved.

Part IV
Applications

Chapter 14
Spatial Filtering in a Regression Framework: Examples Using Data on Urban Crime, Regional Inequality, and Government Expenditures

Arthur Getis

> If autocorrelation is found, we suggest that it be corrected by appropriately transforming the model so that in the transformed model there is no autocorrelation (Gujarati, 1992, p. 373).

14.1 Introduction

In a recent paper Getis (1990), I develop a rationale for filtering spatially dependent variables into spatially independent variables and demonstrate a technique for changing one to the other. In that paper, the transformation is a multi-step procedure based on Ripley's second order statistic (1981). In this chapter, I will briefly review the argument for the filtering procedure and propose a simplified method based on a spatial statistic developed by Getis and Ord (1992). The chapter is divided into four parts: (1) a short discussion of the rationale for filtering spatially dependent variables into spatially independent variables, (2) a review of a Getis–Ord statistic, (3) an outline of the filtering procedure, and (4) three examples taken from the literature on urban crime, regional inequality, and government expenditures.

14.2 Rationale for a Spatial Filter

One of the most difficult problems facing those who develop regression models of spatial series is finding ways to estimate parameters of stochastically autocorrelated variables. A typical stochastically autocorrelated spatial variable is a modified or spatially lagged autocorrelated variable. It is made up of the original autocorrelated variable (y) multiplied by a spatial weight matrix (W) and a spatial autocorrelation coefficient (ρ). ρW_y does not fulfill the required fixed-effects linear regression assumption that correlated variables are not to be stochastic. In this case, since

A. Getis
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*,
Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_14,

ordinary least squares yields biased parameter estimates and R-squared values, other estimation techniques must be considered, such as maximum likelihood estimation. In addition, any remaining spatial dependence in the regression equation, as may be evident in error terms, must be accounted for. In multiple variable cases, it may be necessary to develop a series of W matrices, thus further complicating the meaning of the various tests on the significance of the parameters.

Because of the complexity of the typical spatial regression formulation, I propose that the spatial dependence within each dependent and independent variable be filtered out before the estimation procedure is adopted. This proposal has the following ramification: one must somehow reintroduce into the regression equation the removed spatial dependence in order to avoid misspecification. No spatial dependence should be evident in the error term since supposedly it has been removed from all of the possible sources. Since each variable on the right hand side is no longer stochastically correlated nor spatially autocorrelated, and there is only the usual spherical error, ordinary least squares can be used for estimation.

This argument for filtering from spatially autocorrelated variables the spatial dependence effects is sound only insofar as there is a way to accomplish the task. Before demonstrating the variable filtering procedure, let me briefly describe a statistic that will act as its foundation.

14.3 The G_i Statistic

This statistic measures the degree of association that results from the concentration of weighted points (or areas represented by weighted points) and all other weighted points included within a radius of distance d from the original weighted point.[1] We are given an area subdivided into n regions, $i = 1, 2, \ldots, n$, where each region is identified with a point whose Cartesian coordinates are known. Each i has associated with it a value x (a weight) taken from a variable X. The variable has a natural origin and is positive.[2] The statistic is written as

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j}, \qquad j \neq i, \tag{14.1}$$

where w_{ij} is a one/zero spatial weight matrix with ones for all links defined as being within distance d of a given i; all other links are zero. The numerator is, therefore, the sum of all x within d of i except when i equals j. The denominator is the sum of all x_j except when i equals j. The mean is

$$E(G_i) = \frac{W_i}{(n-1)}, \tag{14.2}$$

where $W_i = \sum_j w_{ij}(d)$.

[1] For a full discussion see Getis and Ord (1992).

[2] A more recent version of this statistic in Ord and Getis (1993) avoids these restrictions.

If we set $\sum_j x_j/(n-1) = Y_{i1}$ and $\sum_j x_j^2/(n-1) - Y_{i1}^2 = Y_{i2}$ then:

$$Var(G_i) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)}\left(\frac{Y_{i2}}{Y_{i1}^2}\right). \tag{14.3}$$

$G_i(d)$ measures the concentration or lack of concentration of the sum of values associated with variable X in the region under study. $G_i(d)$ is to be differentiated from a statistic $G_i^*(d)$ that takes into account the value of x at i, that is, j equal to i. $G_i(d)$ is a proportion of the sum of all x_j values that are within d of i. If, for example, high-valued $x_j's$ are near to the point i, and d includes these high values, so that a large proportion of the sum of all $x_j's$ is within d of i, then $G_i(d)$ is high. Whether the $G_i(d)$ value is statistically significant depends on the statistic's expectation.

For our purposes here, the most important characteristic of the statistic is that it gives the proportion of the summed variable within a specified distance from a particular point i as a part of the entire summed variable. When this value is compared to the statistic's expectation, the difference tells us the degree of clustering of the sum of the x variable in the vicinity of i that is greater or less than chance would have it.

14.4 The Filtering Procedure

The rationale for transforming a spatially dependent variable into a spatially independent variable is that the spatial dependence can be removed from the spatially dependent variable and replaced as a separate independent variable. An easy solution to this problem, but useless, would be to set all values of the spatially dependent variable to the mean. This "variable" would not be spatially dependent and it would not correlate with any other variable. The solution I outline below attempts to adjust the spatial dependent variables only to the point where spatial dependence is no longer embodied in them. That which is filtered from the original variable becomes a new spatial variable. It may be that the autocorrelation filtered from one variable is highly correlated with that which is filtered from another. In a regression equation, in order to avoid multicollinearity, in the final equation it may be necessary to use only one spatial variable rather than all spatial variables extracted from the original variables.

Suppose that within distance d of a point, x_1, there are two other points with values, x_2, and x_3, which when summed are a greater proportion of all x (minus x_1) than what one should expect in a similar spatial configuration with the same d value when all x values are randomly distributed. This means that $G_1(d)$ is greater than the random expectation, $W_1/(n-1)$. Suppose $G_1(d)$ is the proportion 0.40 and $E[G_1(d)] = 0.30$. Then 40% of the sum of all observed x (not counting x_1 itself) is contained within d of x_1, while the expectation is only 30%. We then call 30/40 of x_1 the filtered value of x_1. The difference between the original value and

the filtered value is that which is filtered out due to the spatial clustering of x values in the vicinity of x_1. In this example, the ratio 30/40 represents the degree to which x_1 is similar to its expectation. The degree of dissimilarity, 10/40, represents that which is due to positive spatial association. Negative spatial association is found in like manner. Thus,

$$x_i^* = \frac{x_i(\frac{W_i}{n-1})}{G_i(d)} \tag{14.4}$$

which when solved for all x_i, represents the filtered variable X^*. The difference between X and X^* is a new variable, L, that represents the spatial effects embedded in X.

For realistic filtering it is essential to find an appropriate d value. The value should represent the distance within which spatial dependence is maximized. In Getis (1990), d corresponds to the maximum total sum of squared differences between the observed and the expected G_i values. On reflection, however, a different d value seems more appropriate. This is the value that corresponds to the maximum absolute sum of the normal standard variate of the statistic $G_i(d)$ for all i observations of the variable X. This single value is chosen since it represents overall the distance beyond which no further association or nonassociation effects increase the probability that the observed value is different than the expected value. One might argue that beyond this d value there is an overall cessation of spatial effects for the variable in question. For each variable, we use this value in the examples given below. A more detailed approach, but less general, would be to identify a critical d value for each individual point i. Clearly, more research is needed on this subject.

Our approach to spatial filtering can be given credence by showing that the following four conditions hold:

1. There is no spatial autocorrelation embodied in X^*.
2. If X is a variable with spatial dependence embedded in it, then the difference between X and X^* is a spatially autocorrelated variable (L).
3. In any regression model where all variables have been filtered using an appropriate distance d, residuals are not spatially associated.
4. In a regression equation, appropriate variables should be statistically significant after spatial dependence has been removed from them. Of course, appropriateness in this case requires theoretical justification. In this chapter, we will be satisfied if intuitively correct variables are statistically significant after spatial dependence has been removed.

If these conditions are met, one might conclude that a reasonable estimate of spatial autocorrelation has been found and that ordinary least squares is appropriate for regression modeling where this type of filtering has been used. In the next section, these conditions are demonstrated by way of three examples.

14.5 Filtering Variables: Three Examples

14.5.1 Example 1: Urban Crime

Anselin (1988) provides data on three variables (crime, income, and housing) by neighborhoods (given as points with x, y coordinates) for Columbus, Ohio, for 1980 ($n = 49$). The autocorrelated variable, crime (CR), is constructed from the number of burglaries and thefts per thousand households, and income (IN) and housing (HO) are given by household in thousands of dollars. When $d = 4$ km spatial association is at a maximum for crime, $d = 3$ for housing and income. In the tests that follow, these values for d will be used. The trial OLS model is

$$\begin{array}{cccccc} CR = & 68.62 & - 1.597IN & - 0.274HO \\ (t) & (14.49) & (-4.78) & (-2.65) \end{array} \tag{14.5}$$

The adjusted $R^2 = 0.533$ and the standardized Moran's I on the residuals is 2.765. The criterion for spatial dependence in this and all subsequent tests is the 0.05 level of the normal curve calculated from the / statistic (randomization, distance effect = $1/d^2$) of Moran (Cliff and Ord, 1973). This statistic will not be reviewed here. In this case then, the residuals are spatially statistically significant. In addition, it is important to note that two of the variables reveal strong spatial autocorrelation:

Variable	$Z(I)$
CR	7.345
IN	4.168
HO	1.903

Test 1: There is no spatial autocorrelation embodied in X^*.

This test requires that the transformed variables, CR^*, IN^*, and HO^* are not spatially autocorrelated. The results are as expected:

Variable	$Z(I)$
CR^*	-0.152
IN^*	-0.280
HO^*	-0.454

Test 2: If X is a variable with spatial dependence embedded in it, then the difference between X and X^* is a spatially autocorrelated variable (L).

The results of the test (see below) show that all three L variables are spatially autocorrelated. Although it is not required for the housing variable to have a high $Z(I)$ value, it, too, has spatial dependence embedded within it.

Variable	$Z(I)$
L_{CR}	7.659
L_{IN}	2.059
L_{HO}	3.044

Test 3: In any regression model where all variables have been filtered using an appropriate distance d, residuals are not spatially associated.

In this case we use the linear regression:

$$CR^* = a - IN^* - HO^* + e \tag{14.6}$$

and then test e for spatial association using I. Clearly, this is an inadequate model but it does serve to show how the filtering of variables satisfies our intuition. As expected, the residuals are not spatially associated as $Z(I)$ = 0.161. The correlated variables explain 32% of the variation in the filtered crime variable. In the unfiltered case shown above, the adjusted R^2 is 0.533. This indicates that by removing the spatial association, the model becomes a weaker predictor of the location of the incidence of crime.

Test 4: In a regression model, appropriate variables should be statistically significant after spatial dependence has been removed from them.

Based on the earlier tests, an intuitively appealing model designed to "explain" crime is

$$\begin{array}{lccccc} CR = & 60.15 & -0.96IN^* + 0.94L_{IN} & -0.27HO^* & -0.63L_{CR} \\ (t) & (14.11) & (-3.40) \quad (-2.97) & (-3.28) & (-6.08) \end{array}. \tag{14.7}$$

The adjusted R^2 is 0.719, and the residuals are not spatially associated as $Z(I)$ = 1.1618.

Discussion: All four tests had satisfactory outcomes. The example tells us much about the crime, income, and housing data of Columbus. Note that the adjusted R^2 value decreases when each of the variables is filtered for spatial association (0.533–0.323). When the spatial effects are reintroduced to the filtered equation, the R^2 increases (0.323–0.719). The implication of this is that although filtered equations are free of spatial effects, they must be included in the final model in order to account for the spatial effects. In the case of crime in Columbus, it appears that the configuration of the data units has as much to do with explaining crime as does the income and housing variables. The proliferation of small, spatially autocorrelated units in the high crime areas contributed to the inappropriateness of the original unfiltered equation.

14.5.2 Example 2: Regional Inequality

In a study of regional disparities among the 16 regional divisions of Turkey, Atalik (1990) tested the Cobb-Douglas type model:

$$Y = aP^b I^c S^d A^f \tag{14.8}$$

which can be written as

$$\log Y = \log a + b \log P + c \log I + d \log S + f \log A, \tag{14.9}$$

where Y is GDP per capita, P activity rate (share of the country's active population), I infrastructure rate (share of the country's literate population), S industrial employment rate, and A agglomeration rate (the proportion of the people of the region residing in the largest city).

The trial OLS equation is

$$\begin{array}{ccccccc} \log Y = & 0.81 \log a - & 0.79 \log P + & 1.27 \log I + & 0.63 \log S + & 0.05 \log A \\ (t) & (0.23) & (-0.47) & (2.28) & (2.12) & (0.38) \end{array} \tag{14.10}$$

in which each variable is spatially autocorrelated (see below) but the residuals are not autocorrelated ($Z(I) = -0.8061$; see Fig. 14.1). Log P has the wrong sign, and the $\log P$ and $\log A$ variables are not significant. Log I and $\log S$ are significant predictors of $\log Y$. The adjusted R^2 is 0.797.

In order to give spatial meaning to these results, coordinates were identified that correspond to the population centroid of each district Using the procedure described earlier, the critical d value was found to be 187.5 miles for each variable (increments of 18.75 miles were examined starting at 18.75 miles). The $Z(I)$ values shown below indicate that there is strong spatial autocorrelation in the data. There is a clear east-west trend in the data; the more favorable socioeconomic values are in the west (see Fig. 14.2).

$Variable$	$Z(I)$
$\log Y$	5.608
$\log P$	3.777
$\log I$	4.437
$\log S$	5.603
$\log A$	2.790

Test 1: All filtered variables are free of spatial autocorrelation.

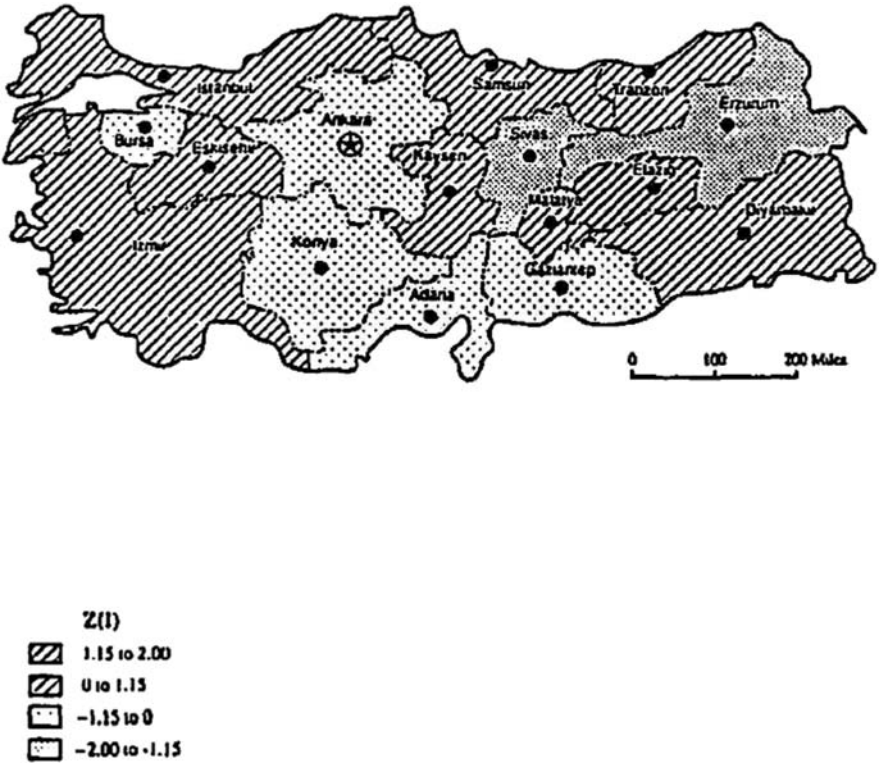


Fig. 14.1 There is no discernible spatial pattern of the residuals of the trial equation. One might get the impression from this that there is no spatial autocorrelation in the data

Variable	$Z(I)$
$\log Y^*$	-1.203
$\log P^*$	-1.720
$\log I^*$	-0.131
$\log S^*$	-0.282
$\log A^*$	-0.065

Test 2: L variables based on spatially autocorrelated variables are spatially autocorrelated.

Variable	$Z(I)$
$L_{\log Y}$	5.396
$L_{\log P}$	3.292
$L_{\log I}$	4.861
$L_{\log S}$	5.223
$L_{\log A}$	4.486

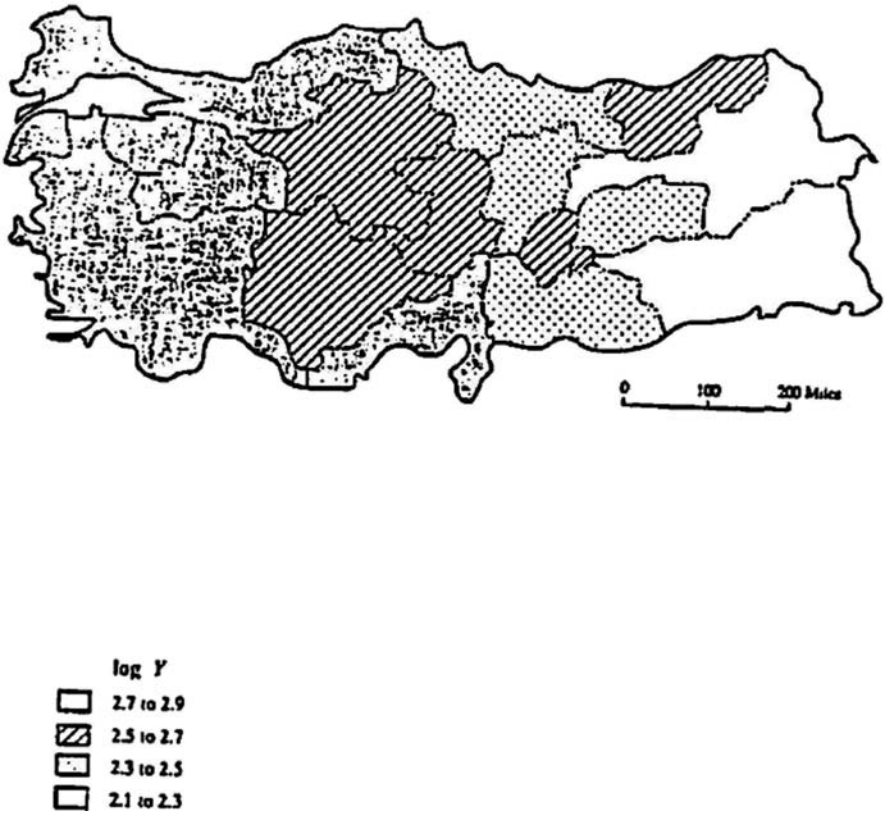


Fig. 14.2 The dependent variable. Log Y, is high in the west and low in the east. AH other variables in the trial equation act similarly

Test 3: Residuals of the filtered regression equation are not spatially associated.

The equation is

$$\log Y^* = \log a + \log P^* + \log I^* + \log S^* + \log A^* + e. \tag{14.11}$$

The $Z(I)$ value of 1.8027 for residuals is not significant.

Test 4: In a regression model, appropriate variables should be statistically significant after spatial dependence has been removed from them.

Taking into account the results of the trial OLS model shown above and multicollinearity, an appropriate equation is

$$\begin{matrix} \log Y = & 0.23 \log a + & 0.92 \log I^* + & 0.53 \log S^* - & 4.51 \log L_{\log S} + e \\ (t) & (0.24) & (1.67) & (2.29) & (-6.91) \end{matrix}. \tag{14.12}$$

The residuals are not spatially autocorrelated as $Z(I) = -1.5282$. The adjusted R^2 is 0.836, an increase over the 0.797 of the trial model.

Discussion: Again, the four tests show how an inadequate model can be recast in order to account for spatial association in the constituent variables. In this case, not only were we able to show that the east-west trend in the data resulted in a poor trial equation, but that a stronger model can be constructed with fewer variables. As in the first example, the adjusted R^2 value decreases when each of the variables is filtered for spatial association (0.797–0.255). When the spatial effects are reintroduced to the filtered equation, the R^2 increases (0.255–0.836). In this case, the filtered infrastructure rate ($\log I^*$), i.e., literacy, and filtered industrial employment rate ($\log S^*$) together with the strong west to east diminishing trend in the industrial employment rate satisfactorily explain the level of GDP per capita in Turkey.

14.5.3 Example 3: Government Expenditures

In the well known econometrics text by (Pindyck and Rubinfeld, 1981, pp. 169–170), an expenditure data set based on Bureau of Census information is given for the states of the United States for 1970. Variables given in the table are transformed by population size by the authors to study the response variable, state and local government expenditures per capita in dollars ($PCEXP = EXP/POP$), using the regressors: federal grants to each state per capita in dollars ($PCAID = AID/POP$), population in thousands ($POP = 1/POP$), and personal income per capita ($PCINC = INC/POP$) (Pindyck and Rubinfeld, 1981, p. 272). The trial OLS equation is

$$\begin{matrix} PCEXP = -405.81 + 1.63PCAID + 25779POP + 0.210PCINC \\ (t) \qquad (-4.01) \qquad (6.18) \qquad (1.30) \qquad (10.08) \end{matrix} \tag{14.13}$$

and the adjusted R^2 is 0.753.

Population centroids of states were estimated. Using increments of 33.3 miles, the critical d values are 333 miles for the first three variables and 267 miles for $PCINC$. The values shown below indicate that only one variable, $PCINC$, is spatially autocorrelated.

Variable	$Z(I)$
$PCEXP$	1.861
$PCAID$	0.291
POP	−0.080
$PCINC$	5.063

Test 1: All filtered variables are free of spatial autocorrelation.

Variable	$Z(I)$
$PCEXP^*$	-1.619
$PCAID^*$	-1.147
POP^*	0.588
$PCINC^*$	-0.250

Although all filtered variables are free of spatial autocorrelation, the only requirement for this test is that $PCINC^*$ not be spatially autocorrelated. In any case, we always expect filtered variables not to be spatially statistically significant.

Test 2: L variables based on spatially autocorrelated variables are spatially autocorrelated.

Variable	$Z(I)$
L_{PCEXP}	3.705
L_{PCAID}	1.238
L_{POP}	1.208
L_{PCINC}	5.274

The requirement that L_{PCINC} autocorrelated holds. Note that $PCEXP$ also contains much embedded spatial dependency.

Test 3: Residuals of the filtered regression equation are not spatially autocorrelated.

For the equation:

$$PCEXP^* = a + PCAID^* + POP^* + PCINC^* + e \tag{14.14}$$

the $Z(I)$ value for the residuals is -0.666 and the adjusted R^2 is 0.439, which is considerably less than the 0.753 value of the trial OLS equation.

Test 4: In a regression model, appropriate variables should be statistically significant after spatial dependence has been removed from them.

By experimentation, the best model for $PCEXP$ is found to be

$$\begin{array}{cccccc} PCEXP = & 369.68 & + 1.72PCAID^* & + 0.19PCINC^* & & \\ (t) & (-3.29) & (7.28) & (9.42) & & \\ & -2.27L_{PCAID} & - 0.26L_{PCINC} & & & \\ & (-4.06) & (-3.84) & & & \end{array} . \tag{14.15}$$

The adjusted R^2 is 0.747 and the $Z(I)$ of the residuals is -0.819. There is essentially no multicollinearity in this formulation, and all variables are highly significant.

Discussion: The explanation for $PCEXP$ is made up of both the non-spatial aspects and the spatial configuration of $PCAID$ and $PCINC$. Note that a considerable degree of questionable explanation (0.753) in the trial model exists when spatial association is not taken into account. The final model is slightly less in explained variance (0.747), but it is free of embedded spatial autocorrelation.

14.6 Conclusions

The results of the tests indicate that for the cases presented here, in every instance the filtering procedure conforms to our expectation. Clearly, these are but three case studies. Many more examples should be carried out. In the cases shown here, the procedure considerably helps in isolating the spatial dependence embedded within spatial series variables. In addition, the methodology aids in the proper specification of multiple regressionrelationships. The simplicity and ease of understanding made available by least squares methodology enable us to avoid estimation procedures that rob us of the convenience of R^2 interpretations.

Acknowledgements I would like to thank Giuseppi Arbia for his suggestions on an earlier version of this chapter. Serge Rey for suggesting the data used in the government expenditures example, and the editors, Luc Anselin and Raymond Florax, for helpful comments.

Chapter 15
Characteristics of the Spatial Pattern of the Dengue Vector, *Aedes aegypti*, in Iquitos, Peru

Arthur Getis, Amy C. Morrison, Kenneth Gray, and Thomas W. Scott

This Chapter was originally published in:
Getis A, Morrison AC, Gray K, Scott TW, 2003. Characteristics of the Spatial Pattern of the Dengue Vector, *Aedes aegypti*, in Iquitos, Peru. *Am J Trop Med Hyg* 69: 494–505

Abstract We determine the spatial pattern of *Aedes aegypti* and the containers in which they develop in two neighborhoods of the Amazonian city of Iquitos, Peru. Four variables were examined: adult *Ae. aegypti*, pupae, containers positive for larvae or pupae, and all water-holding containers. Adults clustered strongly within houses and weakly to a distance of 30 m beyond the household; clustering was not detected beyond 10 m for positive containers or pupae. Over short periods of time restricted flight range and frequent blood-feeding behavior of *Ae. aegypti* appear to be underlying factors in the clustering patterns of human dengue infections. Permanent, consistently infested containers (key premises) were not major producers of *Ae. aegypti*, indicating that larvaciding strategies by themselves may be less effective than reduction of mosquito development sites by source reduction and education campaigns. We conclude that entomologic risk of human dengue infection should be assessed at the household level at frequent time intervals.

15.1 Introduction

Patterns of dengue transmission are influenced by the abundance, survival, and behavior of the principal mosquito vector, *Aedes aegypti* (L.); the level of immunity to the circulating virus serotype in the local human population; density, distribution and movement of humans; and time required for development of virus in *Ae.*

A. Getis (✉) and K. Gray
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: arthur.getis@sdsu.edu

A. C. Morrison and T. W. Scott
Department of Entomology, University of California, Davis, Davis, CA, USA

aegypti (Halstead, 1990). The relative influence of these factors on dynamics of virus transmission is poorly understood, including how they vary through space and time. Although the apparent clustering of human cases of dengue within households has been reported previously (Halstead et al., 1969; Waterman et al., 1985) there has been little formal spatial research on the distribution pattern of *Ae. aegypti* and dengue cases. An exception was the spatial statistics study of a dengue epidemic in Florida, Puerto Rico by Morrison and others (1998). They found that dengue cases clustered within individual households over short periods of time and that a large proportion of the entire municipality of 9,000 people was affected within seven weeks of the first reported case. Presumably the same, or very few, infected adult mosquitoes were causing the household case clusters while infected humans traveling within the town may have facilitated the rapid spread of infections. The most effective dengue control programs rely on entomologic, viral, serologic, and clinical surveillance (Gubler, 1993). Early detection of virus activity allows for more streamlined application of vector control measures. Because there is no vaccine or clinical cure for dengue, mosquito control is the only method of reducing virus transmission. Effective serologic and viral surveillance is often beyond the resources of the majority of affected, developing countries. Consequently, they rely on entomologic surveillance to estimate potential risk for virus transmission and disease.

Traditional *Ae. aegypti* control measures include elimination (source reduction) or treatment of larval habitats to prevent production of adults and insecticidal space spraying to reduce adult population densities (Gubler, 1993; Reiter and Gubler, 1997). Contemporary programs emphasize reducing *Ae. aegypti* populations to levels that prevent or slow virus transmission with the ultimate objective of decreasing the incidence of disease, especially severe, life-threatening illness. However, traditional entomologic surveillance techniques are based on a series of indices that were designed to detect the presence or absence of *Ae. aegypti* larvae. Those methods assume a strong positive correlation between the presence of larvae and adult females in a household: only adult females transmit virus to humans. There are, however, three important reasons to question the strength of the larvae-adult association. First, because larval mortality can be high, adults may not emerge from a container holding immature mosquitoes. Alternative entomologic surveillance methods, especially pupal surveys, were developed to circumvent this shortcoming (Focks and Chadee, 1997). Second, because adults are capable of flight, they can move away and become spatially disassociated from their development sites. Third, independent of the surveillance technique (larvae, pupae, or adult collections) city-wide surveys are often carried out in such a way that the number and location of households selected are derived from standard parametric sample size calculations. The assumption that there is no spatial structure among infested houses must be validated.

The purpose of this study was to characterize the spatial distribution of *Ae. aegypti* populations in two representative neighborhoods in the Amazonian city of Iquitos, Peru over two time periods. Specifically, from complete samples of households in two areas of Iquitos we examined the (1) underlying spatial structure

of *Ae. aegypti* infestations (larvae, pupae, and adult), (2) temporal stability of that structure, and (3) correlation between clusters at different life stages of the mosquito. We conclude by discussing the implications of our findings on estimation of entomologic risk to epidemiologic studies of dengue and routine dengue surveillance.

15.2 Materials and Methods

Study Area

The area chosen for this study consists of two neighborhoods in Iquitos (73.2°W, 3.7°S, and 120 m above sea level), a city that is surrounded on three sides by the Amazon, Nanay, and Itaya Rivers. Because Iquitos is accessible only by air or river, it is a geographically isolated city of approximately 345,000 people in the Amazon forest (Watts et al., 1999) (Fig. 15.1). The major industries in Iquitos are small commercial enterprises, fishing, oil, lumber, and to some extent agriculture.

The two neighborhoods where we carried out entomologic surveys were Maynas, located in the north central part of the city, and Tupac Amaru, situated in the southwestern-most part of the city (Fig. 15.1). We selected these two neighborhoods because they were characterized as areas of high (Maynas) and low (Tupac Amaru) prevalence of human dengue infection in previous informal studies (Morrison, A.C. and Scott, T.W., unpublished data). Although Maynas could be characterized as the wealthier and older of the two neighborhoods, households within both areas vary greatly in socioeconomic status so that well constructed households with piped water and poorly constructed households with no water or sewer services exist in both neighborhoods in a patchwork. Nevertheless, there are some distinct differences between the two neighborhoods. Maynas has a higher proportion than Tupac Amaru of permanent houses constructed with bricks and concrete. Conversely, Tupac Amaru is a community in transition from predominantly temporary wood houses with palm roofs to houses constructed with brick and concrete. Even though Maynas has a better-developed sewer system than Tupac Amaru, the Maynas water supply is inconsistent. Consequently, Maynas residents are more likely than those in Tupac Amaru to store water in containers that are potential development sites for immature *Ae. aegypti*. In contrast, Tupac Amaru has many open sewers but because of close proximity to the city water plant most houses have a stable water supply and are less likely to store water than in Maynas.

Study Design

A unique-house code was painted on the front of each of the 550 houses located on 20 blocks in Maynas and the 510 houses located on 14 blocks in Tupac Amaru. Almost all houses have at least one wall in common with a neighboring house. Beginning in mid-November 1998, five two-person entomology collection teams

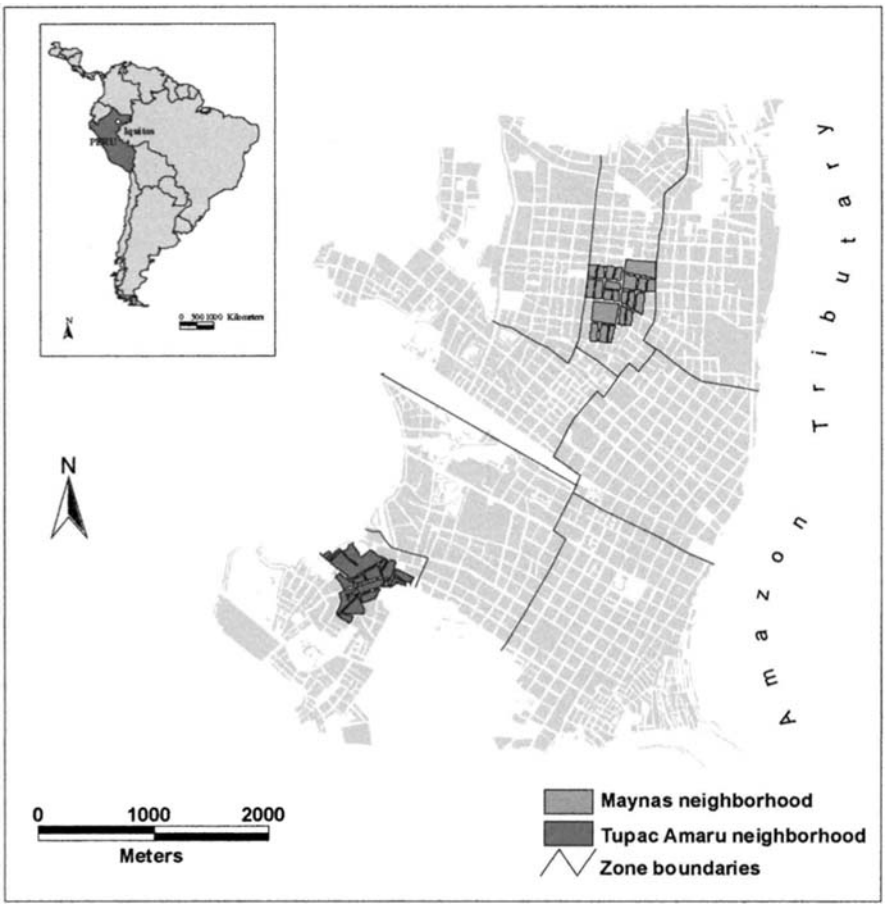


Fig. 15.1 Map of Iquitos, Peru and location of the Maynas and Tupac Amaru study areas

were provided a map of a block to be surveyed with a designated start house. Households were surveyed in sequence daily along the block from the start house between 7:00 a.m. and 1:00 p.m. Unoccupied or closed houses and houses where residents did not provide permission for the survey, businesses, offices, and schools were not sampled. Thus, we were able to survey 95% of the houses in both surveys: 528 in Maynas and 481 in Tupac Amaru. Collecting teams were rotated among blocks each day in an attempt to limit temporal and collector biases. Each day, prior to continuing surveys of unsampled households, an attempt was made to inspect houses that were previously closed or where access had been refused. Access to houses was attempted a minimum of three times. Maynas and Tupac Amaru were surveyed on alternating days. This process was carried out until all the houses in each neighborhood had been surveyed or repeated attempts to gain access failed. In mid-December 1998, immediately after termination of the first survey, the sampling procedure was

repeated. The second survey was completed on January 18, 1999. To differentiate data associated with the four different collections, the two surveys will be referred to as a (November–December) and b (December–January).

Entomologic Surveys

Our survey methodology was based on techniques suggested by Focks and others (1993). Briefly, after asking permission to survey the household, one member of the team administered a demographic survey designed to determine the number of occupants, dimensions of the property, house construction materials, method of cooking, water use patterns, type of sewage disposal, and insecticide use. Simultaneously, the other team member began collecting adult mosquitoes using a backpack aspirator (John W. Hock Company, Gainesville, FL) (Scott et al., 2000a). Aspiration collections were attempted in all rooms of the house (when permitted) including walls, under furniture, and inside closets and other likely adult mosquito resting sites. Aspiration collections were similarly attempted outside the house from outside walls, under eaves, vegetation, and in and around outdoor stored materials.

In our field laboratory, larvae were identified as *Ae. aegypti* by the relative size of the siphon and their movement compared with the other most commonly found Culex species (Consoli and de Oliveira, 1994). *Limatus* larvae were differentiated by the characteristics on the eighth tergite (Consoli and de Oliveira, 1994). All larval samples were cross-checked with the entomology collection sheets provided by the field team. Pupae were counted and placed in plastic emergence vials, $\leq$30 per vial and labeled with the house, container code, and date. Each subsequent day, emerged adults were collected and placed in a -20°C freezer. After 30 min to 1 h, their species was identified, counted by sex, and data were recorded on the entomology collection sheet.

Data Management

A geographic information system (GIS), using ARC/INFO and ArcView software (Environmental Systems Research Institute, Inc., Redlands, CA), was developed for the city of Iquitos. A base map of city blocks in the form of AutoCAD files was obtained from the Peruvian Navy, which they created by digitizing ortho-corrected 1995 aerial photographs. The coordinate system and datum used were Universal Transverse Mercator and WGS-84, respectively. The AutoCAD files were converted to ARC/INFO export files and all polygons (city blocks) were closed using standard ARCEdit procedures. Files were then imported into ArcView and converted to shape files.

We then divided city blocks into individual housing lots that were identified by painted codes. The front end of each house lot was measured and recorded along with the house code and street address on a rough sketch of each block. Based on maps constructed in the field, each digital block in the GIS was split into lots of

appropriate width using the measuring tool in ArcView. Lot length was estimated. Lot geometric centroids were then added to each individual lot and assigned a unique project code that was included on all subsequent survey forms. Construction of maps with resolution to the level of household lots allowed all entomologic data from the four surveys to be joined to geographic coordinates via house codes. Centroids allowed for spatial analysis to be performed from the level of the individual household upwards.

Analysis of the Data

Spatial patterns of four variables were examined (adult *Ae. aegypti*, pupae, all water-holding containers, and water-holding containers positive for larvae and pupae). Variables were explored by identifying the spatial distribution of each of the variables for each of the two time periods. Our study focused on (1) each of the two neighborhoods as a whole, (2) the magnitude of each variable in each household for each neighborhood, and (3) the presence or absence of a variable in a household for each neighborhood. Global K-functions, point and weighted, were used to identify clustering for (1) and the local statistic, G_i^*, was used for (2). These statistics are some of the suite of spatial statistical programs available as part of the Point Pattern Analysis (PPA) program. The program was developed by Arthur Getis with assistance from Laura Hungerford, Dong-Mei Chen, and Jared Aldstadt. An online version is available at http://zappa.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi. For (3), we used chi-square tests to compare similarities and differences among the various patterns.

K-functions

Pattern models are based on the K-function work of Ripley (1981) and Getis (1984). The K-function describes the number of pairs of observations between a point, which is the center of a disk and other points that are distance d away. For a stationary, isotropic process, $\lambda(d)$ is the expected number of points within distance d of an arbitrary point. The estimator of λ is N/A where N is the number of points in the study area A.

The estimator of $K(d)$ is

$$\hat{K}(d) = A/N^2 \Sigma_i \Sigma_j u_{ij}^{-1} I_d(d_{ij} \leq d), \quad i \neq j, \tag{15.1}$$

where d_{ij} is the distance between the ith and jth observed points and $I_d(d_{ij} \leq d)$ is an indicator function that is 1 if d_{ij} is less than or equal to d and 0 otherwise. For a circle centered on i passing through point j, u_{ij} is the proportion of the circumference of the circle that lies within A. When d_{ij} is less than the distance from i to one or more borders of the study area, u_{ij} is 1. The "border correction" makes $\hat{K}(d)$ an approximately unbiased estimator of $K(d)$ provided that d is less than the

circumference of A. A square-root scale makes the function linear and stabilizes the variance. Thus, we have

$$\hat{L}(d) \equiv \sqrt{\hat{K}(d)/\pi} \tag{15.2}$$

which is the estimator of $L(d) \equiv \sqrt{K(d)/\pi}$. The mean of $L(d)$ is d and the approximate variance is $\frac{1}{2}(\pi N^2)$ (Ripley, 1979a). The expectation of $L(d)$ given the hypothesis of complete spatial randomness (CSR) is d. CSR is a homogenous planar Poisson process where all points are independent of all other points and all locations are equally likely to contain a point. For CSR, a plot of $\hat{L}(d)$ against d on similarly scaled axes yields a 45° line beginning at the natural origin. A clustered pattern occurs when $\hat{L}(d)$ is greater than d and a dispersed pattern can be identified when $\hat{L}(d)$ is less than d. In the spirit of a exploratory diagnostic tool, statistical significance at the $P \leq 0.05$ level is assumed to exist when the observed $\hat{L}(d)$ function falls outside of an envelope containing 19 permutations of the location of the N objects where each permutation is based on CSR. $\hat{L}(d)$ is usually calculated for a series of distances d.

Instead of considering each point as a nominal scale variable, points can be weighted according to some measure of size or intensity (Getis, 1984),

$$\hat{L}_w(d) = [\{A\Sigma_i\Sigma_j u_{ij}^{-1} I_d(d_{ij} \leq d)x_i x_j\}/\{\pi[(\Sigma_i x_i)^2 - \Sigma_i x_i^2]\}]^{1/2}, \quad i \neq j, \tag{15.3}$$

where X is a random variable having values x for adult mosquitoes in houses at sites i. Equation (15.3) is the estimator for $L_w(d)$, which is equal to $E[\hat{L}_w(d)]$. In the cases discussed in this paper, the weights are in turn numbers of adult mosquitoes, pupae, water-holding containers, and positive containers. For each x_i, there are $(N-1)$ values x_j. In this case, the numerator of $\hat{L}_w(d)$ represents the product of the pairs of values $x_i\, x_j$ within distance d of each x. The denominator is scaled such that if all x are of equal value, then $\hat{L}(d)$ will be approximately equal to $\hat{L}_w(d)$. Thus, (15.3) represents a measure of clustering or dispersion identified in (15.2). If the number of adult mosquitoes, for example, is independently distributed within the plots of houses, $\hat{L}(d)$ will be approximately equal to $\hat{L}_w(d)$. Upper and lower significance boundaries for $\hat{L}_w(d)$ can be determined by a permutation procedure in which the various observed values for number of adult mosquitoes, x_i, are permuted among the house locations a specified number of times.

We also explored the increments to $\hat{L}(d)$ and $\hat{L}_w(d)$ observed for each equal increase of distance. In a CSR pattern of adult mosquitoes, these successive values will be the same for each equal increase of d. The focus is on the noncumulative properties of these pattern indicators. When the change in $\hat{L}(d)$ is greater or less than the change in $\hat{L}_w(d)$ for a given distance band, the adult mosquitoes are less concentrated or more concentrated, respectively, than that expected in the observed pattern, no matter how clustered the pattern of houses. That is, the number of adult mosquitoes is not randomly distributed among the houses. In essence, we compare $\Delta\hat{L}(d)$ with $\Delta\hat{L}_w(d)$ for a given small change in d.

Table 15.1 Summary of clustering statistics

Test	Purpose	Scale	Cut-off for statistic
$\hat{L}(d)$	To identify the existence of clustering for a 1/0 variable in a neighborhood	d	19 simulations of random occurrence within neighborhood (0.05 level)
$\hat{L}_w(d)$	To identify clustering of a weighted variable in a neighborhood	d	99 simulations of random occurrence within eligible locations of variable (0.01 level)
$G_i^*(d)$	To identify individual observations of a variable who are members of clusters	Z	$> +2.575$ (0.01 level)

$G_i^*(d)$ Statistic

In addition to $L(d)$, we used the local statistic, G_i^* (Ord and Getis, 1995), to identify individual members of clusters. For G_i^* we take each house as a center, one at a time, and search the nearby area for occurrences of more or fewer adult mosquitoes than expected. In this way, specific houses are identified as members or non-members of clusters. This statistic is written as

$$G_i^*(d) = [\Sigma_j w_{ij}(d)x_j - W_i^*\bar{x}]/[s\{[NS_{1i}^* - W_i^{*2}]/(N-1)\}^{1/2}], \quad \text{all } j, \ (15.4)$$

where $w_{ij}(d)$ is the i, jth element of a one/zero spatial weights matrix with ones if the jth house is within d of a given ith house; all other elements are zero; $W_i^* = \Sigma w_{ij}(d)$, where w_{ii} is included, and $S_{1i}^* = \Sigma w_{ij}^2$ (all j). The mean of the adult mosquitoes in houses is $\bar{x}$ and s is the standard deviation. The value of $G_i^*(d)$ is given in normal standard deviates. Note that this statistic has as its expectation, $W_i\bar{x}$, which controls for the number of houses within d of each house. Note, too, that $G_i^*(d)$ is 0 in a pattern where adult mosquitoes are randomly distributed within d of house i. For this study, we arbitrarily define values greater than 2.575 (the 0.01 level of confidence) as representing houses which are members of clusters of adult mosquitoes. The statistics used in the analysis and the test criteria are summarized in Table 15.1.

15.3 Results

We begin the explanation of results from our study by focusing on one neighborhood, Maynas, using data from the initial survey a. We first consider the general, neighborhood (global) spatial pattern of adult mosquitoes and then focus on the pattern of the numbers of *Ae. aegypti* in individual houses (local) followed by an analysis of the presence or absence of adult mosquitoes in households. Next we examine the same processes for immature mosquitoes. Finally, we compare the four entomologic variables in the two neighborhoods and two time periods.

Table 15.2 $L(d)$ values for distances 10–100 m for houses and adult mosquitoes in Maynas a*

Distance (m)	Houses	Adult mosquitoes	House increment	Adult increment
10	16.33	22.86	16.33	22.86
20	27.13	36.79	10.80	13.93
30	38.70	50.58	11.57	13.79
40	52.85	61.13	14.15	10.55
50	65.67	74.24	12.82	13.11
60	76.70	83.94	11.03	9.70
70	88.03	93.71	11.33	9.77
80	100.98	104.12	12.95	10.41
90	111.77	113.10	10.79	8.98
100	122.19	120.57	10.42	7.47

* i does not equal j

Neighborhood Pattern Analysis

The results of the K-function analysis for adult *Ae. aegypti* in Maynas in time period a are shown in Table 15.2. Adult mosquito clustering occurs if values of $\hat{L}(d)$ are higher not only than adult mosquitoes distributed at random in the Maynas neighborhood for a given distance (i.e., d), but also higher than the $\hat{L}(d)$ value for the pattern of houses at that same distance. Clearly, it is not enough that adult mosquitoes were spatially concentrated at the same rate as the spatial concentration of houses. Note that in column 3 in Table 15.2, the $\hat{L}_w(d)$ value for adult mosquitoes at 10 m is 22.86, which is quite a bit higher than the 10.00 (random expectation) shown in column 1. However, houses were much more clustered than random (16.33 vs. 10.00 at 10 m). Even so, adult mosquitoes were more clustered than houses. In addition, using 19 permutations to identify the range of possible values for adult mosquitoes among houses (at the 0.05 level), we find that adult mosquitoes at 22.86 fall outside of that range (low of 11.88 to high of 19.10) at 10 m. This gives strong statistical evidence that adult mosquitoes were clustered in the Maynas neighborhood during time period a. Clustering is at the 10-m level; thus, we can conclude that there is clustering around houses to at least 10 m distant.

Notice that in column 2 of Table 15.2, as distance increases to 20, 30 m, and so on, the $\hat{L}(d)$ values for houses increase at a rate that is not dissimilar from random expectation. This means that although houses are closely spaced at short distances, there is little or no increase in clustering as distance increases. The $\hat{L}_w(d)$ value for adult mosquitoes shown in column 3 at 20 and 30 m, however, increases at a slightly higher rate than houses (column 5 vs. column 4), indicating a continuing of the clustering identified at 10 m to at least 30 m. This pattern of increase changes by 40 m (the increment is 10.55, less than the house increment of 14.15) indicating an end to the increase in clustering. That is, beyond 30 m, any further clustering of adult mosquitoes corresponds to clustering of houses. We conclude that adult mosquitoes cluster heavily at nearest house distances and moderately to approximately 30 m. In

Table 15.3 $L(d)$ values for distances 10–100 m for houses and adult mosquitoes in Maynas a*

Distance (m)	Houses	Adult mosquitoes	House increment	Adult increment
10	21.44	39.30	21.44	39.30
20	30.46	48.65	9.03	9.35
30	41.08	59.67	10.62	11.02
40	54.60	68.75	13.51	9.08
50	67.06	80.52	12.47	11.77
60	77.88	89.46	10.81	8.94
70	89.04	98.60	11.16	9.14
80	101.83	108.44	12.79	9.84
90	112.52	117.00	10.69	8.56
100	122.87	124.17	10.34	7.17

* i may equal j

Maynas, the mean house width was 7 ± 3 m; thus, adult clusters could extend to about two households on each side.

We altered (15.1) and (15.3) to include houses themselves; that is, we allowed i to equal j (Table 15.3; see Getis (1984) for an explanation of the methodology). Our focus now is on houses and their neighbors rather than neighboring houses only. In this circumstance, the clustering of houses (column 2) is inflated to include not only near neighbors at 10 m, but also the houses themselves. The original value of 16.33 at 10 m now increases to 21.44 for houses indicating that in this view, houses are more clustered than was indicated previously (an increase of 31%). More importantly, however, are the results when adult mosquitoes within houses are taken into account. Here the value at 10 m increases to 39.30 from 22.86, an increase of 72%. The implication is that adult mosquitoes are heavily clustered within houses. Note also that as distance increases, the increment to houses and adult mosquitoes is approximately 10, indicating that there is a cessation of clustering beyond 10 m. Again there is additional, albeit weak clustering up to 30 m because the increase in the mosquito value is higher than that for the houses at 20 and 30 m. These results taken together with the earlier ones unequivocally indicate that adult mosquitoes cluster heavily within or among nearest neighboring houses. In addition, there is evidence of further, albeit minor, clustering as far as 30 m. The clustering within houses in the Maynas neighborhood quantitatively overwhelms this further clustering.

Household Pattern Analysis by Numbers of Adult Mosquitoes

After it was evident that there was short distance clustering of adult mosquitoes in Maynas a, we identified the exact houses that could be considered as members of clusters. First, we considered the actual numbers of adult mosquitoes in each house in Maynas a (Fig. 15.2). If clustering was within households, the G_i^* statistic will be above $+2.575$ at short distances, say 1 m at the 0.01 level of statistical significance. If clustering continues to near neighbors within 10 m of a house, the value of G_i^*

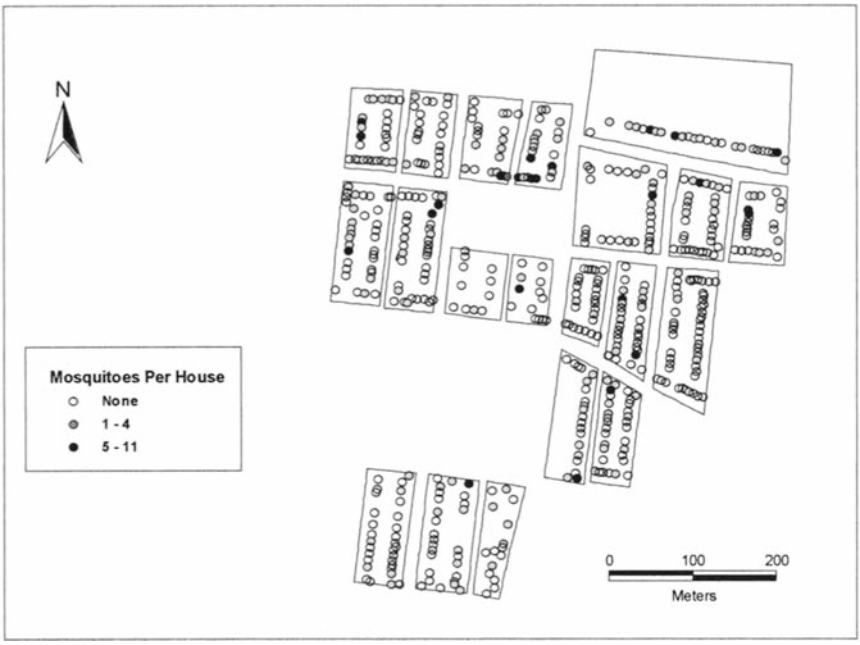


Fig. 15.2 Mosquitoes per house in the Maynas a study

will be higher at 10 m than at 1 m. If values of G_i^* do not increase with increases in distance, then whatever clustering existed at the shorter distance ceases to exist at longer distances. The houses that are members of significant clusters at 1, 10, 20, and 30 m are shown in Fig. 15.3. Note that of the 528 houses in Maynas during time period a, 35 (6.6%) are members of statistically significant clusters of adult mosquitoes. Of the 35, 10 exhibit clustering with near neighbors beyond the house itself. Of these 10, seven show clustering to 10 m, two to 20 m, and one to 30 m. This result reinforces the notion that adult mosquitoes tend to cluster in single households with a modest spread to as far as 30 m.

Pattern of Houses Infested with Adult *Ae. aegypti* (<1 Mosquito)

Figure 15.4 is a map of the presence of one or more mosquitoes in households. One hundred sixty-four (31.1%) of the houses had one or more adult mosquitoes present; however, only 35 of them (21.3%) were members of statistically significant clusters. This indicates that clusters were made up mainly of household concentrations, and that 79.7% of the households with mosquitoes were spread about in a random pattern among all households.

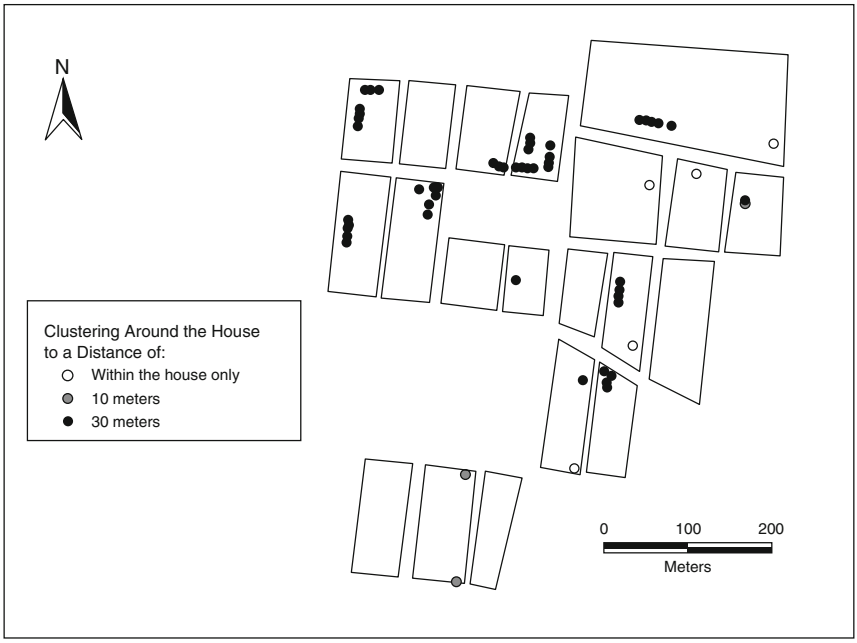


Fig. 15.3 Clusters of *Aedes aegypti* adults in the Maynas a study based on the number of mosquitoes in houses

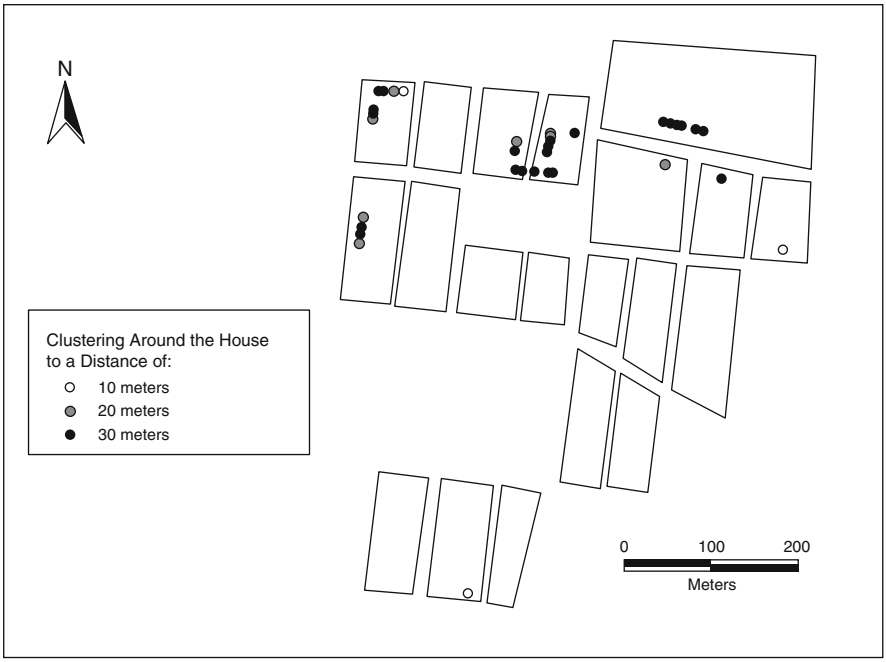


Fig. 15.4 Clusters of *Aedes aegypti* adults in the Maynas a study based on presence or absence of mosquitoes

Neighborhood Pattern Analysis of Immature Mosquitoes

Results in Tables 15.4 and 15.5 allow for the comparison of K-function values for water-holding containers, positive containers, and pupae with house and adult mosquito patterns in the Maynas neighborhood [(15.1) and (15.3)]. The $d = 10\,\text{m}$ row in Table 15.4 shows, as before, that adult mosquitoes cluster more so than houses (22.86–16.33), but the pattern of water-holding containers and positive containers is more nearly like the pattern of houses (16.25–16.33 and 15.40–16.33). Thus, there is evidence of no clustering for these variables. In the case of pupae, however, there is a significantly lower value (12.03), indicating that pupae do not cluster beyond the household and, in fact, are dispersed rather evenly throughout the neighborhood. However, when we allow i to equal j (Table 15.5), pupae increase from 12.03 to 56.13, an extremely high and statistically significant value.

Table 15.4 $\hat{L}(d)$ values for distances 10–100 for houses, adult mosquitoes, pupae, water-holding containers, positive water-holding containers in Maynas a*

Distance (m)	Houses	Adult mosquitoes	Pupae	Containers	Positive containers
10	16.33	22.86	12.03	16.25	15.40
20	27.13	36.79	22.73	27.43	27.03
30	38.70	50.58	36.82	40.03	37.66
40	52.85	61.13	46.40	54.16	51.88
50	65.67	74.24	56.15	66.86	64.55
60	76.70	83.94	70.50	78.42	76.20
70	88.03	93.71	80.66	90.19	86.40
80	100.98	104.12	92.23	102.57	99.59
90	111.77	113.10	102.49	113.17	110.28
100	122.19	120.57	110.86	123.36	119.91

* i does not equal j

Table 15.5 $\hat{L}(d)$ values for distances 10–100 m for houses, adult mosquitoes, pupae, water-holding containers, positive water-holding containers in Maynas a*

Distance (m)	Houses	Adult mosquitoes	Pupae	Containers	Positive containers
10	21.44	39.30	56.13	23.44	29.05
20	30.46	48.65	59.26	32.18	36.53
30	41.08	59.67	65.77	43.36	44.93
40	54.60	68.75	71.41	56.61	57.3
50	67.06	80.52	77.91	68.74	68.92
60	77.88	89.46	88.51	79.93	79.88
70	89.04	98.60	96.56	91.49	69.60
80	101.83	108.44	106.14	103.62	102.31
90	112.52	117.00	114.91	114.12	112.68
100	122.87	124.17	122.22	124.26	122.07

* i may equal j

This indicates that pupae cluster strongly within houses, but households infested with pupae are dispersed rather evenly throughout the neighborhood (Table 15.4).

Because water-holding container spatial data are similar to the house location data (Tables 15.4 and 15.5), we conclude that water-holding containers are ubiquitous in Maynas. That is, nearly all houses have water-holding containers. Conversely, containers positive for pupae and/or larvae are more concentrated in some houses than others and infested houses are dispersed evenly throughout the neighborhood.

Continuing on to 20, 30 m, and further (Tables 15.4 and 15.5), only pupae act differently than containers and positive containers. For both of the container variables, increases mirror those of houses, reinforcing our earlier results that show ubiquitous occurrences of these variables. Pupae values (Table 15.5), however, increase at a much slower rate than houses after 10 m, indicating that households infested with pupae are less common than households with water-holding containers or positive containers, and that the spatial pattern of pupae is characterized by strong clustering within households.

Household Pattern Analysis of Non-adult Mosquitoes

Our G_i^* statistic results show that there is a lack of statistically significant clustering beyond households for container and immature mosquito variables. In the case of pupae, there were 18 households exhibiting clustering with no clustering beyond the household. Of the 24 houses with clusters of containers, only two were clustered to a neighboring distance of 10 m. For positive containers, 23 houses exhibit clustering, but only three of those were clustered beyond the household, 2–10 m, and 1–20 m.

Patterns of Pupae: Presence or Absence in Houses

In this analysis, the concern is less with numbers of pupae in houses and more with their spatial occurrence in houses. Data in Fig. 15.5 were derived from a G_i^* analysis that assigned a 1 to houses with one or more pupae present and 0 for the absence of pupae. We found that 18 (3.4%) of the 528 houses can be considered as members of clusters at the 99% level of confidence. There are two distinct clusters: one in the middle block in the south and a smaller cluster in the north. These concentrations raise the question of the relationship of the location of pupae to adult mosquitoes.

Comparison of Entomologic Spatial Patterns in Maynas a

Does the pattern of adult mosquito clusters correspond to the patterns of the other variables? We answer this question in three ways. First, we consider the overlap of clusters among the four variables. Second, we note the presence (one or more) of each variable occurring simultaneously in individual houses. Third, we focus on the number of water-holding containers, positive containers, pupae, and adult mosquitoes in households

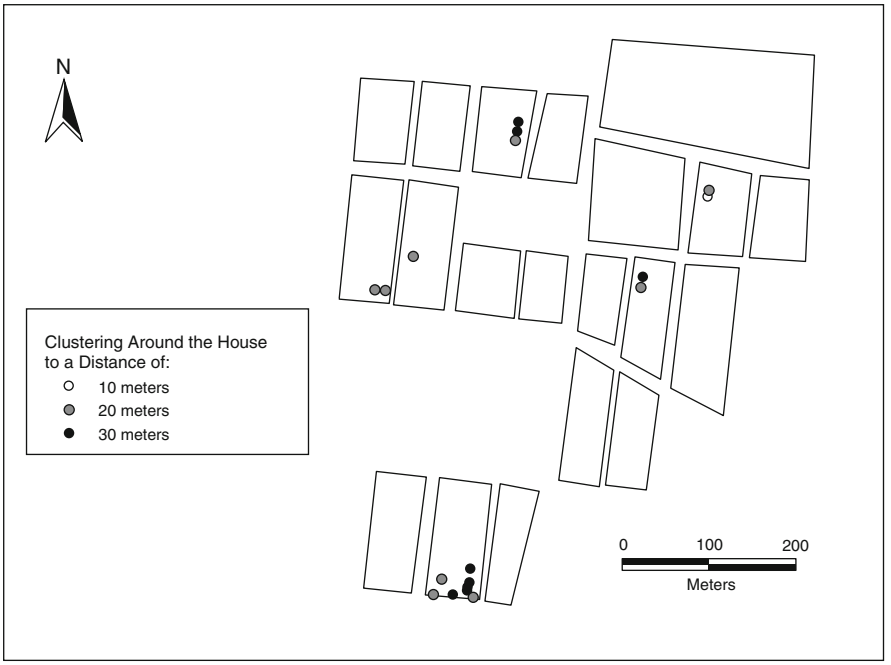


Fig. 15.5 Clusters of *Aedes aegypti* pupae in the Maynas a study based on presence or absence of pupae

Table 15.6 Number of members of clusters in Maynas and Tupac Amaru in time periods a and b

	Maynas	Tupac Amaru
Houses	528	481
Adults in time period a	35	40
Adults in time period b	27	32
Pupae in time period a	18	18
Pupae in time period b	4	24
Adults in a and b	7[a]	2
Pupae in a and b	0	6[a]
Adults in a and pupae in b	0	1
Pupae in a and adults in b	2	3
Adults in a and pupae in a	3	4[b]
Adults in b and pupae in b	0	0

[a] Significant at the 0.01 level
[b] Significant at the 0.05 level

Association Among Clusters

In Table 15.6 we see, as before, that of the 528 houses in Maynas, 35 were members of clusters of adult mosquitoes and 18 were members of clusters of pupae in time period a. Only three houses were constituents of both clusters, a non-statistically significant result at the 0.05 level ($\chi^2 = 1.60$, degrees of freedom $= 1$, Yates'

Table 15.7 One or more adult mosquitoes and/or pupae present in houses in Maynas and Tupac Amaru in time periods a and b

	Maynas	Percent	Tupac Amaru	Percent
Houses	528		481	
Adults in time period a	164	31.06	87	18.09
Adults in time period b	151	28.60	92	19.13
Pupae in time period a	155	29.36	86	17.88
Pupae in time period b	134	25.38	65	13.51
	Maynas		Tupac Amaru	
	Observed	Expected	Observed	Expected
Adults in a and b	67	47[a]	20	15
Pupae in a and b	70	39[a]	25	11[a]
Adults in a and pupae in b	53	42[b]	14	11
Pupae in a and adults in b	50	44	20	15
Adults in a and pupae in a	66	48[a]	25	14[a]
Adults in b and pupae in b	50	38[a]	15	11

[a] Significant at the 0.01 level
[b] Significant at the 0.05 level

correction for small expectations). There was not a significant correlation between pupal and adult abundance within household or neighborhood clusters detected during the same survey.

Association Among Households Having One or More of Each Variable Present

The analysis summarized in Table 15.7 reveals the overlap of households that have as few as one mosquito or one pupae present. Note that of the 528 houses in Maynas, 164 had at least one mosquito present and 155 had at least one pupae present in time period a. Expectation from a chi-square two-by-two contingency test indicate that the two types of occurrence come together in households 48 times. A total of 66 households were infested with both pupae and adults, demonstrating that the presence of these two variables are not independent ($P < 0.01$).

Association of Water-Holding Containers and Adult Mosquitoes and Pupae

Because there are water-holding containers in every household in Maynas, we compared the relative abundance of positive containers, pupae and adult mosquitoes. Table 15.8 shows the results of Spearman's rank correlation test where the number of water-holding containers per household were ranked from 1 to 14. Ranks 15 and 16 were made up of 15–19 and 20–35 containers, respectively. The final two ranks were grouped because of the few numbers of observations at these high levels. The

Table 15.8 Spearman's rank correlations of the number of containers per house with the number of mosquitoes and pupae per house

Location of containers	Mosquitoes	Pupae
Maynas a	$+0.615^{a}$	$+0.487$
Maynas b	$+0.682^{b}$	$+0.594^{a}$
Tupac Amaru a	$+0.284$	$+0.486$
Tupac Amaru b	-0.199	$+0.481$

[a] Significant at the 0.05 level
[b] Significant at the 0.01 level

Table 15.9 $\hat{L}(d)$ values for 10 m for Maynas and Tupac Amaru for time periods a and b*

	Maynas a	Maynas b	Tupac Amaru a	Tupac Amaru b
Houses	21.44	21.44	25.00	25.00
Mosquitoes	39.30	51.06	76.64	51.08
Pupae	56.13	71.42	80.34	76.14
Containers	23.44	23.43	27.68	27.87
Positive containers	29.05	31.00	38.56	44.30

* i may equal j

mean number of adult mosquitoes per house was ranked for each container level. The result was a moderately high positive correlation for adults ($+0.615$, $P < 0.05$), and a modest correlation for pupae ($+0.487$, not significant). Our analysis indicates that elevated numbers of water-holding containers in houses increase the likelihood for elevated numbers of adult mosquitoes and/or pupae to be present.

Maynas Vs. Tupac Amaru

Although non-spatial measures of *Ae. aegypti* population densities decreased in both sites in the second surveys, they were higher in both surveys in Maynas than in Tupac Amaru. For example, the house index (percentage of surveyed houses with $\geq$1 positive container) was 45% in Maynas a, 38% in Maynas b, 29% in Tupac Amaru a, and 23% in Tupac Amaru b.

Clustering patterns of adult mosquitoes and pupae were consistent among the four surveys, but the level of clustering was greatest during the first Tupac Amaru survey. Table 15.9 shows the $\hat{L}(d)$ values (i may equal j) for each of the four surveys for 10 m. Houses in Tupac Amaru were slightly more clustered than in Maynas (25.00–21.44). Note also that in both neighborhoods water-holding containers are distributed much the same as were houses, but positive containers tend to cluster. Maynas with 29.05 and 31.00 in the two time periods are approximately 8–10 L units higher than the pattern of houses. Tupac Amaru with 38.56 and 44.30 are about 13–19 units higher than the pattern of houses. This implies that positive containers were more clustered in Tupac Amaru than Maynas, which may be a reflection of lower infestation rates in Tupac Amaru. Nevertheless, in both sites the level of clustering was relatively low.

Time Period *a* Vs. *b*

The objective of carrying out back-to-back surveys in two sites was to account for variability in collector aptitude; a commonly cited limitation of entomologic surveys (Reiter and Gubler, 1997). Despite only three weeks separating surveys, the number of water-holding containers and immature mosquito indices decreased between the two sampling periods. Reasons for this are not known, but the possibility that our survey methodology affected immature populations must be considered. During the first survey, small containers not used for water storage were tipped over and homeowners may have cleaned or drained larger containers that our field team identified as being infested with larvae or pupae. Following a reduction in immature mosquitoes, we would expect a decrease in emergence of adults and in turn a measurable reduction in adult population density. Curiously, a reduction in adult density was only detected in Tupac Amaru, where the number of adults per household decreased from 0.4 to 0.3. In Maynas, the number of adult *Ae. aegypti* per household was 0.7 in both surveys. In the second surveys the number of water-holding containers decreased by 13% in Tupac Amaru compared with only 3% in Maynas.

Overall Patterns of Adult Mosquito and Pupae Household Clustering

Table 15.6 shows the number of houses that were members of statistically significant clusters of pupae and adult *Ae. aegypti*. The number of houses included in clusters for pupae in Maynas decreased from 18 to 4 from time period a to b. Interestingly, the location of adult clusters changed between the two surveys. Twenty-eight households were members of adult clusters in the first Maynas survey that were not members of clusters in the second, a statistically significant finding that was not the case in Tupac Amaru. Only seven households were members of adult clusters in both Maynas surveys. Twenty Maynas households were members of clusters in the second but not first survey. The same type of result, changing cluster locations, was evident with member houses of pupae clusters. In Maynas none of the houses were members of pupae clusters in both surveys, whereas six households were part of pupae clusters during both time periods in Tupac Amaru (Table 15.6). This result indicates that the spatial distribution of entomologic data varies greatly within short periods of time.

Association Among Households Having One or More of Each Variable Present in Each Neighborhood over Time

Although clusters of positive containers, pupae, and adult mosquitoes identified by G_i^* were not consistent with time, *Ae. aegypti* infestations of individual households were clearly a risk factor for future infestation. That is, there is evidence of repeat offenders. Table 15.7 shows the number of houses observed to be infested with either

pupae or adults in survey a, survey b, or both. Pupae in a are again found in the same houses in b in both neighborhoods between 29% and 45% of the time, a statistically significant result. The implication is that for unknown reasons mosquitoes are more likely to lay eggs in containers on some house lots than others. Another risk factor for infestation is the number of water-holding containers in a household. Results in Table 15.8 indicate that there is a tendency for houses in both neighborhoods and both time periods to contain more pupae when more water-holding containers are present.

15.4 Discussion

Historically, entomologic surveillance for dengue was dominated by the use of larval surveys, in large part because *Ae. aegypti* control grew out of an eradication paradigm that promoted complete, thorough and repeated coverage of infested areas (Reiter and Gubler, 1997). In 1994, however, the Pan American Health Organization declared *Ae. aegypti* eradication an unattainable goal and promoted *Ae. aegypti* control, which they defined as the "cost effective utilization of limited resources to reduce vector populations to levels at which they are no longer of significant public health importance" (PAHO, 1994). Although this recommendation intuitively makes sense, it is not specific enough for public health officials to use as a guideline to control dengue. For example, experience with yellow fever and recent computer simulation estimates indicate that entomologic thresholds below which dengue transmission will cease are low (Reiter and Gubler, 1997; Focks and Chadee, 1997; Focks et al., 1995), but threshold values have not been systematically derived or tested (Reiter and Gubler, 1997). Empirically defined thresholds will require prospective, longitudinal studies in which investigators simultaneous monitor relationship between dengue virus transmission in a human cohort and *Ae. aegypti* population densities. Interpretation of data from those kinds of studies will require careful consideration of (1) spatial auto-correlation and scale in statistical analyses; (2) the most appropriate measure of entomologic risk-should absolute numbers or indices be measured and what life stage of the mosquito provides the best estimate for risk of human dengue virus infection; and (3) survey design, including the extent of data collection. Our study contributed to an improved understanding for each of these issues.

The lack of spatial structure for immature forms of *Ae. aegypti* supports recommended vector surveillance strategies where standard sample size calculations and resource limitations are used to determine in a systematic way the number of houses to be sampled, typically every ith house. Our K-function analysis indicates that individual households are the appropriate spatial unit for entomologic surveys. From a temporal perspective because water-holding containers were ubiquitous in Iquitos, all households are at risk of infestation over any considerable period of time. Our results, however, imply that as the number of containers on a premise increases so does the risk of *Ae. aegypti* pupae and adult infestations. In other words, positive

containers and pupae cluster within individual households, but the location of clusters changes through time. Biologically this makes sense. Infestation of a household is largely a function of container management practices by the occupants of the property and the ecology of *Ae. aegypti* egg-laying behavior. We did not detect larger scale structure that might have been affected by other factors (data not presented or discussed in this paper) such as the availability of piped water, local temperature, rainfall patterns, or garbage disposal.

Identification of "key premises" or households that are superproducers of *Ae. aegypti* has been proposed as a way to streamline surveys (Tun-Lin et al., 1995). The idea is that the presence of pupae or adults during an initial survey is a significant risk factor for observing the same life stage at the same location during subsequent surveys. If we adopt the notion of controlling key premises as a way of reducing but not eliminating *Ae. aegypti* populations, the fundamental need to refine our understanding of entomologic thresholds is reinforced. Until we quantitatively define the relationship between mosquito density and risk of virus transmission, we cannot predict the effect that eliminating key premises will have on the risk of human infection and disease. For example, eliminating key premises may not reduce the adult mosquito population below the threshold density and, depending on the nature of the relationship between virus transmission and vector density, the pattern of human infections could continue unabated. Interestingly, the transient pattern of immature mosquito cluster locations observed in our study indicates that even if key premises can be identified and eliminated there may still be a sufficient number of *Ae. aegypti* to sustain dengue virus transmission. It should be noted, however, that because Iquitos has a relatively low percentage *Ae. aegypti* production in permanent water holding containers, our results may be site specific. The same kind of thorough examination may need to be carried out (large sample sizes and spatial analysis) at other locations.

Although small, there was significant spatial structure of adult mosquito populations compared with pupae and positive containers. Adults cluster most to distances of approximately 10 m and to a lesser extent out to 30 m, which could include neighboring houses. This finding is consistent with our conclusion to use the household as the basic unit of entomologic surveillance. It also superficially supports focal insecticide treatments for dengue control, a practice in which households are treated with insecticides within a 50–100 m radius of the residence of a detected dengue case (PAHO, 1994). There are, however, at least three shortcomings to focal treatments that extend beyond spatial patterns of adult *Ae. aegypti*. The approach does not take into account (1) the time delay between when a person is infective to mosquitoes and they are detected as being clinically ill with dengue, (2) that infected people can transport virus rapidly over greater distances than flying infected mosquitoes, and (3) that viremic people can have an inapparent infection or may not seek medical assistance, the homes and surrounding areas of many people infective to mosquitoes will not be sprayed.

Our statistical approach corroborates results from mark-release-recapture experiments on the dispersal of adult *Ae. aegypti*. Most researchers have concluded that the typical flight range of this species is short (<100 m). Rodhain and Rosen (1997)

stated that spontaneous dispersal of adult *Ae. aegypti* averages from 30 to 50 m per day, so that females are rarely expected to visit more than two or three houses in their lifetime. The length of an *Ae. aegypti* lifetime is difficult to estimate, but is generally believed to range from 8 to 16 days (Focks et al., 1993). Ordonez and others (1997) reported minimum and maximum daily flight distance for *Ae. aegypti* of 8 and 120 m, respectively, with a mean of 30.5 m. In a Kenyan village, McDonald (1977) found that most adult *Ae. aegypti* dispersed to less than 20 m and the majority of those recaptured were collected in the same house where they were released. Edman and others (1998) similarly collected most of their recaptured *Ae. aegypti* in Puerto Rico from their release house. In Kenya, Trpis and Hausermann (1986) reported 57 m as the mean daily flight distance for females, with a maximum dispersal of 154 m. Sixty percent of their recaptured females were collected in 11 houses that were within 50 m from their release point. Our spatial analysis agrees with the preponderance of evidence that in a place such as Iquitos most adult *Ae. aegypti* do not fly far from the container where they developed as larvae and pupae.

Spatial referencing of our adult survey data and application of statistical tools, such as K-function and G_i^*, provided insights into adult dispersal behavior that help explain patterns of human dengue infections. We propose that over short periods of time the restricted flight range and frequent blood-feeding behavior of *Ae. aegypti* (Scott et al., 2000b) are underlying factors in the clustering patterns of human dengue infections. In addition to the studies cited above on *Ae. aegypti* dispersal, several researchers have reported spatial and temporal clusters of clinically ill dengue patients in the same household or adjacent houses (Halstead et al., 1969; Waterman et al., 1985; Chan, 1985; Gubler, 1997). In the first spatial statistics analysis of this phenomenon, Morrison and others (1998) found that dengue cases reported within a three-day interval during an epidemic in Florida, Puerto Rico clustered up to 10 m. With regard to blood-feeding behavior, *Ae. aegypti* is know to frequently and preferentially imbibe human blood meals (Scott et al., 2000b; Harrington et al., 2001) and infected females can transmit dengue virus to as many as 20 consecutive hosts, one after another (Putnam and Scott, 1995). It is conceivable that a single or very few infected *Ae. aegypti* that remain in the same general area could bite and transmit virus to several susceptible family members or their immediate neighbors within a period of a few days.

Upon further investigation, we may discover that the extent to which infected humans are clustered is influenced by house construction and distribution. For example, households in our study area were small and often located close together; most were row houses with common walls. Although features of housing in Iquitos might facilitate *Ae. aegypti* movement, we do not expect that the tendency for adult females to disperse will be dramatically different at other locations. In Iquitos, water-holding containers were found in all households surveyed, something that is expected to decrease the probability of female dispersal (Edman et al., 1998).

Abundance of adult female mosquitoes should be the most appropriate measure of entomologic risk because they are in the life stage from which viruses are transmitted. Interestingly, in at least one previous study adult *Ae. aegypti* abundance was correlated with diagnosed dengue cases (Rodriguez-Figueroa et al., 1995). The

value of larval indices was recently challenged because their relationship with adult densities is questionable (Reiter and Gubler, 1997). Pupal indices are now being considered as alternatives to traditional larval indices (Focks and Chadee, 1997; Focks et al., 1993). Pupal indices are attractive for three reasons. First, it is theoretically possible to make absolute counts of their abundance, something that cannot be done for flying and difficult to capture adults. Second, pupal mortality is low. The magnitude of the pupal population should, therefore, be directly and relatively easily correlated with adult densities. Third, because the pupa is the life stage that directly precedes the virus-transmitting adult, pupae should be a more direct measure of transmission risk than larvae, which are a developmental step removed from adults.

Results from our spatial analyses, however, identified some limitations of pupal indices. The transient nature and high variability of containers positive for pupae can lead to misleading survey results, especially if the goal is to identify "key premises" and if only a single survey is carried out. Examination of spatial correlations among water-holding containers, larvae, pupae, and adults reveal significant correlations between life stages that are directly linked in their developmental sequence. For example, larval clusters correlated with pupal clusters and pupal with adults, but larval clusters were not correlated with adult clusters. This indicates that many containers exhibited a cohort effect. That is to say, cohorts of mosquitoes in a given container move in synchrony through the different stages of their life cycle without overlapping other cohorts. A noteworthy observation in that regard is that we did not consistently collect all stages of mosquitoes at the same time in the same household. This indicates that containers in Iquitos are not in equilibrium with the mosquito population. Instead houses are positive for a limited period of time as mosquitoes develop, disperse, and the household reverts to being negative. Other households subsequently become positive and the process repeats itself. In locations where positive containers are ubiquitous and permanent a different pattern of cluster spatial stability may emerge.

We conclude that pattern analysis can efficiently describe local *Ae. aegypti* populations and substantially aid in our understanding of dengue epidemiology and the development of dengue surveillance and control strategies. We argue that development of long-term entomologic risk assessment strategies requires thorough surveys of all mosquito life stages. Our results highlight the importance of scale when investigating the dynamics of dengue transmission. In Iquitos, the appropriate scale for assessing mosquito vector density is the household level at frequent time intervals.

This work is being extended with more extensive studies in additional areas of Iquitos, including an entire city study of the affinity that *Ae. aegypti* may have for particular types of water-holding containers and the relationship of various measures of mosquito density to human dengue infection. In addition, related work is underway in Thailand, which will allow comparison of concepts and processes described for Iquitos to results from an ecologically and epidemiologically distinct study area.

Acknowledgements We thank the residents of Maynas and Tupac Amaru, Iquitos, Peru for allowing us to work with them in their homes. We greatly appreciate support of the Loreto Regional Health Department, including Dr. Carlos Calampa, Dr. Jorge Reyes, Dr. Ruben Naupay, Dr. Carlos Vidal, Dr. Martin Casapia, and Dr. Moises Sihuincha, who have all facilitated our work in Iquitos.

Helvio Astete and Gerson Perez Rodriguez supervised the collection and processing of mosquitoes. Entomologic surveys were carried out by Jimmy Maykol Castillo Pizango, Rusbel Inapi Tamani, Juan Luiz Sifuentes Rios, Nestor Jose Nonato Lancha, Federico Reategui Viena, Victor Elespuru Hidalgo, Edson Pilco Mermao, Abner Enrique Varzallo Lachi, Fernando Chota Ruiz, Angel Puertas Lozano, Guillermo Inapi Huaman, and Manuel Ruiz Rioja. Jimmy Roberto Espinoza Benevides and Fernando Espinoza Benevides entered data into our database. Jose Elespuro Bastos checked data entry for accuracy.

Financial support: This study was supported by grant AI-42332 from the National Institutes of Health.

Chapter 16
Spatial Filtering and Missing Georeferenced Data Imputation: A Comparison of the Getis and Griffith Methods

Daniel Griffith

Abstract Spatial filtering first introduced independently by Getis and by Griffith is beginning to mature, with a third version now being developed by Legendre and his colleagues. Like the Getis formulation, this newest version is distance-based; like the Griffith formulation, it uses eigenfunctions, but extracted from a modified distance matrix – it is a mixture of the other two. Bivand (2002) comments that "the Getis filtering approach ... seems to admit prediction to new data locations .... The Griffith eigenfunction decomposition approach ... does not ...." Missing data prediction equations are presented for each of these two original formulations, and then compared with several popular datasets.

16.1 The Imputation Problem

The *E*stimation–*M*aximization (EM) algorithm (Dempster et al., 1977), an iterative procedure for computing maximum likelihood estimates when data sets are incomplete, is a useful device for helping to solve model-based small geographic area estimation problems. Flury and Zoppè (2000, p. 209) emphasize:

> it can not be stressed enough that the E-step does not simply involve replacing missing data by their conditional expectations (although this is true for many important applications of the algorithm).

But model-based small geographic area estimation problems desire just this type of imputation output from the algorithm. The purpose of this paper is to illustrate the use of the eigenvector- and G_i-statistic-based spatial filtering specifications to compute such imputations.

Descriptions of the EM algorithm may be found in Flury and Zoppè (2000), Meng (1997), and McLachlan and Krishnan (1997), among others. For only missing response variable values, where missing is at random or completely at random, the

D. Griffith
The University of Texas at Dallas, Richardson, TX, USA
e-mail: dagriffith@utdallas.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*,
Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_16,

EM algorithm can be implemented as a regression problem in the following way (see Yates, 1933):

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \begin{pmatrix} \mathbf{1}_o & \mathbf{X}_o \\ \mathbf{1}_m & \mathbf{X}_m \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{o,m} \\ -\mathbf{I}_{m,m} \end{pmatrix} (\mathbf{y}_m) + \begin{pmatrix} \epsilon_o \\ \mathbf{0}_m \end{pmatrix}, \tag{16.1}$$

where the subscript o denotes observed data, and the subscript m denotes missing data, $\mathbf{Y}$ is the vector of response values, $\mathbf{1}$ is a vector of ones, $\mathbf{X}$ is a matrix of covariates, $\mathbf{0}$ is a vector of zeroes, $\mathbf{I}$ is the identity matrix, $\mathbf{y}_m$ is the vector of missing values, α is the intercept term, β is the vector of covariate regression coefficients, and ϵ is the vector of independent and identically distributed normal random errors. Of note is that the estimates of $\mathbf{y}_m$ are conditional expectations, resulting in their corresponding residuals being zero.

The unknown $\mathbf{y}_m$ are subtracted from both sides of a conventional linear regression equation to obtain (16.1). This subtraction results in the $\mathbf{0}_m$ subvector appearing in the response variable, and the set of m indicator variables appearing in the right-hand side of the equation. This subtraction is why $\mathbf{I}_{m,m}$ has a negative sign.

Table 16.1 and Fig. 16.1 respectively tabulate and portray results obtained for selected example dataset EM algorithm results reported in the literature. The normality diagnostics appearing in Table 16.1 imply that the linear regression normal probability model seems reasonable to employ. The near-perfect alignment of estimates depicted in Fig. 16.1 confirms that (16.1) and the EM algorithm yield exactly the same imputation results; slight deviations from roughly the middle of the trend line are attributable to the use of multiple imputation results by Schafer. The bivariate linear regression equation describing the scatterplot in Fig. 16.1 is as follows:

Table 16.1 Sources of data for using (16.1) to construct Fig. 16.1

Data source	Page	n_o	n_m	# Co-variates	Parameters	$P(S-W)$
McLachlan and	49	8	2	1	$\hat{\mu}_2$, $\hat{\sigma}_{12}$, $\hat{\sigma}_{22}$	0.1667
Krishnan (1997)	53	7	2	2	$\hat{y}(-1,-1)$, $\hat{y}(0,-1)$	0.2205
	54	34	2	7	$\hat{y}_{23}$, $\hat{y}_{51}$	0.4388
	137	12	6	2	$\hat{\sigma}_{\hat{\mu}_2}$	0.5672
Little and Rubin (1987)	31	13	2	8	$\hat{\mu}_1$, $\hat{\mu}_2$	0.9894
	101	12	6	1	$\hat{\mu}_2$	0.6661
Schafer (1997)	34	19	5	1	$\hat{y}_{2,L_8}$, $\hat{y}_{2,L_9}$, $\hat{y}_{2,L_{10}}$, $\hat{y}_{2,L_{11}}$, $\hat{y}_{2,L_{12}}$	0.9973
	195	19	9	2	$\hat{y}_{3,2}$, $\hat{y}_{3,4}$, $\hat{y}_{3,5}$, $\hat{y}_{3,10}$, $\hat{y}_{3,13}$, $\hat{y}_{3,16}$, $\hat{y}_{3,18}$, $\hat{y}_{3,23}$, $\hat{y}_{3,25}$	0.2485
Montgomery and Peck (1982)	145	25	4	2	$\hat{y}_{26}$, $\hat{y}_{27}$, $\hat{y}_{28}$, $\hat{y}_{29}$	0.2711
http://missingdata.org.uk	www	9	1	1	$\hat{y}_{10}$	0.4244

Note: P(S-W) is the probability under the null hypothesis of normality of the Shapiro–Wilk normality diagnostic test statistic calculated for regression residuals

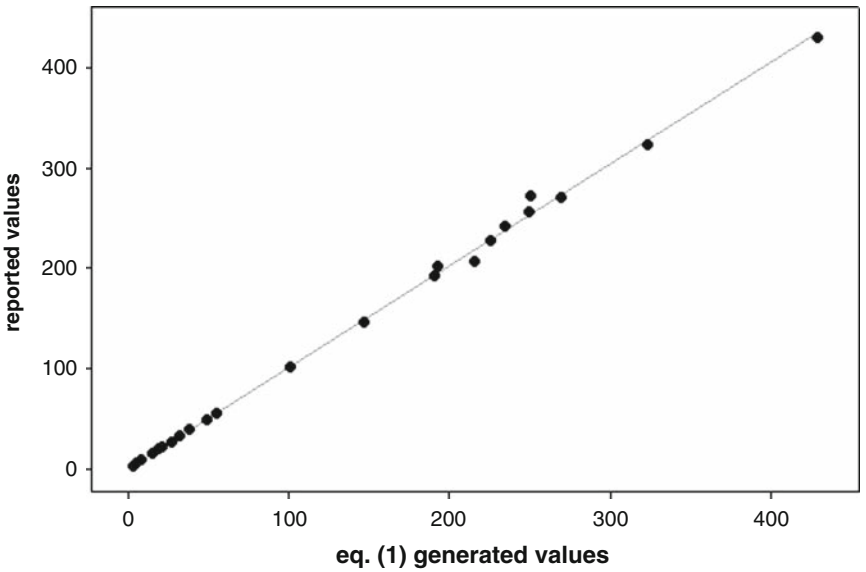


Fig. 16.1 Scatter plot of EM algorithm results: reported published vs. (16.1) generated estimates

$$\begin{array}{rcl} \text{reported} = & -0.15609 & + 1.01245 \times \text{equation}(16.1) + \text{e}, \\ & (t = -0.13) & (t = 1.66) \end{array}$$

which has an R^2 value of 0.998.

16.2 Interpolation as Missing Data Imputation

Haining et al. (1984) outline a spatial EM algorithm for estimating missing georeferenced variable values. Consider the following partitioned spatial covariance matrix:

$$\Sigma = \begin{pmatrix} \Sigma_{oo} & \Sigma_{om} \\ \Sigma_{mo} & \Sigma_{mm} \end{pmatrix} = \begin{pmatrix} V_{oo} & V_{om} \\ V_{mo} & V_{mm} \end{pmatrix}^{-1} \sigma^2,$$

where, as before, the subscript o denotes observed data, the subscript m denotes missing data. For a multivariate normal probability model, the maximum likelihood estimate of missing data is given by

$$\hat{Y}_m = X_m\beta + \Sigma_{mo}\Sigma_{oo}^{-1}(Y_o - X_o\beta) \tag{16.2}$$

which is the kriging equation of geostatistics (see Griffith, 1993). Using the preceding notation, Haining, Griffith and Bennett show that for an autoregressive model

specification, (16.2) becomes:

$$\hat{Y}_m = X_m\beta - V_{mm}^{-1}V_{mo}(Y_o - X_o\beta)$$

which reduces to the following equation for the conditional autoregression (CAR) model specification based upon a binary geographic connectivity matrix, C, and spatial autocorrelation parameter ρ:

$$\hat{Y}_m = X_m\beta + \rho(I - \rho C_{mm})^{-1}C_{mo}(Y_o - X_o\beta).$$

In other words, the spatial EM and geostatistical kriging solutions are identical, a finding that is consistent with results reported in the preceding section.

Returning to the form of (16.1), and considering the simultaneous autoregressive (SAR) model based upon the row-standardized version of matrix C, namely matrix W,

$$\begin{pmatrix}\mathbf{Y}_o\\ \mathbf{0}_m\end{pmatrix} = \rho W\begin{pmatrix}\mathbf{Y}_o\\ \mathbf{0}_m\end{pmatrix} + (I - \rho W)X\beta + \sum_{m=1}^{M} y_m(-I_m + \rho {W_{om}}^*) + \epsilon,$$

where again missing values in the vector Y are replaced by 0_s, I_m is the indicator variable vector for missing value m that contains $n - 1$ 0_s and a single 1 in each of its m columns, ${W_{om}}^*$ is the column of the geographic weights matrix W associated with the mth missing value, and M is the number of missing values.

16.3 Eigenvector- and G_i-Based EM Solutions for Georeferenced Data

Getis (Getis and Griffith, 2002) notes that the geostatistical range furnishes one way to determine the distance parameter of the G_i statistic. Because missing data complicate a situation, this seems to be a preferred way to determine this distance parameter. Once d has been identified with a semivariogram plot and model, the following quantity needs to be computed for each location on a map:

$$\frac{G_i}{E(G_i)} = \frac{\sum_{j=1}^{n} c_{ij}(d)y_j}{\sum_{i=1}^{n} y_i - y_i} \Bigg/ \frac{n-1}{\sum_{j=1}^{n} c_{ij}(d)}, \tag{16.3}$$

where the operator E denotes the expected value, and c_{ij} is the entry in cell (i, j) of matrix C, with c_{ij} = 1 if the distance between locations i and j is less than or equal to d. This quantity becomes the spatial covariate for imputation purposes.

Similar to the way (16.2) was formulated on the basis of (16.1), the following bivariate regression equation can be formulated based upon (16.3), using a single

missing value in this instance for illustrative purposes:

$$\begin{pmatrix} Y_{o|o} \\ Y_{o|m} \\ 0 \end{pmatrix} = \alpha\mathbf{1} + \beta \begin{pmatrix} \left\langle y_i \Big/ \dfrac{\sum_{j=1}^{n} c_{ij}(d) y_j}{\sum_{j=1}^{n} y_i - y_i + y_m} \right\rangle_{o|o} \\ \left\langle y_i \Big/ \dfrac{\sum_{j=1}^{n} c_{ij}(d) y_j + y_m}{\sum_{j=1}^{n} y_i - y_i + y_m} \right\rangle_{o|m} \\ y_m \Big/ \dfrac{\sum_{j=1}^{n} c_{mj}(d) y_j}{\sum_{j=1}^{n} y_i} \end{pmatrix} - y_m I_m + \epsilon, \quad (16.4)$$

where $o|o$ and $o|m$ respectively denote observed values given observed values and observed values given both observed and missing values, and $\left\langle \quad \right\rangle$ denotes a subvector. Estimation of (16.4) requires the use of nonlinear least squares, because the missing value y_m appears in both the numerator and the denominator of fractions constructed according to (16.3).

In contrast, the eigenvector-based spatial filter EM solution may be written as

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \alpha\mathbf{1} + X\beta_x - \sum_{m=1}^{M} y_m I_m + \sum_{k=1}^{K} E_k \beta_{E_k} + \epsilon, \quad (16.5)$$

where β_x is the vector of regression coefficients for the set of X attribute variable covariates, K eigenvectors, denoted by E_k, are selected from the candidate set extracted from matrix

$$(I - 11^T/n)C(I - 11^T/n)$$

an expression that appears in the numerator of the Moran Coefficient, and β_{E_k} is the regression coefficient for the kth selected eigenvector.

One advantage of (16.5) is that it can be extended to the generalized linear model. For example, a binomial regression takes on the form

$$\begin{pmatrix} \left\langle LN\left(\frac{p}{1-p}\right) \right\rangle_o \\ \mathbf{0}_m \end{pmatrix} = \alpha\mathbf{1} + X\beta_x - \sum_{m=1}^{M} y_m I_m + \sum_{k=1}^{K} E_k \beta_{E_k}, \quad (16.6)$$

where LN denotes natural logarithm, and the vector 0_m is obtained by substituting p = 1/2 into the expression $LN\left(\frac{p}{1-p}\right)$. If p is based upon a total, then for missing values its unknown numerator initially would be set to 50% of this total.

16.4 Empirical Examples

Consider the two empirical examples of coal ash data presented in Cressie (1993) and vandalized turnip field plot yields presented in Hand et al. (1994). Distances for the G_i statistic, obtained by estimating spherical semivariogram models[1] with

[1] Coal ash: $C_0 = 0.31$, $C_1 = 1.14$, RESS = 41.1%. Turnips yield: $C_0 = 7.54$, $C_1 = 23.35$, RESS = 33.6%.

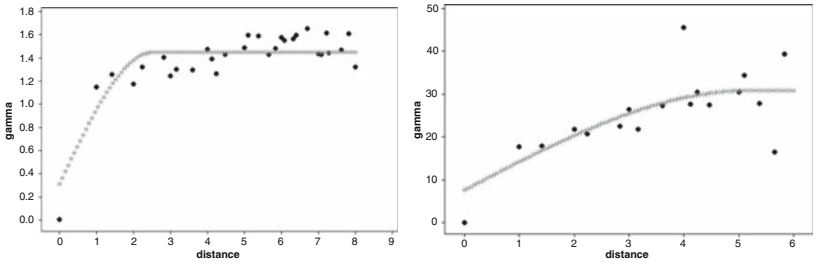


Fig. 16.2 Spherical model plots (denoted by *asterisks*) superimposed on experimental semivariograms (denoted by *solid circles*). (a) *left:* coal ash data ($n = 208$). (b) *right:* vandalized turnip field plot residuals ($n = 33$)

Table 16.2 Selected model-based imputations for two selected empirical data sets containing missing values: Pennsylvania coal ash, and vandalized turnip field plots

Estimator	Coal ash missing %	Vandalized turnip filed plot yields		
		Field(6,5)	Field(5,6)	Field(6,6)
Conventional EM algorithm	9.78	28.9	18.8	27.8
Cressie	10.27			
Spherical	10.62			
Gaussian	10.18			
Exponential	10.12			
SAR	10.17	29.99	17.66	28.26
Eigenvector-based spatialfilter	10.71	24.31	13.62	23.93
G_i-based spatial filter: constant d	9.59	0	0	0
G_i-based spatial filter: variable d_i	10.62	28.9	26.8	38.8

these data (see Fig. 16.2), respectively are 2.53 and 5.22 lattice units (these data are distributed over regular square lattice grids based upon interval scale geocoding).

Missing value estimation results are reported in Table 16.2. Of note is that the constant d values obtained as semivariogram spherical model ranges do not perform well. Rather, estimating a d_i for each location tends to render more meaningful estimates. These variable d_i values are obtained by increasing d for location i until its relative G_i statistic (i.e., the ratio of the expected and observed values for location i) begins to decrease. Both the eigenfunction- and G_i-based spatial filter missing value estimates for the coal ash data are greater than most of the semivariogram and the SAR estimates. In contrast, the eigenfunction-based spatial filter estimates for the missing turnip yields appear too small, whereas the G_i-based spatial filter estimates for two locations appear reasonable, while the third estimate appears too large.

16.5 Conclusions and Implications

A comparison of the two spatial filter missing value estimation techniques is summarized in Table 16.3. A serious drawback of the G_i-based procedure is the possible need to additionally approximate missing data location d_i values; this extra

Table 16.3 Comparisons of the two spatial filter estimators

	G_i-based	Eigenvector-based
Conceptual basis	Kriging	EM algorithm
Required covariates	Approximation of d_m plus computed $G_i s$	Selected eigenvectors
Estimation technique	Nonlinear LS (NLS)	OLS or GLM
Value restrictions	Positive numbers with a natural origin	Real numbers
Spatial context	Inter-pointdistances	Tessellation topology or inter-point distances

calculation may not be avoidable by resorting to semivariogram model range estimation. A second serious drawback is that only positive numbers with a natural origin can be used to compute G_i statistics.

In conclusion:

1. The Getis spatial filter approach enables missing georeferenced values to be imputed.
2. The imputation procedure may well require an additional estimation to be made – namely, the distance threshold for each missing value.
3. Resulting imputations appear to be reasonably consistent with those obtained with other procedures, including Griffith's eigenvector-based one.
4. Findings reported here supplement those reported in Getis and Griffith (2002).

Overall, Bivand's (2002) contention that G_i-based spatial filtering can be used for imputation purposes appears to be correct, although in need of more extensive study; his contention that eigenfunction-based spatial filtering cannot be used to calculate imputations is incorrect.

Chapter 17
Spatial Patterns of Fertility in Rural Egypt

John R. Weeks

Abstract The Getis–Ord G_i^* statistic and the Getis spatial filtering method are shown in this paper to be very useful geospatial tools for uncovering the spatial patterns of human reproduction in a rural governorate in Egypt that had been assumed by many to be a spatially homogeneous area. We apply the G_i^* statistic to dasymetrically mapped data from the 1976, 1986 and 1996 censuses of Egypt to show that there were very distinct spatial patterns in fertility over time in this predominantly rural region of the Nile Delta. The spatial filtering technique allows us to conclude as well that the spatial component became more important over time as a predictor of fertility levels. Improvements in education represent a key feature of the changing rural social environment driving these spatial changes in fertility. There is evidence as well that increases in contraceptive utilization contributed to this change, but we are unable to evaluate its spatial component. Nonetheless, the research illustrates and illuminates the underlying conceptual framework that demographic behavior is a joint function of who people are and where they are.

17.1 Background

Demographic literature is rich in studies that compare rural with urban areas, and in the former women invariably have more children on average than do women in the latter. It is nearly an iron law. One of the problems with this type of comparison, however, is that even if rural places have higher fertility than urban places within the same country, rural fertility may be higher in some countries than in others. We may well find that rural fertility in a richer country is lower than urban fertility in a less-rich country. This points to what might be thought of as the cultural and geographic relativity of the urban–rural dichotomy and, at the same time, the underlying social nature of human reproduction. A second problem is that the dichotomy ignores any variability that might occur within either the rural or the urban areas,

J. R. Weeks
Department of Geography, San Diego State University, San Diego, CA, USA
e-mail: john.weeks@sdsu.edu

L. Anselin and S.J. Rey (eds.), *Perspectives on Spatial Data Analysis*,
Advances in Spatial Science, DOI 10.1007/978-3-642-01976-0_17,

assuming instead that fertility is uniformly higher in rural than in urban areas. Weeks and his associates have shown that this assumption may be very wrong both in rural areas (Weeks et al., 2000) and in urban areas (Weeks et al., 2004). In Greater Cairo in 1986, for example, the average neighborhood-level total fertility rate (TFR) calculated by indirect methods described below was 3.1 children per woman, whereas in the rural governorate of Menoufia, just to the north of Cairo, the average level per village was 5.8. That clear distinction hides considerable overlap, however. The lowest TFR in Menoufia was 3.2 and only 25% of Cairo neighborhoods had a level that was lower than that. The highest TFR in Cairo was 7.7 and only 1% of Menoufia's villages had a rate higher than that.

Very little attention has been given in the literature to the social causes and consequences of fertility levels at these local levels. There is a vast literature on fertility differentials, to be sure, but attention is paid largely to characteristics of individuals without regard to where they live. Entwisle, Casterline and Sayed (Entwisle et al., 1989) have demonstrated that village contexts can be important influences on contraceptive behavior (and thus on fertility) in rural Egypt, but their analysis was limited spatially to a distinction between villages in upper and lower Egypt. There is also a nearly universal finding that fertility differs by social class (defined by income differences, occupational prestige, educational attainment, and often incorporating some element of race/ethnicity). And, since there is a tendency for there to be a geographic sorting process by social class, the local spatial dimension of fertility is implicitly incorporated into that model. Yet, that is almost never measured by demographers, and little attention has otherwise been given to the demographic and social variability in fertility across space. That is to say, little attention is paid to the ecology of fertility, even among social ecologists. Rather, the emphasis is on examining fertility levels at the individual level, using data from surveys that by and large do not permit more than a very generalized spatial analysis. These studies of necessity focus attention on national comparisons or on regional differences within a country.

From a purely geographic perspective, one could argue that this is simply a scale issue. Variability may exist at any spatial scale and the only issue is whether we have the tools to measure it. But from a broader social science perspective, the scale issue matters because it fits into the human ecological approach that suggests that the behavior of people is influenced by their personal characteristics (who they are) and also by their locational characteristics (where they are). This is the underlying premise of spatial demography (also known as geodemographics), as discussed by Weeks (2004a). If we are to understand the patterns and changes in human fertility, we must take into account all aspects of the social world in which people live. Characteristics such as education, occupation, income, ethnicity and religion all play a role in shaping behavior, but the likelihood that a given person will be at one end or another of the continuum on each variable may well depend upon where they live. This is not to be interpreted as geographic determinism, but rather as an acknowledgement that we are social creatures who are influenced by those around us. If we live in an area where, for whatever reason, education is not valued, then the probability that we will value education is commensurately low and our life will probably

turn out very differently than if we live in an area where education is highly valued and sought after. This is the essence of spatial autocorrelation as it applies to human society.

The model that guides our research thus incorporates the assumptions that (a) the social environment influences the social and human capital variables that more directly influence the demand for children; (b) the reproductive behavior of some people within a village may influence the behavior of others, even net of the human capital opportunities that objectively exist in the village; and (c) these influences operate on reproductive levels through the mechanisms of the proximate determinants of fertility, such as age at marriage and the use of contraceptives within marriage, to determine fertility at the local level; but (d) changes in reproductive behavior at the local level may be influenced by changes in, and reciprocally influence changes in, other neighboring regions, resulting in spatial patterns of fertility transition; the consequences of which (e) ultimately determine the overall societal level of reproduction, thus creating the wider phenomenon of a fertility transition.

Our goal in this research is to use this conceptual framework to build upon earlier work that examined the spatial component of fertility in the rural governorate of Menoufia, Egypt. That study (Weeks et al. 2000) examined data from the 1976 and 1986 censuses of Egypt, but was published prior to the release of the 1996 census data. That study concluded by making prognostications for the 1986–1996 period with respect to fertility levels, as follows:

> The period from 1976 to 1986 was a period of overall relative stability in fertility levels in rural Egypt and not until the 1996 census data become available at the village level will we be in a position to track significant changes in fertility over time. However, it is clear that at least by 1976 there were clear spatial patterns to fertility in Menoufia and our analysis suggests that these spatial patterns were even more definitive in 1986 than they had been in 1976. This seems to suggest the existence of some momentum for change, which we hypothesize will be observable when the 1996 data become available. The southern portion of the governorate was more obviously the location of higher than average fertility in 1986 than had been true in 1976 and we would predict that the clustering of lower fertility will have exhibited a southward drift or diffusion by 1996. The results from our spatial filtering procedure suggest that some of this effect will be due solely to where villages are located, regardless of any changes in female education. The analysis also suggests that improving levels of female education will have been the most important human capital influence on fertility between 1986 and 1996 (Weeks et al., 2000, p. 712).

In order to evaluate the correctness of those forecasts, we first revisit the data sets in order to improve the dasymetric mapping so that the point-pattern spatial analysis, which is based on distances between villages, is optimized. In that process we are able to harmonize the administrative boundary changes that took place between censuses. Furthermore, we have now been able to use the Egypt Demographic and Health Survey data to derive usable estimates of underenumeration in the census, from which we can calculate improved fertility estimates for each village in Menoufia. With these methodological refinements in hand, our research questions become (1) Did fertility decline more between 1986 and 1996 than it had between 1976 and 1986? (2) Did fertility decline more rapidly in the southern part of the governorate than in the northern part? And (3) was the decline due both

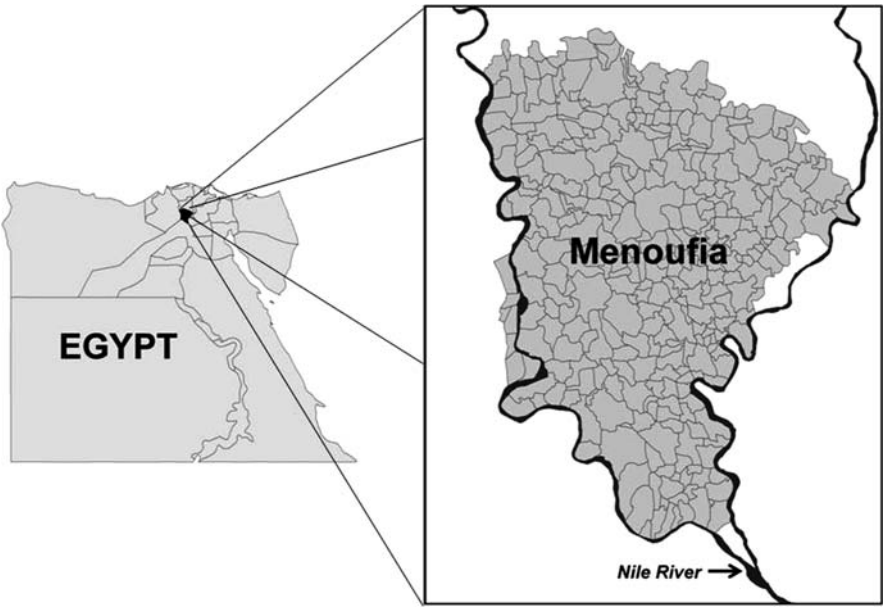


Fig. 17.1 The study site of Menoufia governorate, Egypt

toimprovements in female literacy as well as being conditioned by the geographic location of a village within the governorate?

17.2 The Study Site

The study site is the governorate of Menoufia, in the Nile delta region of Egypt, just to the northwest of Cairo (see Fig. 17.1). Menoufia is one of the 26 governorates that form the administrative regions of Egypt, roughly equivalent to states in the United States or to counties in the United Kingdom. The 1996 Census of Egypt enumerated 2.8 million people in Menoufia, representing about 4% of the total population of Egypt. The officially-defined rural population accounted for 80% of the governorate's population in 1996. The southern region of Menoufia is situated just below (to the north of) the Barrage that controls the flow of water from the Nile as it enters the Delta, which has permitted perennial irrigation in the region since the early nineteenth century. But even for millennia prior to that this was a rich agricultural area with traditions that almost certainly contribute to the maintenance of low levels of education and higher than average levels of fertility.

We chose Menoufia as a study site less for its specific demographic characteristics than for the fact that it has been relatively well studied in nonspatial analyses and thus there are comparative studies by which to judge the spatial analysis produced

by our own work (see especially Gadalla, 1978; Weeks et al., 2000). Menoufia does have some advantages for spatial analysis including its essentially flat landscape in the Delta region of the Nile, which means that elevation is not an issue that needs to be dealt with. Partly for this reason, there are more than 300 villages, with an average population per village in 1976 of 6,036, increasing to 7,840 in 1986, and up to 9,349 by 1996. Thus, in 1996 the average village had more than half again as many people living in it as in 1976, a situation that almost certainly was going to induce some changes into village life.

17.3 Data and Methods

17.3.1 The Variables in the Statistical Models

We use census data aggregated at the qurah level. In its most literal translation from Arabic, a "qurah" is a city, but in the Egyptian census definitions it refers to the administrative boundaries of a village, as illustrated in Fig. 17.1. The dependent variable in our analysis is an estimate of the fertility rate derived from age structure data in the census. The independent variables include those that are measured comparably over the three censuses that we are analyzing: (1) marital status as a proxy for the average age at marriage; (2) female education (illiteracy); and (3) two measures of the local neighborhood context, including the sex ratio as a proxy for the impact of the outmigration of males to Cairo or, more likely, to Gulf States for temporary employment, and the size of the village as a proxy for its level of urbanness.

As is true in many less-rich countries, the amount of detailed information collected in the census is limited, and so we employ an indirect measure of the total fertility rate (TFR – the number of children a woman would have in her lifetime if reproduction remained at the current level) as our dependent variable. We derived the TFR from the age structure data in the census, applying the CBR-TFR population analysis spreadsheet developed by the International Programs Center of the US Census Bureau (Arriaga, 1994). This spreadsheet estimates the crude birth rate and total fertility rate using the total population, the female population in childbearing ages by 5-year age groups, general fertility rate, and empirical patterns of age-specific fertility rates included in the program.

Before making these calculations, we dealt with the issue of the accuracy of the age and sex structure as enumerated in the census, since if we are to estimate fertility indirectly from the age structure, we need to be confident in that source of data. We have employed information from Demographic and Health Surveys (DHS) in Egypt as a source of comparative data. The age groups that are of importance for indirectly estimating fertility are the children aged 0–4, which are notoriously the least well enumerated, and women of reproductive age (15–49). In general we assume that the interviewers of the DHS were likely to obtain more accurate information than would have been obtained by enumerators in the census. The sampling error in the DHS

is sufficiently low so that any large disparities between the DHS and the census for both girls and boys are likely to be statistically significant beyond the 0.01 level. The 1995 DHS included 548 households from 17 different villages in Menoufia governorate, and from the household listing in the DHS we reconstructed the age and sex structure as reported for each household, representing a sample household population in Menoufia in the 1995 DHS of 3,196 persons. These data were then compared with the 1996 census age distribution for Menoufia to assess potential errors in the census age distribution. We also made the assumption that the 1-year difference between the census and the DHS would not affect our comparison in any meaningful way.

From this comparison we concluded that boys aged 0–4 were underenumerated by 18% in the census, and girls that age were underenumerated by 7%. We thus made an across-the-board adjustment in each village to increase the number of children aged 0–4 by those amounts. We then rejuvenated the population of girls and boys aged 0–4 from the census by dividing by the respective sex-specific survivorship rates. Survivorship rates are calculated from life tables derived from ${}_nM_x$ data compiled by the Cairo Demographic Center (2001), adjusted for likely underregistration of deaths. From these estimates we calculated the number of births over the prior five years, and dividing that by 5 and then dividing by the rejuvenated average number of women of reproductive age produced an estimate of the average single year general fertility rate. This value was combined with data on the female population by 5-year age group, and the total population in each village to estimate the total fertility rate based on empirically derived relationships between the GFR, the female population, the total population, and age-specific fertility rates in the population analysis spreadsheets. Our calculations produced a TFR for Menoufia in 1996 of 3.6 children per woman.

We repeated the procedure for 1986, comparing data from the 1988 DHS with the 1986 census data for Menoufia. The DHS in 1988 included a sample of 379 households in 10 villages, with a total household population of 2,449. We did not find that the differences were statistically significant, so no age adjustments were undertaken. We then constructed a life table for Egypt for 1986, building on the 1996 life table, but incorporating higher death rates, especially at the younger ages, as estimated from DHS data, in order to rejuvenate the population, as described above for the 1996 calculations, in order to complete our indirect estimation procedure. We did not have a comparable fertility survey for the period near the 1976 census, and so we assumed that no adjustment was necessary, since that had been true in 1986. We then rejuvenated the 1976 census data based on the life table for Egypt available from the International Programs Center at the US Census Bureau (http://www.census.gov/ipc/www/idbacc.html).

The predictor variables are limited in number, especially given the need to have comparability across all three census dates. We measure the human capital variable in the village in terms of female education. The data are available only for all women aged 15 years and older, regardless of specific age or other characteristic. In addition, because of the limited educational attainment of women in Menoufia, the educational variable was measured as the percentage of women aged 15 years

and older who are illiterate. Our expectation is that lower levels of illiteracy will be associated with lower levels of fertility, and that declines in illiteracy over time will be associated with declining fertility. We also anticipate some interaction between education and marital status. The literature suggests that the early impetus for fertility decline in Arab countries has come from a delay in marriage (occurring in the general absence of any offsetting rise in out-of-wedlock births) (Rashad, 2000). In a society where virtually all women eventually marry, the proportion of women who are currently married should be an index of the relative age at marriage from one place to another. And, since out-of-wedlock births are relatively rare in Muslim countries such as Egypt, we would expect that fertility will be lower where the proportion of married women is lower.

We also take into account that in rural areas there may be migration out of the village and it will disproportionately affect males. As a result, the relative absence of men could have a dampening influence on fertility. We control for this effect by calculating the sex ratio of males aged 25–44 years to females aged 20–39 years as a covariate in the analysis. The final covariate introduced into the model is the total population size of the village, serving as a proxy for the relative degree of urbanness of the place.

17.4 Dasymetric Mapping of the Villages

The use of point pattern spatial analysis with data that are aggregated at an administrative level such as the qurah in Menoufia requires that an assumption be made about the point that will best represent the area for which the data are aggregated. The easiest and most common solution is to assume that data are uniformly distributed within the area and that the geographic center (centroid) of that polygon adequately describes the average location of people to whom the data refer. This approach may, however, compromise the accuracy of the spatial analysis. For decades, if not centuries, Menoufia has been one of the most rural and most densely populated rural areas of Egypt (Gadalla, 1978). It has been, and remains, predominantly agricultural, with the population congregated into rural villages from which people go out each day to work in the fields. Because most of the area in each qurah is devoted to agriculture, the assumption of a uniform distribution of the population within each qurah is certainly not accurate, and there is no reason to believe that the geographic center of the area defined as the village is a good representation of where the population actually resides. This is one component of the well-known modifiable areal unit problem (MAUP) Openshaw (1984). “The MAUP concerns the fact that varying the scale of data aggregation, and/or aggregating data using different aggregation boundaries at a single scale, may affect the results of spatial statistical analysis” Mennis (2002).

A dasymetric approach to the data helped to deal with two additional problems that confronted us with the Menoufia data: (1) the administrative boundaries of several qurah actually cut right through the middle of built areas and so we had data for

what seemed to be two or more places when in fact the data really only referred to a single village; and (2) the administrative boundaries had changed slightly between 1976 and 1996, leading to the need to harmonize the data over time. The dasymetric approach attempts to improve on the default method (the geographic center of the entire administrative area) by more accurately locating the point in each polygon based on ancillary information about where the people are located. Our ancillary data are based on the classification of satellite imagery into those areas that represent a built environment (the villages) or not (the agricultural fields and other non-village areas). For this purpose we used an Indian Remote Sensing IRS-1C LISS-III 24-m resolution multispectral image covering bands 2, 3, and 4 (green, red, and near infra-red) satellite image acquired in 1996. Although the village boundaries may have enlarged somewhat between 1976 and 1996, especially as a result of the significant population growth discussed above, we assume that the geographic center of the built area is the same for all three census dates. The classification methods used with the imagery are discussed elsewhere (Weeks et al., 2004, 2005; Weeks, 2004b; Rashed et al., 2001, 2003, 2005). Once the imagery was processed and the built area identified, we calculated the centroid of the built area and used it to represent the data for the village, rather than the geographic center of the entire administratively defined area. If more than one built area existed within the village administrative boundaries, the weighted mean center of all built areas was found, using the areal extent of each built area as its weight.

Especially important in this process was the identification of single villages that had been split into multiple administrative units. Figure 17.2 illustrates how the default placement of points could artifactually create spatial autocorrelation in the data because the hypothetical village shown is administratively divided into four parts, for each of which a separate set of census tables will have been created. Applying the geographic center to those data would then produce data allocated to four different points. Since the demographic characteristics are likely to be similar for all four parts of the village, this situation would appear at first glance to refer to four similar villages next to one another – a classic case of spatial clustering. In reality, the data are all associated with different segments of the same built area, a fact that we discover only with the use of ancillary data, in this case the satellite imagery, and which is corrected for through the dasymetric approach. This process reduced the number of points associated with villages from 314 to 286.

17.5 Statistical Analysis

Our approach to answering the research questions posed in this paper is to employ multiple regression techniques, taking into account any observed spatial patterns. Assuming that a spatial pattern exists in the data, we next want to know exactly where the clustering occurs. Where are the "hot spots" in which high levels of fertility are clustered and where are the "cold spots" in which lower levels of fertility are clustered? The local spatial statistic we utilize is the $G_i^*(d)$ statistic (Getis,

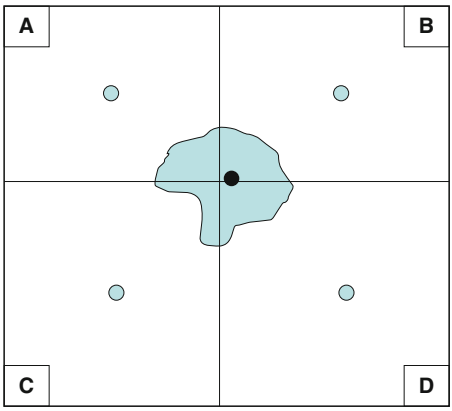


Fig. 17.2 Situations improved by dasymetric mapping

1995b; Ord and Getis, 1995). We then use the results from the clustering statistics to conduct a spatially filtered regression analysis. This is a method that allows us to quantify the role that spatial autocorrelation is playing in the observed variability of the dependent variable (Getis, 1995b; Getis and Griffith, 2002; Weeks et al., 2004). Anselin and Rey (1991) have differentiated between two forms of spatial dependence; that which is a nuisance, and that which represents a substantive spatial process. As a nuisance, it can be controlled with a properly designed weights matrix within a spatially autoregressive model. However, when the spatial dependence is a subject of inquiry, as it is in this research, it is useful to be able to quantify the role that it plays within each of the predictor variables. Two such filtering approaches are currently available – the Getis filtering method (see Getis, 1995b) and the Griffith eigenfunction decomposition method (see Griffith, 2000). Both of these methods are capable of identifying the spatial effects within a regression framework (Getis and Griffith, 2002), but in this research, we employ the Getis method. We thus use the $G_i^*(d)$ statistic as a spatial filter to extract the spatially autocorrelated portion of each of the variables in a regression model. The difference between the original variable x_i and the filtered variable x_i^f becomes a new variable x_i^{sp}, that represents the spatial effects embedded in x_i (Getis, 1995b). These two variables, x_i^f and x_i^{sp} replace the original variable x_i in the regression equation to produce a spatially filtered regression model in which the contribution of the spatial and filtered (non-spatial) components of each factor can be determined by the beta coefficients in the resulting model. These techniques of spatial filtering were developed originally by Getis (1995b) and have been modified into a Fortran program by Scott (1999) in the format that will be used in this project. In this format, the regression model to be tested is as follows:

$$\begin{aligned}\text{TFR} =& \{\text{illiteracy filtered}\} + \{\text{illiteracy spatial}\} \\ &+ \{\text{marital status filtered}\} + \{\text{marital status spatial}\} \\ &+ \{\text{sex ratio of adults filtered}\} + \{\text{sex ratio of adults spatial}\} \\ &+ \{\text{village population filtered}\} + \{\text{village population spatial}\} + \text{error}\end{aligned}$$

By solving the equation with the filtered and spatial components separated, the spatial autocorrelation is removed from the residuals and incorporated into the model as a component helping to predict variation in the dependent variable.

The final model fit has been shown to be comparable whether using spatial filtering or autoregressive models (Getis and Griffith, 2002), but the spatial filtering technique has the advantage of giving us intermediate information about the effect of spatial dependence on the dependent variable that is not available within an autoregressive framework.

17.6 Results

We organize our results around the three research questions posed above, which require that we look at the spatial pattern and regression results for 1976, then 1986, and 1996, and then examine the changes over time.

17.6.1 Did Fertility Decline More Between 1986 and 1996 than Between 1976 and 1986?

Fertility did indeed decline more in Menoufia between the 1986 and 1996 censuses than it had in the previous decade. This pattern was at odds with the country as a whole, which experienced a more rapid decline between 1976 and 1986 than between 1986 and 1996, as can be seen in Table 17.1. In 1976, the TFR for Menoufia was slightly less than for the country as a whole, but Menoufia experienced only a slight decline between then and 1986, as Weeks and associates (2000) have already shown. However, the country was experiencing a rather rapid decline during that time and by 1986 the TFR in Menoufia was nearly one child higher than for the country as a whole. The decline in Egypt was led by the urban areas, which had about a 10-year head start on rural areas in the fertility transition (Weeks et al., 2004). Between 1986 and 1996 Menoufia experienced a very rapid drop in fertility and thus in 1996 it was once again at parity with the nation.

The data in Table17.1 lead us clearly to expect that changes in the educational level of women were playing a role in Menoufias fertility decline. Between 1976 and 1996 there was little overall change in the proportion of women who were married, suggesting that there were few observable shifts in the pattern of marriage. However, female illiteracy dropped substantially, from 77% down to 52%. At the same time,

Table 17.1 Fertility decline in Egypt and Menoufia, 1976–1996

	Egypt		Menoufia					
	TFR	Change in TFR	TFR	Change in TFR	Proportion married	Female illiteracy	Sex ratio at reproductive ages	Village population size
1976	6.05		5.84		0.61	0.77	0.86	6,036
1986	4.51	1.54	5.40	0.44	0.63	0.67	0.88	7,840
1996	3.57	0.94	3.56	1.84	0.63	0.52	0.92	9,349

Sources: Data for Egypt are from the International Database of the International Programs Center of the US Census Bureau; Menoufia data were calculated by the authors from Egyptian census data

the sex ratio at the reproductive ages was increasing, probably due especially in the 1990s to the Gulf War, which forced many Egyptian men back to their villages from the oil fields in Kuwait and Iraq. All other things equal, we would expect this to have created an upward pressure on fertility as men returned to more frequent intercourse with their wives. If age at marriage was not rising, and women were more likely to have their husbands around, the likely explanation for the decline would almost have to be an increase in contraceptive utilization among women, and we will look for that evidence later using data from the Egypt Demographic and Health Survey.

Having now laid out the case for the overall pattern of change in fertility in Menoufia, the remainder of this analysis is devoted to examining at a finer spatial scale what was going on Menoufia to create this fertility transition that was timed differently from the country as a whole, so that we can improve our understanding of the demographic changes in Egypt.

17.6.2 Spatial Patterns of Fertility in Menoufia in 1976

In 1976 the average woman in Menoufia was having children at a rate that would produce nearly six children over the course of her lifetime. Although fertility was quite high in most places throughout the governorate, the distribution of fertility by village was negatively skewed, indicating that there were several places with significantly below average fertility levels. The left panel of Fig. 17.3 shows the spatial pattern of total fertility rates by village. The substantial level of spatial autocorrelation is evidenced by the global Moran's I statistic of 0.35, with $z(I)$ being equal to 9.39.

The right panel of Fig. 17.3 shows the statistically significant clusters with respect to fertility, based on the Getis–Ord G_i^* statistic, as discussed above. There are scattered clusters of low fertility, especially in the center of the governorate around the city of Shbin El Kom, the capital and largest of the governorate's handful of urban areas and home to Menoufia University which was, in fact, founded in 1976. In particular, low fertility is found in the northern area formed by the triangle

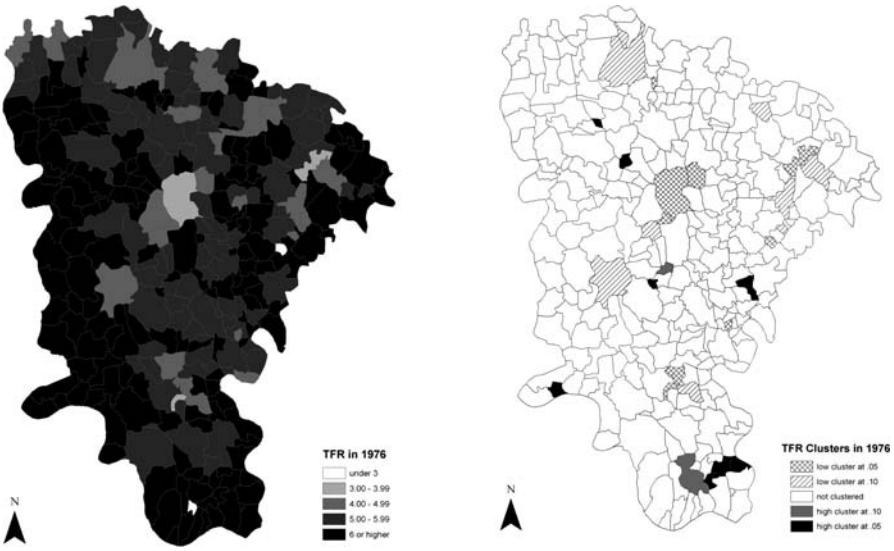


Fig. 17.3 Fertility levels in Menoufia in 1976

between Shibn El Kom, a larger urban center, Tanta, which is just across the administrative boundary in the adjacent northern governorate of Gharbia (not shown in Fig. 17.3), and another urban center, Banha, which is just to the east in the adjacent governorate of Qalyubbia (also not shown). Highways and railroad lines link these three places and form, in essence, the more urban or cosmopolitan corridor in Menoufia. The highest fertility is in the southern part of the governorate nearest to the Nile Delta Barrage. As the crow flies, this area is geographically closer to Cairo, but it is less connected to Cairo by way of transportation networks than is the northern part of the governorate.

Although we have only a limited number of variables available to us, the ordinary least-squares model shown in Table 17.2 reveals that the four predictor variables account for 47% of the village to village variability in fertility levels. Of these, the proportion of women who are married is clearly the most important, and as that proportion goes up, so does the fertility rate. But the other three variables are also statistically significant predictors, with higher illiteracy being associated with higher fertility, a higher sex ratio equating to higher fertility and a larger population size in the village correlating with lower fertility. However, the model has a high level of spatial autocorrelation in the residuals, suggesting that the model needs better specification to account for the spatial component.

The lower panel of Table 17.2 shows the results of the spatially filtered regression, undertaken as described above in the methods section. The spatial and non-spatial components of the proportion married are nearly equally important predictors of fertility, suggesting that both the level itself and being in the neighborhood of similarly situated villages affects the level of the TFR. However, only the non-spatial

Table 17.2 Regression models for fertility in Menoufia, 1976

Predictor variables	Initial OLS model			
	Standardized beta coefficient	t-score	p-value	Moran's $z(I)$
(Constant)		−1.463	0.144	
Proportion married	0.502	8.555	0	3.891
Female adult illiteracy	0.139	2.382	0.018	5.39
Sex ratio at reproductive ages	0.15	3.239	0.001	2.276
Village population	−0.159	−3.412	0.001	1.091
Adjusted R^2 = 0.47				
$Z(I)$ for residuals	2.997			
	Spatially filtered model			
(Constant)		−0.245	0.807	
Female illiteracy non-spatial	0.135	2.423	0.016	
Female illiteracy spatial	0.037	0.6	0.549	
Proportion married non-spatial	0.357	6.594	0	
Proportion married spatial	0.412	5.94	0	
Sex ratio non-spatial	0.149	3.193	0.002	
Sex ratio spatial	0.113	2.121	0.035	
Village population	−0.174	−3.706	0	
R^2 = 0.48				

Dependent variable is village TFR

component of female illiteracy is statistically significant. We posit that the spatial component is not significant because the high level of illiteracy in 1976 meant that nearly every village was likely to be in the midst of other villages with generally high levels of female illiteracy. The non-spatial component of the sex ratio was somewhat more important than the spatial component, although the latter was statistically significant. The village population size had not exhibited a spatial pattern and so it was not spatially filtered. It remains a statistically significant predictor of fertility. The R^2 is essentially the same for both the initial and the filtered models, but the filtered model has provided additional information about the spatial nature of the predictors of fertility in Menoufia.

17.6.3 Spatial Patterns of Fertility in Menoufia in 1986

As has been anticipated from the earlier study Weeks et al. (2000), the fertility pattern in 1986 is not dramatically different from that in 1976. The left side of Fig. 17.4 reveals a spatial pattern very similar to that in Fig. 17.3, and the Moran's I of 0.37 (with a normalized z-score of 9.85) confirms the visual impression of a non-random distribution of fertility levels around the governorate. The right side of Fig. 17.4 does show, however, that the hot spots of high fertility in the northern part of the governorate were no longer visible in 1986. The implication is that the northern villages in the governorate were the ones most involved in the relatively modest decline in fertility between 1976 and 1986.

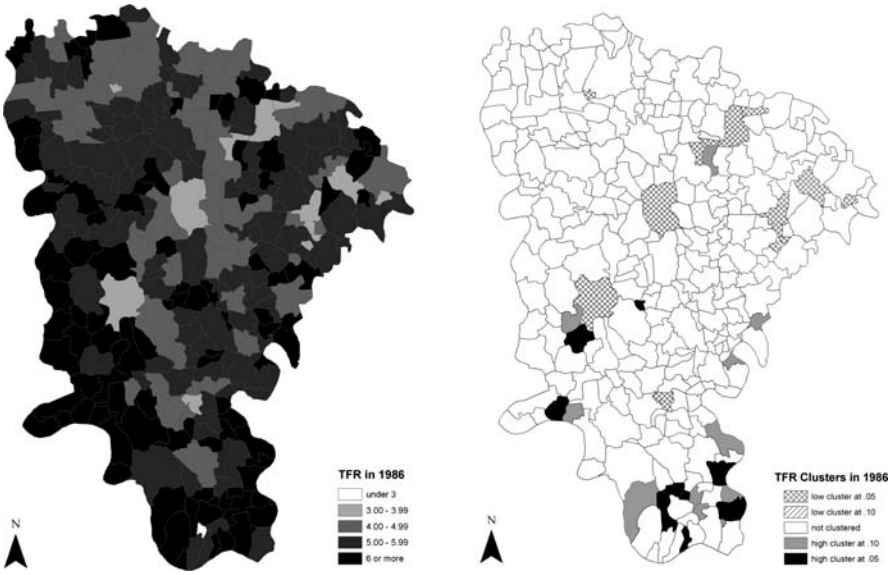


Fig. 17.4 Spatial pattern of fertility in 1986

The regression models for 1986 are shown in Table 17.3. Two things stand out in these results. The first is that female adult illiteracy emerges in 1986 as the most important predictor of the TFR in Menoufia's villages. The second is, that largely as a result of the emergence of education as a key predictor of fertility, the R^2 goes up to 0.59, which is considerably improved over the 1976 results. The sex ratio is again a predictor of fertility levels, but in 1986 the size of the village's population is no longer a factor.

The residuals showed a significant level of spatial autocorrelation, as they had in 1976, and so we applied the spatial filtering process to these data. The non-spatial component of illiteracy was a stronger predictor of fertility than was the spatial component, but both are the top two factors influencing fertility levels in 1986. The spatial component of the proportion married was slightly more important than the non-spatial, whereas the non-spatial component of the sex ratio was more important than the spatial, as had been true in 1976.

17.6.4 Spatial Patterns of Fertility in Menoufia in 1996

By 1996 the spatial pattern of fertility in Menoufia was clearly altered from previous years, as shown in Fig. 17.5. Dramatically lower fertility levels are nearly the norm throughout the governorate, so much so that there are very few clusters of low fertility. The distribution of TFR by village is now positively skewed and the more unusual villages are now those that persist in their high fertility. These places are most noticeably in the southern part of the governorate.

Table 17.3 Regression models for fertility in Menoufia, 1986

Predictor variables	Initial OLS model			
	Standardized beta coefficient	t-score	p-value	Moran's $z(I)$
(Constant)		−3.679	0	
Proportion married	0.244	5.297	0	5.294
Female adult illiteracy	0.589	12.587	0	5.223
Sex ratio at reproductive ages	0.11	2.813	0.005	4.907
Village population	0.019	0.464	0.643	1.735
Adjusted R^2 = 0.594				
$Z(I)$ for residuals	2.559			
	Spatially Filtered Model			
(Constant)		−2.168	0.031	
Female illiteracy non-spatial	0.48	10.83	0	
Female illiteracy spatial	0.347	6.361	0	
Proportion married non-spatial	0.167	4.005	0	
Proportion married spatial	0.263	4.339	0	
Sex ratio non-spatial	0.101	2.588	0.01	
Sex ratio spatial	0.042	0.907	0.365	
Village population	0.01	0.249	0.804	
$R^2 = 0.605$				

Dependent variable is village TFR

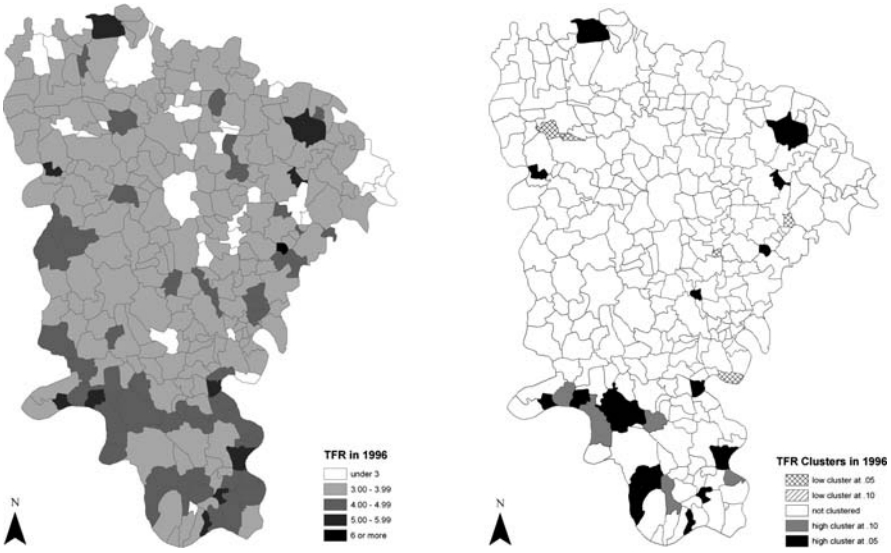


Fig. 17.5 Spatial pattern of fertility in 1996

Table 17.4 Regression models for fertility in Menoufia, 1996

Predictor variables	Initial OLS model			
	Standardized beta coefficient	t-score	p-value	Moran's $z(I)$
(Constant)		−4.509	0	
Proportion married	0.331	6.419	0	7.09
Female adult illiteracy	0.393	7.712	0	6.234
Sex ratio at reproductive ages	0.197	4.26	0	9.229
Village population	−0.014	−0.299	0.765	1.825
Adjusted $R^2 = 0.45$				
$z(I)$ for residuals	2.544			
	Spatially filtered model			
(Constant)		−3.34	0.001	
Female illiteracy non-spatial	0.296	6.246	0	
Female illiteracy spatial	0.291	5.046	0	
Proportion married non-spatial	0.248	5.241	0	
Proportion married spatial	0.214	3.605	0	
Sex ratio non-spatial	0.179	3.849	0	
Sex ratio spatial	0.118	2.435	0.016	
Village population	−0.014	−0.297	0.767	
$R^2 = 0.45$				

Dependent variable is village TFR

In 1996 female illiteracy was the most important predictor of fertility, as it had been in 1986, although the proportion married was nearly as important, as seen in Table 17.4. The sex ratio continued to be a significant factor, whereas population size was not. Again, the spatial autocorrelation in the residuals led us to engage in spatial filtering and the results suggest that the non-spatial component of female illiteracy was slightly more important than the spatial component, and this was true as well for the proportion married, and also for the sex ratio. In all cases, however, both the spatial and non-spatial components of those variables were statistically significant. Overall, the predictor variables in 1996 were able to explain 45% of the intervillage variability in the TFR in Menoufia.

Comparisons of the three different census dates suggests that over time illiteracy became an ever-more important predictor of village-level fertility rates and that the spatial component emerged out of the shadows between 1976 and 1986 to assume greater importance in understanding fertility patterns within Menoufia.

17.6.5 Did Fertility Decline More in the South than in the North?

We examined the overall change in fertility between 1976 and 1996 to see if there was a spatial pattern to the change. As can be seen in Fig. 17.6 there is a pattern to the change, and the z-normalized value of Moran's I is 5.03. But the decline in fertility

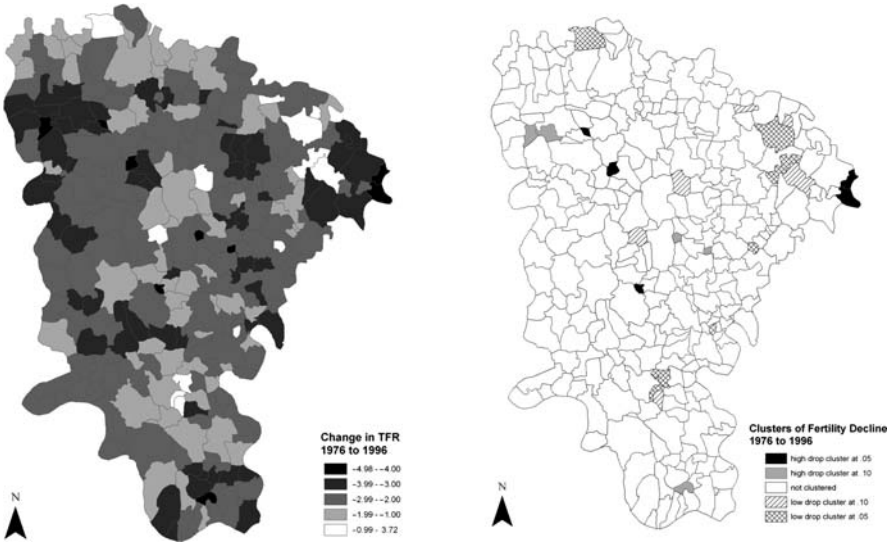


Fig. 17.6 Spatial pattern of fertility change between 1976 and 1996

was generally greater in the north than in south, contrary to what we had anticipated. The data are shown in terms of absolute decline, regardless of the starting point, so this would have privileged the villages in the south, where the starting level of fertility was highest. Yet, even given this advantage to the south, the absolute decline was higher in the north. The right-hand panel of Fig. 17.5 shows the clusters of fertility change. Although the clusters are scattered about the governorate, there are more "high decline" clusters in the north than in the south.

17.6.6 Was the Fertility Decline Due Both to Improvements in Female Literacy and to Location Within Menoufia?

The data show that fertility underwent a dramatic decline in this governorate between 1986 and 1996, a result that had been foreshadowed by the changes taking place between 1976 and 1986, especially the improving level of education among women, which was part of a nationwide program pushed by the government in Egypt (Fargues, 1997). Between 1976 and 1986 the percent of adult women who were illiterate dropped from 81% to 67%, even though there was little change in fertility. It seemed unlikely that a shift of that size in female education would not lead eventually to a decline in fertility and, indeed, that drop did occur between 1986 and 1996, as the level of education of women continued to climb, and as women began to catch up with men in terms of educational attainment. On the other hand, there was little change in the percent married, suggesting little change in the average age

at marriage, and the sex ratio in the adult ages increased, which we would posit should encourage, rather than discourage fertility.

The regression models of fertility change between 1976 and 1996 quantify these changes, as shown in Table 17.5. We included the fertility level in 1976 as an endogenous variable in the model, because we wanted to see how the other predictors behaved after controlling for the fact that the 1976 fertility level might influence the subsequent fertility decline. As we expected, the change in the proportion married did not predict a change in fertility, largely because there was not much change in the proportion married. The largest standardized beta coefficient, other than the initial fertility level, was the change in levels of illiteracy. The unstandardized coefficients (not shown in the table) suggest that a 10 percentage point drop in illiteracy is associated with a decline of one child in the total fertility rate. The sex ratio was also significantly related to a drop in fertility, with a decline in the sex ratio (indicating fewer men per woman) being associated with a decline in fertility. Change in population size of villages was not associated with a change in fertility.

It was noted above that fertility dropped more in the north than in the south, and that is also where educational levels were changing most rapidly for women. Figure 17.7 shows that in 1976 in nearly all villages at least 50% of adult women

Table 17.5 Regression models of fertility change between 1976 and 1996

Predictor variables	Initial OLS model			
	Standardized beta coefficient	t-score	p-value	Moran's $z(I)$
(Constant)		7.597	0	
Change in proportion married	0.053	1.164	0.245	11.91
Change in female adult illiteracy	0.134	3.341	0.001	3.06
Change in sex ratio at reproductive ages	0.097	2.421	0.016	7.52
Change in village population	0.009	0.226	0.821	2.21
TFR in 1976	−0.69	−16.601	0	9.39
Adjusted $R^2 = 0.62$				
$z(I)$ for residuals = 3.98				
	Spatially filtered model			
(Constant)		−0.336	0.737	
Change in female illiteracy non-spatial	0.12	2.92	0.004	
Change in female illiteracy spatial	0.113	1.741	0.083	
Change in proportion married non-spatial	0.039	0.91	0.364	
Change in proportion married spatial	0.073	1.16	0.247	
Change in village population non-spatial	−0.013	−0.358	0.721	
Change in village population spatial	0.09	2.346	0.02	
Change in sex ratio non-spatial	0.085	2.202	0.028	
Change in sex ratio spatial	−0.02	−0.353	0.724	
TFR in 1976 non-spatial	−0.659	−15.838	0	
TFR in 1976 spatial	−0.338	−5.639	0	
Adjusted R^2 = 0.64				

Dependent variable is change in village TFR

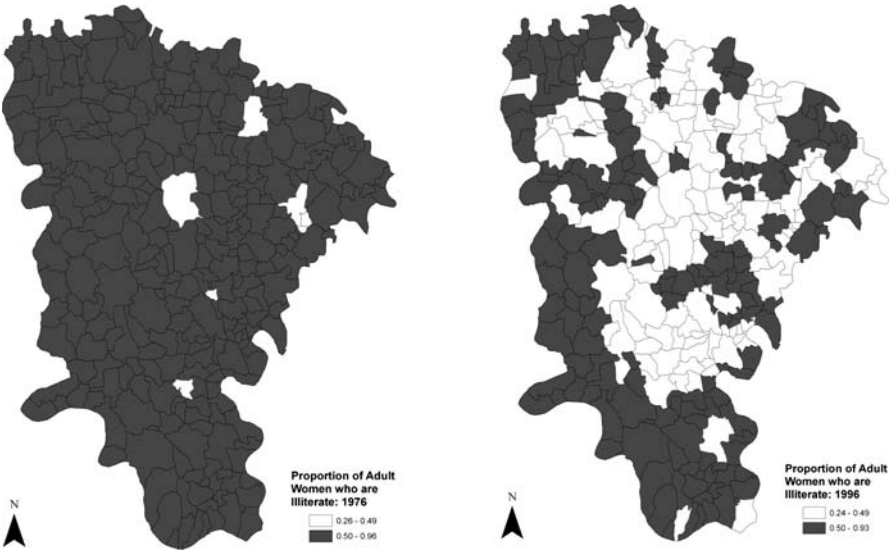


Fig. 17.7 Spatial pattern of illiteracy change between 1976 and 1996

were illiterate. The exceptions to this rule were in some, but not all, of the more urban parts of the governorate, especially Shbin El Kom, in the central northern part of the governorate. By 1996 there had been a huge swath of villages that had dropped below 50% illiteracy and they were heavily concentrated in the northern part of the governorate. This pattern was sufficiently widespread so that only the non-spatial component of illiteracy was statistically significant in the spatially filtered regression (bottom panel of Table 17.5). It is likely that the spatial component of illiteracy was subsumed under the spatial pattern of the TFR in 1976, given the overall relationship between education and reproduction.

17.7 Discussion and Conclusion

We have used a better measure of fertility and a more spatially precise dasymetric mapping approach to confirm that between 1976 and 1986 there was little change in fertility in the rural governorate of Menoufia, Egypt, but there was considerable spatial variability in both of those years. As Weeks et al. (2000) had predicted, we found that in 1986 the governorate was poised for a rapid drop in fertility because of the rapid rise in female literacy that had not yet, in 1986, produced any clear decline in the average number of children being born to women. Our results show that the central and northern parts of the governorate were the sites of the most dramatic declines in fertility, rather than the southern part, as had been anticipated. This is due in large part, we assume, because the decline in illiteracy was more dramatic

in those parts of the governorate, and the results do confirm the expectation that education was a key predictor of changing fertility.

By itself, of course, education cannot directly determine fertility levels. Education is a distal, not a proximate determinant of fertility, which include especially age at marriage (which in Muslim countries is almost always the same as the age at first intercourse), contraception, and abortion (which is only available under very restricted circumstances in Egypt). It is noteworthy that the proportion married did not emerge as a variable helping to explain the decline in fertility in Menoufia, because we noted above that a change in the age at marriage has been posited as an important element of the Arab fertility transition. In 1976, 1986, and 1996 it helped to explain the spatial variability in fertility in the governorate, but it was not a change in marriage behavior that seems to have accounted for the decline in fertility between 1986 and 1996. In fact, there was practically no change in the percent married in most villages during that 10-year interval. The obvious implication is that fertility was accomplished by means of contraceptive utilization, rather than delayed marriage.

The census data themselves provide no clues about the possible role of family planning, but we can gain some insights using data from the Egypt Demographic and Health Surveys (http://www.measuredhs.com). In 1988, there were 345 married women of reproductive age included in the DHS in Menoufia sampled from 10 different villages, and in 1995 the sample included 507 women from 17 different villages. Between 1985 and 1995 there was a slight increase in the average age at marriage among women in the sampled villages, but most noticeably the percentage of women who had ever-used a modern method of contraception increased from 62% to 75% (see Table 17.6). The increase was especially noticeable among younger women. For example, women aged 20–24 increased their ever-use of modern contraceptives from 40% to 62%. Although the sample sizes are fairly small, that difference is large enough to be statistically significant. The current use of modern contraception increased from 39% in 1988 to 49% in 1995, and again it was the younger woman among whom the increase was most notable, increasing from 24% to 47% among women aged 20–24.

This rise seems plausible given the high percentage of women (three-fourths at both dates) who indicated that their husband approved of birth control. The data in Table 17.6 also show the likely source of the increase in the use of contraception, namely "family planning effort." Married women in Menoufia were switching from the pharmacy as a source of contraception to a hospital or clinic. This was part of a government effort to promote an increase in the use of contraception in order to lower the birth rate. The decline in mortality was fairly substantial during this time in Egypt. The US Census Bureau's International Programs Center estimates that the infant mortality rate (deaths during the first year per 1,000 live births) was 132 in 1976, 89 in 1986, and 49 in 1996 (US Census Bureau International Programs Center, n.d.). Furthermore, in 1976 childhood mortality rates were consistently higher for females than for males (Makinson, 1986) but that pattern had abated by 1986.

As the death rate goes down among children without a commensurate drop in fertility, the result is that an increasing fraction of children survive to adulthood,

Table 17.6 Results for Menoufia from the 1988 and 1995 Egyptian demographic and health surveys

	1988	1995
Age at marriage	17.7	18.8
Age at first birth	19.5	20.3
Percent ever having used a method of birth control	62	75
Percent aged 20–24 ever having used a method of birth control	40	62
Percent currently using a modern method of contraception	39	49
Percent aged 20–24 currently using a modern method of contraception	24	47
Percent whose husband approves of family planning	76	76
N of living children at first use	3.16	2.76
Source of last method = pharmacy	43	14
Source of last method = private physician	24	24
Source of last method = hospital or clinic	29	60
Percent of women with more than primary education-DHS	17	33
Percent of husbands with more than primary education-DHS	24	49
Percent of women with more than primary education-Census	7	17
Percent married-census	63	63

forcing families and the villages in which they live to adjust to ever larger numbers of young people who need to be clothed, fed, housed, and provided with a job. In 1976, the combination of fertility and mortality in Menoufia meant that the average women could expect to have 4.03 children survive to adulthood. By 1986 this had risen to 4.52 because death rates had dropped dramatically, but birth rates had not. Recognizing this, the Egyptian government implemented a family planning program in many governorates, including Menoufia, to promote the use of contraception. As we have seen, the birth rate did then drop dramatically, but even so a woman in Menoufia in 1996 could expect to have 4.24 children survive to adulthood – a greater number than 20 years earlier, despite the drop in fertility.

Was the drop in fertility between 1986 and 1996 due to this government-sponsored push to encourage the use of contraception, or was it due to the government-sponsored push to improve literacy among the rural villagers, in order to improve their economic productivity? Almost certainly both of those factors were at work, but we do not have the data to decompose their relative contributions. In general, more educated women are more likely to be contraceptors, so we can anticipate that type of interaction. We also know that there is a spatial clustering of villages where illiteracy dropped between 1986 and 1996 and fertility also dropped in those parts of the governorate. We can infer a cause and effect relationship, but we cannot confirm it. Importantly, though, we know that no matter how widespread both government programs might have been, villagers in different parts of the governorate responded differently. There are clusters of high fertility and low fertility, clusters of rapid change and clusters of slow or no change and our spatially filtered regression results suggest that some portion of the spatial pattern is a product of being in the neighborhood of villages where these phenomena are occurring, whether or not your own village may be very similar to those other villages. We have

thus shown that at this geographic scale it is important to know where a village is, not just what its demographic characteristics might be, if we are to understand the level of reproduction of women living in that place.

We end by noting that the story of fertility in Menoufia is not yet completely told. We know that fertility changed very little between the 1976 and 1986 censuses, but the data presented here show clearly that a rather dramatic fertility decline was occurring between 1986 and 1996. Keep in mind that our fertility data refer to behavior that was occurring on average 2.5 years prior to the census, so from a calendar perspective, we can say more accurately that there was little evidence of change in fertility between the 1970s and the early 1980s, but there was evidence of substantial change between the early 1980s and the early 1990s. Interestingly enough, the 1996 census seems to have captured the point at which fertility had at least temporarily leveled off. Our calculations from the 1995 DHS for Menoufia show that there is virtually no difference in the TFR as calculated from births in the year preceding the survey compared to the five years preceding the survey, suggesting no trend in reproductive behavior. By contrast, the 1988 DHS data showed that the TFR based on the year prior to that survey was lower than that based on the five years preceding the survey, implying that there was a downward trend over time to the data. Consistent with this idea is the finding from the Egyptian DHS for 2000 that fertility levels were not much different in 2000 in Menoufia than they had been in 1995 (El-Zanaty and Way, 2001). In fact, our analysis of the data for Menoufia (not shown) suggests, if anything, an upward trend in fertility based on births in the year preceding the 2000 DHS compared to the five years preceding the survey. The 2006 census of Egypt has not been completed as of this writing, but when those data become available we will be able to determine whether, for example, fertility stopped declining in villages where illiteracy had not decline as rapidly as in other villages. From such a fact we may infer that the village's fertility decline had been due to the government's family planning program (which was later substantially reduced in funding as a result of changing United States foreign assistance priorities), rather than to a more systemic change in the education of women.

References

Abler RF (1987) The National Science Foundation National Center for Geographic Information and Analysis. Int J Geogr Inf Syst 1:303–336

Ahuja N, Schachter BJ (1983) Pattern models. Wiley, New York

Aldstadt J, Getis A (2006) Using amoeba to create a spatial weights matrix and identify spatial clusters. Geogr Anal 38:327–343

Almeida-Neto M, Lewinsohn TM (2004) Small-scale spatial autocorrelation and the interpretation of relationships between phenological parameters. J Veg Sci 15:561–568

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht

Anselin L (1990) What is special about spatial data? Alternative perspectives on spatial data analysis. In: Griffith DA (ed) Spatial statistics, past, present and future. Institute of Mathematical Geography, Ann Arbor, MI, pp 63–77

Anselin L (1991) SpaceStat: a program for the analysis of spatial data. Department of Geography, University of California, Santa Barbara, CA

Anselin L (1995) Local indicators of spatial association-LISA. Geogr Anal 27:93–115

Anselin L (1996) The Moran scatterplot as an exploratory spatial data analysis tool to assess local instability in spatial association. In: Fischer MM, Scholten H, Unwin D (eds) Environmental modeling with GIS. Oxford University Press, Oxford, pp 454–469

Anselin L, Getis A (1992) Spatial statistical analysis and geographic information systems. Ann Reg Sci 26:19–33

Anselin L, Griffith DA (1988) Do spatial effects really matter in regression analysis? Pap Reg Sci 65:11–34

Anselin L, Rey SJ (1991) Properties of tests for spatial dependence in linear regression models. Geogr Anal 23:112–131

Anselin L, Florax RJGM, Rey SJ (2004a) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin

Anselin L, Florax RJGM, Rey SJ (2004b) Econometrics for spatial models: recent advances. In: Anselin L, Florax R, Rey S (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin, pp 1–25

Anselin L, Kim YW, Syrabi I (2004c) Web-based spatial analysis tools for the exploration of spatial data. J Geogr Syst 6:197–218

Arbia G (1989) Spatial data configuration in statistical analysis of regional economic and related problems. Kluwer, Dordrecht

Arctur D, Zeiler M (2004) Designing geodatabases: case studies in GIS data modeling. ESRI, Redlands, CA

Arriaga EE (1994) Population analysis with microcomputers. Working paper, United States Bureau of the Census, Washington, DC

Atalik G (1990) Some effects of regional differentiation on integration in the European community. Pap Reg Sci Assoc 69:11–19

Atkins DE, Droegemeier KK, Feldman SI, Garcia-Molina H, Klein ML, Messerschmitt DG, Messina P, Ostriker JP, Wright MH (2003) Revolutionizing science and engeineering through cyberinfrastructure. Working paper, National Science Foundation

Bailey TC (1990) GIS and simple systems for visual, interactive, spatial analysis. Cartogr J 27: 79–84

Bailly A, Gibson LJ (2004) Applied geography: a world perspective. Kluwer, Dordrecht

Barberis IM, Tanner EVJ (2005) Gaps and root trenching increase tree seedling growth in Panamanian semi-evergreen forest. Ecology 86:667–674

Bartels CPA (1979) Operational statistical methods for analysing spatial data. In: Bartels CPA, Ketellapper RH (eds) Exploratory and explanatory statistical analysis of spatial data. Martinus Nijhoff, Boston

Bartlett MS (1950) Periodogram analysis and continuous spectra. Biometrika 37:1–16

Bartlett MS (1963) The spectral analysis of point processes. J R Stat Soc B 25:264–296

Bartlett MS (1964) The spectral analysis of two-dimensional point processes. Biometrika 51: 299–311

Becker RA, Chambers JM, Wilks AR (1988) The new S language: a programming environment for data analysis and graphics. Wadsworth, Pacific Grove, CA

Bennett RJ (1979) Spatial time series. Pion, London

Berglund S, Karlström A (1999) Identifying local spatial association in flow data. J Geogr Syst 1:219–236

Berry BJL, Marble DF (1968) Spatial analysis. Prentice-Hall, Englewood Cliffs, NJ

Besag JE (1977a) Discussion following Ripley. J R Stat Soc B 39:193–195

Besag JE (1977b) Efficiency of pseudo-likelihood estimation for simple gaussian fields. Biometrika 64:616–618

Besag JE (1977c) Errors in variable estimation for Gaussian lattice schemes. J R Stat Soc B 39: 73–78

Bivand R (1990) Spatial statistics: front-end interference support for GIS. In: Proceedings Third Scandinavian Research Conference on Geographical Information Systems. Helsingor, Denmark, pp 244–254

Bivand R (1991) SYSTAT-compatible software for modelling spatial dependence among observations. In: Paper Presented at the 7th European Colloquium on Theoretical and Quantitative Geography. Stockholm, Sweden

Bivand R (2002) Spatial econometrics functions in R: classes and methods. J Geogr Syst 4:405–421

Bolduc D, Laferriere R, Santarossa G (1992) Spatial autoregressive error components in travel flow models. Reg Sci Urban Econ 22:371–385

Bolduc D, Laferriere R, Santarossa G (1995) Spatial autoregressive error components in travel flow models: an application to aggregate mode choice. In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics. Springer, Berlin, pp 96–108

Bonham-Carter G, Agterberg F, Wright D (1988) Integration of geological datasets for gold exploration in Nova Scotia. J Photogramm Remote Sens 54:1585–1592

Boots BN (1985) Voronoi (Thiessen) polygons, Catmog no 45. Geo Books, Norwich

Boots BN, Dufournaud C (1994) A programming approach to minimizing and maximizing spatial autocorrelation statistics. Geogr Anal 26:54–66

Boots BN, Getis A (1987) Point pattern analysis. Sage, Newbury Park

Boots BN, Tiefelsdorf M (2000) Global and local spatial autocorrelation in bounded regular tessellations. J Geogr Syst 2:319–348

Brandsma AS, Kelletaper RH (1979) A biparametric approach to spatial autocorrelation. Environ Plan A 11:51–58

Burrough PA (1990) Methods of spatial analysis in GIS. Int J Geogr Inf Syst 4:221–223

Busing RT (1996) Estimation of tree replacement patterns in an Appalachian *Picea Abies* forest. J Veg Sci 7:685–694

Call LJ, Nilsen ET (2003) Analysis of spatial patterns and spatial association between the invasive tree-of-heaven (*Ailanthus altissima*) and the native black locust (*Robinia pseudoacacia*). Am Midl Nat 150:1–14

Camarero JJ, Gutierrez E, Fortin MJ, Ribbens E (2005) Spatial patterns of tree recruitment in a relict population of *Pinus uncinata*: forest expansion through stratified diffusion. J Biogeogr 32:1979–1992

Chan KL (1985) Singapore's dengue haemorrhagic fever control programme: a case study on the successful control of *Aedes aegypti* and *Aedes albopictus* using mainly environmental measures as a part of integrated vector control. Southeast Asia Medical Information, Tokyo

Chapeau-Blondeau F, Monir A (2002) Numerical evaluation of the Lambert W function and application to generation of generalized Gaussian noise with exponent 1/2. IEEE Trans Signal Process 50:2160–2165

Chen C (2006) CiteSpace II: detecting and visualizing emerging trends and transiet patterns in scientific literature. J Am Soc Inf Technol 57:359–377

Chen JQ, Bradshaw GA (1999) Forest structure in space: a case study of an old growth spruce-fir forest in Changbaishan Natural Reserve, PR China. For Ecol Manage 120:219–233

Chen D, Getis A (1998) Point pattern analysis (PPA). Software manual. Department of Geography, San Diego State University, San Diego, CA

Chicago Area Transportation Study (1980) 1979 Traveland household characteristics survey. Chicago Area Transportation Study, Chicago

Clark DA, Clark DB (1984) Spacing dynamics of a tropical rain forest tree: evaluation of the Janzen–Connell model. Am Nat 124:769–788

Clark JS, Silman M, Kern R, Macklin E, HilleRisLambers J (1999) Seed dispersal near and far: patterns across temperate and tropical forests. Ecology 80:1475–1494

Cleveland WS, McGilI ME (1988) Dynamic graphics for statistics. Wadsworth, Pacific Grove, CA

Cliff AD, Ord JK (1969) The problem of spatial autocorrelation. In: Scott A (ed) Studies in regional science. Pion, London, pp 25–55

Cliff AD, Ord JK (1973) Spatial autocorrelation. Pion, London

Cliff AD, Ord JK (1975) Model building and the analysis of spatial pattern in human geography. J R Stat Soc B 37:297–348

Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion, London

Cliff AD, Martin RL, Ord JK (1974) Evaluating the friction of distance parameter in gravity models. Reg Stud 8:281–286

Cole RG, Syms C (1999) Using spatial pattern analysis to distinguish causes of mortality: an example from kelp in north-eastern New Zealand. J Ecol 87:963–972

Condit R, Ashton PS, Baker P, Bunyavejchewin S, Gunatilleke S, Gunatilleke N, Hubbell SP, Foster RB, Itoh A, LaFrankie JV, Lee HS, Losos E, Manokaran N, Sukumar R, Yakamura T (2000) Spatial patterns in the distribution of tropical tree species. Science 288:1414–1418

Connell JH (1971) On the role of natural enemies in preventing competitive exclusion in some marine animals and rain forest trees. In: den Boer PJ, Gradwell GR (eds) Dynamics of numbers in populations. Centre for Agricultural Publishing and Documentation, Wageningen, pp 298–312

Connell JH, Tracey JG, Webb LJ (1984) Compensatory recruitment, growth, and mortality as factors maintaining rain forest tree diversity. Ecol Monogr 54:141–164

Consoli GB, de Oliveira RL (1994) Principais Mosquitos de Importancia Sanitaria no Brasil Rotrauta. Working paper. FIOCRUZ, Rio de Janeiro

Continental Illinois National Bank (1978) Industrial labor sheds: suburban area of metropolitan Chicago. Area Development Division, Continental Illinois National Bank, Chicago

Cooper CF (1961) Pattern in ponderosa pine forests. Ecology 42:493–499

Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE (1996) On the Lambert W function. Adv Comput Math 5:329–359

Corless RM, Jeffrey DJ, Knuth DE (1997) A sequence of series for the Lambert W function. In: Proceedings of the 1997 international symposium on Symbolic and algebraic computation. ACM, New York, pp 197–204

Couclelis H (1991) Requirements for planning-relevant GIS: a spatial perspective. Pap Reg Sci 70:9–19

Couteron P, Kokou K (1997) Woody vegetation spatial patterns in a semi-arid savanna of Burkina Faso, West Africa. Plant Ecol 132:211–227
Cowen DJ (1988) GIS versus CAD versus DBMS: what are the differences. Photogramm Eng Remote Sens 54:1551–1555
Cox DR, Miller HD (1965) The theory of stochastic processes. Methuen, London
Cressie N (1985) Fitting variogram models by weighted least squares. Math Geol 17:563–586
Cressie N (1986) Kriging nonstationary data. J Am Stat Assoc 81:625–634
Cressie N (1993) Statistics for spatial data. Wiley, New York
Cressie N, Chan N (1989) Spatial modeling of regional variables. J Am Stat Assoc 84:393—401
Cressie N, Head THC (1989) Spatial data analysis of regional counts. Biom J 31:699–719
Crook DA, Robertson AI, King AJ, Humphries P (2001) The influence of spatial scale and habitat arrangement on diel patterns of habitat use by two lowland river fishes. Oecologia 129:525–533
Crosier SJ, Goodchild MF, Hill LL, Smith TR (2003) Developing an infrastructure for sharing environmental models. Environ Plann B Plann Des 30:487–501
Csillag F (1991) Merging GIS and spatial statistics. In: Workbook for IGU-GIS conference on multiple representations and multiple uses. Masaryk University, Brno, Czechoslovakia
Dacey MF (1965) A review on measures of contiguity for two and k-color maps. Technical report no 2, spatial diffusion study. Department of Geography, Northwestern University, Evanston
Dale MRT, Powell RD (2001) A new method for characterizing point patterns in plant ecology. J Veg Sci 12:597–608
Davis JC (1986) Statistics and data analysis in geology. Wiley, New York
Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38
Deng M (2008) An anisotropic model for spatial processes. Geogr Anal 40(1):26–51
Deutschman DH, Bradshaw GA, Childress WM, Daly KL, Grunbaum D, Pascual M, Shumaker NH, Wu J (1993) Mechanisms of patch formation. In: Levin S, Powell T, Steele J (eds) Patch dynamics. Springer, New York, pp 184–209
Diggle PJ (1979) Statistical methods for spatial point patterns in ecology. In: Cormack RM, Ord JK (eds) Spatial and temporal analysis in ecology, vol 8. International Cooperative Publishing House, Fairland, MD, pp 95–150
Diggle PJ (1981) Some graphical methods in the analysis of spatial point patterns. In: Barnett V (ed) Interpreting multivariate data. Wiley, Chichester
Diggle PJ (1983) Statistical analysis of spatial point patterns. Academic, London
Diggle PJ (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. J R Stat Soc A 153:349–362
Diggle PJ, Rowlingson BS (1994) A conditional approach to point process modelling of elevated risk. J R Stat Soc A 157:433–440
Ding Y, Fotheringham AS (1991) The integration of spatial analysis and GIS. National Center for Geographic Information and Analysis, Buffalo, NY
Ding Y, Fotheringham AS (1992) The integration of spatial analysis and GIS. Comput Environ Urban Syst 16:3–19
Dixon PM (2002) Ripely's K function. In: El-Shaawawi AH, Piegorsch WW (eds) Encyclopedia of environmentrics. Wiley, Chichester
Donnegan JA, Rebertus AJ (1999) Rates and mechanisms of subalpine forest succession along an environmental gradient. Ecology 80:1370–1384
Druckenbrod DL, Shugart HH, Davies I (2005) Spatial pattern and process in forest stands within the Virginia piedmont. J Veg Sci 16:37–48
Durbin J (1960) Estimation of parameters in time-series regression models. J R Stat Soc B 22: 139–153
Edman JD, Scott TW, Costero A, Morrison AC, Harrington LC, Clark GG (1998) *Aedes aegypti* (Diptera: Culicidae) movement influenced by availability of oviposition sites. J Med Entomol 35:578–583
El-Zanaty FH, Way AA (2001) Egypt demographic and health survey, 2000. Ministry of Health and Population, National Population Council (Egypt), Cairo, Egypt

Entwisle B, Casterline JB, Sayed HAA (1989) Villages as contexts for contraceptive behavior in rural Egypt. Am Sociol Rev 54:1019–1034

Fargues P (1997) State policies and the birth rate in Egypt: from socialism to liberalism. Popul Dev Rev 23:115–138

Farley JA, Limp WF, Lockhart J (1990) The archaeologist's workbench: integrating GIS, remote sensing, EDA and database management. In: Allen KMS, Green FSW, Zubrow EBW (eds) Interpreting space: GIS and archaeology. Taylor & Francis, London, pp 141–164

Fellmann JD, Getis A, Getis J (2008) Human geography, landscapes of human activities, 10th edn. McGraw-Hill, New York

Fischer MM (1997) Computational neural networks: a new paradigm for spatial analysis. Environ Plan A 30:1873–1891

Fischer MM (2002) Learning in neural spatial interaction models: a statistical perspective. J Geogr Syst 4:287–299

Fischer MM, Getis A (1997a) Advances in spatial analysis. In: Fischer MM, Getis A (eds) Recent developments in spatial analysis: spatial statistics, behavioral modeling, and computational intelligence. Springer, Berlin, Heidelberg and New York, pp 1–14

Fischer MM, Getis A (1997b) Recent developments in spatial analysis: spatial statistics, behavioral modeling and computational intelligence. Springer, Berlin, Heidelberg and New York

Fischer MM, Leung Y (1998) A genetic-algorithms based evolutionary computational neural network for modelling spatial interaction dataNeural network for modelling spatial interaction data. Ann Reg Sci 32:437–458

Fischer MM, Reggiani A (2004) Spatial interaction models: from the gravity to the neural network approach. In: Cappello R, Nijkamp P (eds) Urban dynamics and growth. advances in urban economics. Elsevier, Amsterdam, pp 319–346

Fischer MM, Reismann M, Scherngell T (2006a) Spatial interaction and spatial autocorrelation. In: Paper presented at the International Workshop on Spatial Econometrics and Statistics. Rome, Italy

Fischer MM, Scherngell T, Jansenberger E (2006b) The geography of knowledge spillovers between high-technology firms in Europe: evidence from a spatial interaction modeling perspective. Geogr Anal 38:288–309

Florax RJGM, de Graaff T (2004) The performance of diagnostic tests for spatial autocorrelation in linear regression models: a meta-analysis of simulation studies. In: Anselin L, Florax R, Rey S (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin, pp 29–65

Florax RJGM, Rey SJ (1995) The impacts of misspecified spatial interaction in linear regression models. In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics. Springer, Berlin

Flowerdew R, Green M, Kehris E (1991) Using areal interpolation methods in geographic information systems. Pap Reg Sci 70:303–315

Flury B, Zoppè A (2000) Exercises in EM. Am Stat 54:207–209

Focks DA, Chadee DD (1997) Pupal survey: an epidemiologically significant surveillance method for *Aedes aegypti*: an example using data from Trinidad. Am J Trop Med Hyg 56:159–167

Focks DA, Haile DG, Daniels E, Mount GA (1993) Dynamic life table model for *Aedes aegypti* (Diptera: Culicidae): simulation results and validation. J Med Entomol 30:10018–28

Focks DA, Daniels E, Haile DG, Keesling JE (1995) A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. Am J Trop Med Hyg 53:489–506

Fonseca MG, Martini MZ, dos Santos FAM (2004) Spatial structure of Aspidosperma polyneuron in two semi-deciduous forests in Southeast Brazil. J Veg Sci 15:41–48

Forget PM, Mercier F, Collinet F (1999) Spatial patterns of two rodent-dispersed rain forest trees *Carapa procera* (Meliaceae) and *Vouacapoua americana* (Caesalpinaceae) at Paracou, French Guiana. J Trop Ecol 15:301–313

Fortin MJ, Dale MRT (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge

Foster SA, Gorr WL (1986) An adaptive filter for estimating spatially-varying parameters: application to modeling police hours spent in response to calls for service. Manage Sci 32:878–889
Fotheringham AS (1983) A new set of spatial-interaction models: the theory of competing distances. Environ Plan A 15:15–36
Fotheringham AS (1997) Trends in quantitative methods I: stressing the local. Prog Hum Geogr 21:88–96
Fotheringham AS, Rogerson P (1994) Spatial analysis and GIS. Taylor & Francis, London
Fotheringham AS, Charlton ME, Brunsdon C (1996) The geography of parameter space: an investigation of spatial non-stationarity. Int J Geogr Inf Syst 10:605–627
Fotheringham AS, Brunsdon C, Charlton ME (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, New York
Fowler N (1986) The role of competition in plant-communities in arid and semiarid regions. Annu Rev Ecol Syst 17:89–110
Franklin J, Rey SJ (2007) Spatial patterns of tropical forest trees in Western Polynesia suggest recruitment limitations during secondary succession. J Trop Ecol 23:1–12
Franklin J, Michaelsen J, Strahler AH (1985) Spatial analysis of density dependent pattern in coniferous forest stands. Vegetatio 64:29–36
Gadalla MS (1978) Is there hope? Fertility and family planning in a rural Egyptian community. Carolina Population Center, University of North Carolina, Chapel Hill, NC
Gaines KF, Bryan Jr AL, Dixon PM (2000) The effects of drought on foraging habitat selection of breeding wood storks in coastal Georgia. Waterbirds 23:64–73
Gatrell A (1987) On putting some statistical analysis into geographic information systems: with special reference to problems of map comparison and map overlay. Working Paper Research Report No 5. Northern Regional Research Laboratory
Geary RC (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–145
Getis A (1957) A geographical analysis of rail freight shipments in Pennsylvania. Pa Bus Surv 51:4–5
Getis A (1963) The determination of the location of retail activities with the use of a map transformation. Econ Geogr 39:14–22
Getis A (1964) Temporal land use pattern analysis with the use of nearest neighbor and quadrat method. Ann Assoc Am Geogr 54:391–399
Getis A (1969) Residential location and the journey to work. Proc Assoc Am Geogr 1:55–59
Getis A (1983) Second-order analysis of point patterns: the case of Chicago as a multi-center urban region. Prof Geogr 35:73–80
Getis A (1984) Interaction modeling using second-order analysis. Environ Plan A 16:173–183
Getis A (1985a) Energy costs and land use patterns in metropolitan Chicago. In: Checkoway B, Patton CV (eds) The metropolitan Midwest: policy problems and prospects for change. University of Illinois Press, Urbana, IL, chap. 5
Getis A (1985b) A second-order approach to spatial autocorrelation. Ont Geogr 25:67–73
Getis A (1985c) Urban population spacing analysis. Urban Geogr 6:3–12
Getis A (1989a) A spatial association model approach to the identification of spatial dependence. Geogr Anal 21:251–259
Getis A (1989b) A spatial causal model of economic interdependency among neighboring communities. Environ Plan A 21:115–120
Getis A (1990) Screening for spatial dependence in regression analysis. Pap Reg Sci Assoc 69: 69–81
Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. Environ Plan A 23:1269–1277
Getis A (1993a) Introduction: mathematical models in geography. Pap Reg Sci 72:201–202
Getis A (1993b) Scholarship, leadership, and quantitative methods. Urban Geogr 14:517–525
Getis A (1995a) Spatial filtering in a regression framework: examples using data on urban crime, regional inequality, and government expenditures. In: Anselin L, Florax R (eds) New directions in spatial econometrics. Springer, Berlin, pp 172–188

Getis A (1995b) Spatial filtering in a regression framework: examples using data on urban crime, regional inequality, and government expenditures. In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics. Springer, Berlin
Getis A (1995c) The tyranny of data. San Diego State University Press, San Diego, CA
Getis A (1999) Some thoughts on the impact of large data sets on regional science. Ann Reg Sci 33:145–150
Getis A (2004a) A geographic approach to identifying disease clusters. In: Janelle DG, Warf B, Hansen K (eds) Worldminds: geographical perspectives on 100 problems. Kluwer, Dordrecht, pp 81–86
Getis A (2004b) The role of geographic information science in applied geography. In: Bailly A, Gibson LJ (eds) Applied geography: a world perspective. Kluwer, Dordrecht, pp 95–112
Getis A (2007) Reflections on spatial autocorrelation. Reg Sci Urban Econ 37:491–496
Getis A, Aldstadt J (2004) Constructing the spatial weights matrix using a local statistic. Geogr Anal 36:90–105
Getis A, Boots BN (1978) Models of spatial processes: an approach to the study of point, line, and area patterns. Cambridge University Press, Cambridge
Getis A, Franklin J (1987) Second-order neighborhood analysis of mapped point patterns. Ecology 68:473–477
Getis A, Getis J (1968) Retail store spatial affinities. Urban Stud 5:317–332
Getis A, Griffith DA (2002) Comparative spatial filtering in regression analysis. Geogr Anal 34:130–140
Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24:189–206
Getis A, Ord JK (1996) Local spatial statistics: an overview. In: Longley P, Batty M (eds) Spatial analysis: modelling in a GIS environment. Geoinformation International, Cambridge, UK, pp 261–278
Getis A, Ord JK (1998) Spatial modelling of disease dispersion using a local statistic: the case of aids. In: Griffith DA, Amrhein CG, Huriot JM (eds) Econometric advances in spatial modelling and methodology: essays in honour of Jean Paelinck. Kluwer, Dordrecht
Getis A, Drummy P, Gartin J, Gorr WL, Harries K, Rogerson P, Stoe D, Wright R (2000) Geographic information science and crime analysis. URISA J 12:7–14
Getis A, Getis J, Quastler I (2001) The United States and Canada: the land and the people, 2nd edn. McGraw-Hill, New York
Getis A, Morrison A, Gray K, Scott TW (2003) Characteristics of the spatial pattern of the Dengue vector *Aedes Aegypti*, in Iquitos, Peru. Am J Trop Med Hyg 69:494–505
Getis A, Anselin L, Lea A, Ferguson M, Miller H (2004a) Spatial analysis and modeling in a GIS environment. In: McMaster RB, Usery EL (eds) A research agenda for geographic information science. CRC, Boca Raton, FL, pp 157–196
Getis A, Mur J, Zoller HG (2004b) Spatial econometrics and spatial statistics. Palgrave Macmillan, London
Getis A, Getis J, Getis V, Fellmann JD (2008) Introduction to geography, 11th edn. Mc-Graw-Hill, New York
Glass L, Tobler WR (1971) Uniform distribution of objects in a homogeneous field: cities on a plain. Nature 233:67–68
Good BJ, Whipple SA (1982) Tree spatial patterns – South-Carolina bottomland and swamp forests. Bull Torrey Bot Club 109:529–536
Goodchild MF (1987) A spatial analytical perspective on geographical information systems. Int J Geogr Inf Sci 1:327–334
Goodchild MF (1992) Geographical data modeling. Comput Geosci 18:401–408
Goodchild MF, Brusegard D (1989) Spatial analysis using GIS: seminar workbook. National Center for Geographic Information and Analysis, Santa Barbara, CA
Goodchild MF, Gopal S (1989) The accuracy of spatial databases. CRC, Boca Raton
Goodchild MF, Haining RP, Wise S (1992) Integrating GIS and spatial data analysis: problems and possibilities. Int J Geogr Inf Sci 6:407–423

Goreaud F, Pelissier R (1999) On explicit formulas of edge effect correction for ripley's k-function. J Veg Sci 10:433–438
Gould P (1981) Letting the data speak for themselves. Ann Assoc Am Geogr 71:166–176
Granger CWJ (1969) Spatial data and time series analysis. In: Scott A (ed) Studies in regional science. Pion, London, pp 1–24
Grieg-Smith P (1983) Quantitative plant ecology, 3rd edn. Blackwell, Oxford
Griffin P, Getis A, Griffin E (1996) Regional patterns of affirmative action compliance costs. Ann Reg Sci 30:321–340
Griffith DA (1981) Evaluating the transformation from a monocentric to a polycentric city. Prof Geogr 33:189–196
Griffith DA (1988a) Advanced spatial statistics: special topics in the exploration of quantitative spatial data series, vol 12. Kluwer, Dordrecht
Griffith DA (1988b) Estimating spatial autoregressive model parameters with commercial statistical packages. Geogr Anal 20:176–186
Griffith DA (1993) Advanced spatial statistics for analyzing and visualizing geo-referenced data. Int J Geogr Inf Syst 7:107–123
Griffith DA (1996) Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In: Arlinghaus SL, Griffith DA (eds) Practical handbook of spatial statistics. CRC, Boca Raton, http://www.loc.gov/catdir/enhancements/fy0731/95024710-d.html
Griffith DA (2000) Eigenfunction properties and approximations of selected incidence matrices employed in spatial anlaysis. Linear Algebra Appl 321:95–112
Griffith DA (2003) Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin
Griffith DA, Peres-Neto PR (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analysis. Ecology 87:2603–2613
Griffith DA, Bennett RJ, Haining RP (1989) Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data. Environ Plan A 21:1511–1523
Griffith DA, Lewis R, Li B, Vasiliev I, McKnight S, Yang X (1990) Developing Minitab software for spatial statistical analysis: a tool for education and research. Oper Geogr 8:28–33
Gu WD, Kuusinen M, Konttinen T, Hanski I (2001) Spatial pattern in the occurrence of the lichen *Lobaria pulmonaria* in managed and virgin boreal forests. Ecography 24:139–150
Gubler DJ (1993) Dengue and dengue haemorrhagic fever in the Americas. In: Thoncharoen P (ed) Monograph on dengue/dengue hemorrhagic fever. WHO regional publication SEARO no 22. World Health Organization, New Delhi, pp 9–22
Gubler DJ (1997) Dengue and dengue hemorrhagic fever: its history and resurgence as a global public health problem. In: Gubler DJ, Kuno G (eds) Dengue and dengue hemorrhagic fever. CAB International, Wallingford, Oxon, UK, pp 1–22, http://www.loc.gov/catdir/enhancements/fy0605/97013184-d.html
Guerra MA, Walker ED, Kitron U (2001) Canine surveillance system for *Lyme borreliosis* in Wisconsin and northern Illinois: geographic distribution and risk factor analysis. Am J Trop Med Hyg 65:546–552
Gujarati D (1992) Essentials of econometrics. McGraw-Hill, New York
Haase P (1995) Spatial pattern-analysis in ecology based on Ripley K-function – introduction and methods of edge correction. J Veg Sci 6:575–582
Haase P, Pugnaire FI, Clark SC, Incoll LD (1996) Spatial patterns in a two-tiered semi-arid shrubland in southeastern Spain. J Veg Sci 7:527–534
Haining RP (1977) Model specification in stationary random fields. Geogr Anal 9:107–129
Haining RP (1978) Estimating spatial-interaction models. Environ Plan A 10:305–320
Haining RP (1990a) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge, http://www.loc.gov/catdir/description/cam024/90001361html
Haining RP (1990b) The use of added variable plots in regression modelling with spatial data. Prof Geogr 42:336–344

Haining RP (1991) Bivariate correlation with spatial data. Geogr Anal 23:210–227
Haining RP, Griffith DA, Bennett. RJ (1984) A statistical approach to the problem of missing spatial data using a first-order Markov model. Prof Geogr 36:338–345
Halstead SB (1990) Global epidemiology of dengue hemorrhagic fever. Southeast Asian J Trop Med Public Health 21:636–41
Halstead SB, Scanlon JE, Umpaivit P, Udomsakdi S (1969) Dengue and Chickungunya virus infection in man in Thailand, 1962–(1964) IV. Epidemiologic studies in the Bangkok metropolitan area. Am J Trop Med Hyg 18:997–1033
Hand D, Daly F, Lunn A, McConway K, Ostrowski. E (1994) A handbook of small data sets. Chapman & Hall, New York
Harrington LC, Edman JD, Scott TW (2001) Why do female *Aedes aegypti* (Diptera: Culicidae) feed preferentially and frequently on human blood? J Med Entomol 38:411–422
Harrison D, Rubinfeld D (1978) Hedonic housing prices and the demand for clean air. J Environ Econ Manage 5:81–102
Haslett J, Wills G, Unwin A (1990) SPIDER – an interactive statistical tool for the analysis of spatially distributed data. Int J Geogr Inf Sci 4:285–296
Haslett J, Bradley R, Craig P, Unwin A, Wills G (1991) Dynamic graphics for exploring spatial data with application to locating global and local anomalies. Am Stat 45:234–242
Hawkins D, Olwell D (1998) Cumulative sum charts and charting for quality improvement. Springer, Berlin
Haynes KE, Fotheringham AS (1984) Gravity and spatial interaction models. Sage, Newbury Park, CA
He FL, Duncan RP (2000) Density-dependent effects on tree survival in an old-growth douglas fir forest. J Ecol 88:676–688
Hepple LW (1995) Bayesian techniques in spatial and network econometrics: 2 Computational methods and algorithms. Environ Plan A 27:615–644
HMSO (1987) Handling geographic information (the chorley report). London
Hoeffding W (1951) A combinatorial central limit theorem. Ann Math Stat 22:558–566
Howe HF (1986) Seed dispersal by fruit-eating birds and mammals. In: Murray J (ed) Seed dispersal. Academic, Sydney, pp 123–187
Hubert L (1977) Generalized proximity function comparisons. Br J Math Stat Psychol 31:179–182
Hubert L (1979) Matching models in the analvsis of cross-classifications. Psychometriha 44:21–41
Hubert L, Golledge RG (1982) Measuring association between spatially defined variables: Tjostheim's index and some generalizations. Geogr Anal 14:273–278
Hubert L, Golledge RG, Costanzo C (1981) Generalized procedures for evaluating spatial autocorrelation. Geogr Anal 13:224–233
Hubert L, Golledge RG, Costanzo C, Gale N (1985) Measuring association between spatially defined variables: an alternative procedure. Geogr Anal 17:36–46
Ida H (2000) Treefall gap disturbance in an old-growth beech forest in southwestern japan by a catastrophic typhoon. J Veg Sci 11:825–832
Jaffe AB, Trajtenberg M (2002) Patents, citations, and innovations: A window on the knowledge economy. MIT, Cambridge, MA, http://www.loc.gov/catdir/toc/fy032/2001056257html
Janzen DH (1970) Herbivores and the number of tree species in tropical forests. Am Nat 104: 501–528
Johnson NL, Kotz S, Kemp AW, Johnson NL (1992) Univariate discrete distributions, 2nd edn. Wiley, New York, http://www.loc.gov/catdir/description/wiley031/92011685html
Kashian DM, Turner MG, Romme WH, Lorimer CG (2005) Variability and convergence in stand structural development on a fire-dominated subalpine landscape. Ecology 86:643–654
Kehris E (1990a) A geographical modelling environment built around ARC/INFO. Working Paper Research Report No 13. North West Regional Research Laboratory, Lancaster University
Kehris E (1990b) Spatial autocorrelation statistics in ARC/INFO. Working Paper Research Report No 16. North West Regional Research Laboratory, Lancaster University

Kelejian H, Robinson D (2004) The influence of spatially correlated heteroscedasticity on tests for spatial autocorrelation. In: Anselin L, Florax R, Rey S (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin, pp 79–97

Kenkel NC (1988) Pattern of self-thinning in jack pine: testing the random mortality hypothesis. Ecology 69:1017–1024

Kooijman S (1976) Some remarks on the statistical analysis of grids especially with respect to ecology. Ann Syst Res 5:113–132

Laessle AM (1965) Spacing and competition in natural stands of sand pine. Ecology 46:65–72

LaFrankie JV, Saw LG (2005) The understorey palm *Licuala* (Arecaceae) suppresses tree regeneration in a lowland forest in Asia. J Trop Ecol 21:703–706

Larsen DR, Bliss LC (1998) An analysis of structure of tree seedling populations on a lahar. Landsc Ecol 13:307–322

Lawson A (1993) On the analysis of mortality events associated with a prespecified fixed point. J R Stat Soc A 156:363–377

Leemans R (1991) Canopy gaps and establishment patterns of spruce (*Picea abies* (l) karst) in 2 old-growth coniferous forests in central Sweden. Vegetatio 93:157–165

LeSage JP (2004) A family of geographically weighted regression models. In: Anselin L, Florax R, Rey S (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin, pp 241–264

LeSage JP, Pace R (2005) Spatial econometric modeling of origin-destination flows. In: 52nd Annual North American Meetings of the Regional Science Association International. Las Vegas, NV

Liang Y, Guo LD, Ma KP (2004) Genetic structure of a population of the ectomycorrhizal fungus *Russula vinosa* in subtropical woodlands in southwest China. Mycorrhiza 14:235–240

Little R, Rubin D (1987) Statistical analysis with missing data. Wiley, New York

Longley PA, Goodchild MF, Maguire DJ, Rhind DW (1999) Geographical information systems, 2nd edn. Wiley, New York

Lookingbill TR, Zavala MA (2000) Spatial pattern of *Quercus ilex* and *Quercus pubescens* recruitment in *Pinus halepensis* dominated woodlands. J Veg Sci 11:607–612

MacDougall EB (1991) A prototype interface for exploratory analysis of geographic data. Working paper, Department of Landscape Architecture and Regional Planning, University of Massachusetts

Maguire DJ (1991) An overview and definition of GIS. In: Longley PA, Goodchild MF, Maguire DJ, Rhind DW (eds) Geographical information systems: principles and applications. Longman Scientific and Technical, Harlow, pp 9–20

Maguire DJ, Dangermond J (1991) The functionality of GIS. In: Maguire DJ, Goodchild MF, Rhind DW (eds) Geographical information systems: principles and applications, vol 1. Longman Scientific & Technical, Harlow, pp 319–335

Maguire DJ, Michael B, Goodchild MF (2005) GIS, spatial analysis and modelling. ESRI, Redlands, CA

Makinson C (1986) Sex differentials in infant and child mortality in Egypt. PhD thesis, Department of Sociology, Princeton University, Princeton, NJ

Maling DH (1989) Measurements from maps: principles and methods of cartometry. Pergamon, New York

Malkinson D, Kadmon R, Cohen D (2003) Pattern analysis in successional communities – an approach for studying shifts in ecological interactions. J Veg Sci 14:213–222

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Mapping Science Committee NRC (1993) Toward a coordinated spatial data infrastructure for the nation. Working paper, National Academy Press, Washington, DC

Mast JN, Veblen TT (1999) Tree spatial patterns and stand development along the pine-grassland ecotone in the colorado front range. Can J For Res – Revue Canadienne De Recherche Forestiere 29:575–584

Mast JN, Wolf JJ (2004) Ecotonal changes and altered tree spatial patterns in lower mixed-conifer forests, Grand Canyon National Park, Arizona, USA. Landsc Ecol 19:167–180
Matern B (1971) Doubly stochastic poisson processes in the plane. In: Patil GP, Pielou EC, Waters WE (eds) Statistical ecology. Pennsylvania State University Press, State College, PA, pp 195–213
McDonald PT (1977) Population characteristics of domestic *Aedes aegypti* (Diptera: Culicidae) in villages on the Kenya coast I. J Med Entomol 14:49–53
McDonald RI, Peet RK, Urban DL (2003) Spatial pattern of Quercus regeneration limitations and *Acer rubrum* invasion in a Piedmont forest. J Veg Sci 14:441–450
McLachlan G, Krishnan T (1997) The EM-algorithm and extensions. Wiley, New York
McMillen D, McDonald J (2004) Locally weighted maximum likelihood estimation: Monte Carlo evidence and an application. In: Anselin L, Florax R, Rey S (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin, pp 225–239
Meng X (1997) The EM algorithm. In: Kotz S, Read C, Banks D (eds) Encyclopedia of statistical sciences. Wiley, New York, pp 218–227
Mennis J (2002) Using geographic information systems to create and analyze statistical surfaces of population and risk for environmental justice analysis. Soc Sci Q 83:281–297
Miller C, Urban DL (1999) Forest pattern, fire, and climatic change in the Sierra Nevada. Ecosystems 2:76–87
Mitchell A (1999) The ESRI guide to GIS analysis. ESRI, Redlands, CA
Moellering H, Tobler WR (1972) Geographical variances. Geogr Anal 4:34–50
Moeur M (1993) Characterizing spatial patterns of trees using stem-mapped data. For Sci 39:756–775
Montgomery D, Peck E (1982) Introduction to linear regression analysis. Wiley, New York
Moran PAP (1948) The interpretation of statistical maps. J R Stat Soc B 10:243–251
Moravie MA, Roberts A (2003) A model to assess relationships between forest dynamics and spatial structure. J Veg Sci 14:823–834
Morrison A, Astete H, Chapilliquen F, Ramirez-Prada C, Diaz G, Getis A, Gray K, Scott TW (2004a) Evaluation of a sampling methodology for rapid assessment of aedes aegypti infestation levels in Iquitos, Peru. J Med Entomol 41:502–510
Morrison A, Gray K, Getis A, Astete H, Shihuincha M, Fochs D, Watts D and Scott TW (2004b) Temporal and geographic patterns of *Aedes aegypti* (diptera: Culicidae) production in Iquitos, Peru. J Med Entomol 41:1123–1142
Morrison AC, Getis A, Santiago M, Rigau-Perez JG, Reiter P (1998) Exploratory space–time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991–1992. Am J Trop Med Hyg 58:287–298
Mosteller F, Tukey JW (1977) Data analysis and regression: a second course in statistics. Addison-Wesley series in behavioral science. Addison-Wesley, Reading, MA
Mugglestone MA, Renshaw E (1996) A practical guide to the spectral analysis of spatial point processes. Comput Stat Data Anal 21:43–65
Munyekenye OG, Githeko AK, Zhou GF, Mushinzimana E, Minakawa N, Yan GY (2005) *Plasmodium falciparum* spatial analysis, western Kenya highlands. Emerg Infect Dis 11:1571–1577
Nathan R, Muller-Landau HC (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. Trends Ecol Evol 15:275–285
Navas ML, Goulard M (1991) Spatial pattern of a clonal perennial weed. Rubia Peregrina (Rubiaceae) in vineyards of southern France. J Appl Ecol 28:1118–1129
Nebert D (1993) Implementation of wide area information server (WAIS) software to disseminate spatial data on the internet. In: International ESRI User Conference. Palm Springs
Neyman J (1939) A new class of 'contagious' distributions, applications for entemology and bacteriology. Ann Math Stat 10:35–57
Ng EG, Peyton BW (1993) Block sparse Cholesky algorithms on advanced multi-processor computers. SIAM J Sci Comput 14:1034–1056
Nicotra AB (1998) Sex ratio variation and spatial distribution of *Siparuna grandiflora*, a tropical dioecious shrub. Oecologia 115:102–113

Nijkamp P (1988) The use of information systems for regional planning. R Econ Reg Urb 15: 759–780

Nijkamp P (1990) Geographical information systems in perspective. In: Scholten HJ, Stillwell JCH (eds) Geographical information systems for urban and regional planning, vol 17. Kluwer, Dordrecht

Nijkamp P, Rietveld P (1984) Information systems for integrated regional planning, vol. 149. North-Holland, Amsterdam

Nyerges TL (1993) Understanding the scope of GIS: its relationship to environmental modeling. In: Goodchild MF, Parks B, Steyaert L (eds) Environmental modeling with GIS. Oxford University Press, New York, pp 75–93

O'Brien D, Kaneene J, Getis A, Lloyd J, Rip M, Leader R (2000) Spatial and temporal distribution of selected canine cancers in michigan, USA, 1964–1994. Prev Vet Med 47:187–204

Okabe A, Yamada I (2001) The K-function method on a network and its computational implementation. Geogr Anal 33:271–290

Okabe A, Boots BN, Sugihara K, Chiu SN (2000) Spatial tessellations: concepts and applications of Voronoi diagrams, 2nd edn. Wiley, Chichester, http://www.loc.gov/catdir/description/wiley033/99013149html

Oliver CD, Larson BC (1990) Forest stand dynamics. McGraw-Hill, New York

Oliver MA, Webster R (1990) Kriging: a method of interpolation for geographical information systems. Int J Geogr Inf Sci 4:313–332

Openshaw S (1977) Optimal zoning systems for spatial interaction models. Environ Plan A 9: 169–184

Openshaw S (1984) The modifiable area unit problem, CATMOG 38. Geoabstracts, Norwich

Openshaw S (1990) Spatial analysis and geographical information systems: a review of progress and possibilities. In: Scholten HJ, Stillwell JCH (eds) Geographical information systems for urban and regional planning, vol 17. Kluwer, Dordrecht

Openshaw S, Taylor P (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N, Bennett RJ (eds) Statistical applications in the spatial sciences. Pion, London

Openshaw S, Charlton ME, Wymer C, Craft A (1987) A Mark I geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inf Syst 1:335–358

Openshaw S, Charlton ME, Craft A (1988) Searching for Leukaemia clusters using a geographical analysis machine. Pap Reg Sci 64:95–106

Openshaw S, Cross A, Charlton ME (1990) Building a prototype geographical correlates exploration machine. Int J Geogr Inf Sci 4:297–311

Openshaw S, Brunsdon C, Charlton ME (1991) A spatial analysis toolkit for gis. In: Proceedings of the Second European Conference on Geographical Information Systems. Brussels, Belgium, pp 788–796

Ord JK (1975) Estimation methods for models of spatial interaction. J Am Stat Assoc 70:120–126

Ord JK, Getis A (1993) Distributional issues concerning distance statistics. In: Paper Presented at the 40th North American Meeting of the Regional Science Association International. Houston, TX

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 27:286–306

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. J Reg Sci 41:411–432

Ordonez JG, Fernandez Salas I, Flores-Leal A (1997) Monitoring dispersal of marked *Aedes aegypti* females under field conditions using sticky ovitraps in Monterrey, northeastern Mexico. J Am Mosq Control Assoc 13:121

Pacala S, Silander J Jr (1985) Neighborhood models of plant population dynamics. I. Single-species models of annuals. Am Nat 125:385–411

Pace R, Gilley O (1997) Using the spatial configuration of the data to improve estimation. J Real Estate Fin Econ 14:333–340

PAHO (1994) Dengue and dengue haemorrhagic fever in the Americas. Guidelines for prevention and control. Pan American Health Organization Scientific Publication no 548. Pan American Health Organization

Palmiotto PA, Davies SJ, Vogt KA, Ashton MS, Vogt DJ, Ashton PS (2004) Soil-related habitat specialization in dipterocarp rain forest tree species in Borneo. J Ecol 92:609–623

Pancer-Koteja E, Szwagrzyk J, Bodziarczyk J (1998) Small-scale spatial pattern and size structure of *Rubus hirtus* in a canopy gap. J Veg Sci 9:755–762

Parish R, Antos JA, Fortin MJ (1999) Stand development in an old-growth subalpine forest in southern interior british columbia. Can J For Res – Revue Canadienne De Recherche Forestiere 29:1347–1356

Parker AJ, Parker KC, McCay DH (2001) Disturbance-mediated variation in stand structure between varieties of *Pinus clausa* (sand pine). Ann Assoc Am Geogr 91:28–47

Peet RK, Christensen NL (1987) Competition and tree death. Bioscience 37:586–595

Pelissier R (1998) Tree spatial patterns in three contrasting plots of a southern Indian tropical moist evergreen forest. J Trop Ecol 14:1–16

Pelissier R, Goreaud F (2001) A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation. J Veg Sci 12:99–108

Perry GLW, Miller BP, Enright NJ (2006) A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. Plant Ecol 187:59–82

Peuquet D (1984) A conceptual framework and comparison of spatial data models. Cartographica 21:66–113

Peuquet D (1988) Representations of geographic space: toward a conceptual synthesis. Ann Assoc Am Geogr 78:375–394

Pickles J (1995) Ground truth: the social implications of geographic information systems. Guilford, New York

Pielou E (1977) Mathematical ecology. Wiley, New York

Pindyck R, Rubinfeld D (1981) Econometric models and economic forecasts. McGraw-Hill, New York

Pinske J (2004) Moran-flavored tests with nuisance parameters: examples. In: Anselin L, Florax R, Rey S (eds) Advances in spatial econometrics: methodology, tools and applications. Springer, Berlin, pp 67–77

Pitman EJG (1937) The 'closest' estimates of statistical parameters. Biometrika 58:299–312

Plane D (1981) The geography of urban commuting fields: some empirical evidence from New England. Prof Geogr 33:182–188

Potvin F, Boots BN, Dempster A (2003) Comparison among three approaches to evaluate winter habitat selection by white-tailed deer on Anticosti Island using occurrences from an aerial survey and forest vegetation maps. Can J Zool – Revue Canadienne De Zoologie 81:1662–1670

Prior I, Muncke C, Parton R, Hancock J (2003) Direct visualization of Ras proteins in spatially distinct cell surface microdomains. J Cell Biol 160:165–170

Putnam J, Scott TW (1995) The effect of multiple host contacts on the infectivity of dengue-2 virus infected *Aedes aegypti*. J Parasitol 81:170–174

R Development Core Team (2004) R: a language and environment for statistical computing

Rashad H (2000) Demographic transition in Arab countries: a new perspective. J Popul Res 17: 83–101

Rashed T, Weeks JR, Gadalla MS, Hill AG (2001) Revealing the anatomy of cities through spectral mixture analysis of multispectral imagery: a case study of the greater Cairo region, Egypt. Geocarto Int 16:5–16

Rashed T, Weeks JR, Roberts D, Rogan J, Powell R (2003) Measuring the physical composition of urban morphology using multiple endmember spectral mixture models. Photogramm Eng Remote Sens 69:1011–1020

Rashed T, Weeks JR, Stow D, Fugate D (2005) Measuring temporal compositions of urban morphology through spectral mixture analysis: toward a soft approach to change analysis in crowded cities. Int J Remote Sens 26:699–718

Rayner JN (1971) An introduction to spectral analysis. Pion, London

Rayner JN, Golledge RG (1972) Spectral analysis of settlement patterns in diverse physical and economic environments. Environ Plan 4:347–371

Rayner JN, Golledge RG (1973) The spectrum of US Route 40 re-examined. Geogr Anal 5: 338–350

Reiter P, Gubler DJ (1997) Surveillance and control of urban dengue vectors. In: Gubler DJ, Kuno G (eds) Dengue and dengue hemorrhagic fever. CAB International, Wallingford, Oxon, UK, pp 425–462, http://www.loc.gov/catdir/enhancements/fy0605/97013184-d.html

Rey SJ, Anselin L (2007) PySAL: a Python library of spatial analytical methods. Rev Reg Stud 37:5–27

Rey SJ, Janikas MV (2006) STARS: space–time analysis of regional systems. Geogr Anal 38: 67–86

Ridley HN (1930) The dispersal of plants throughout the world. L Reeve, Ashford, Kent

Ripley BD (1976) The second-order analysis of stationary point processes. J Appl Probab 13: 255–266

Ripley BD (1977) Modelling spatial patterns (with discussion). J R Stat Soc B 39:172–212

Ripley BD (1978) Spectral analysis and the analysis of pattern in plant communities. J Ecol 66:965–981

Ripley BD (1979a) The analysis of geographic maps. In: Bartels CPA, Ketellapper RH (eds) Exploratory and explanatory statistical analysis of spatial data. Martinus Nijhoff, The Hague, pp 53–72

Ripley BD (1979b) Tests of 'randomness' for spatial point patterns. J R Stat Soc B 41:368–374

Ripley BD (1981) Spatial statistics. Wiley, New York

Robinson A, Sale R, Morrison J, Muehrcke P (1984) Elements of cartography, 5th edn. Wiley, New York

Rodhain F, Rosen L (1997) Mosquito vectors and dengue virus-vector relationships. In: Gubler DJ, Kuno G (eds) Dengue and dengue hemorrhagic fever. CAB International, Wallingford, Oxon, UK, pp 61–88, http://www.loc.gov/catdir/enhancements/fy0605/97013184-d.html

Rodriguez-Figueroa L, Rigau-Perez JG, Suarez E, Reiter P (1995) Risk factors for dengue infection during an outbreak in Yanes, Puerto Rico in 1991. Am J Trop Med Hyg 52:496–502

Rowlingson BS, Diggle PJ (1993) Splancs: spatial point pattern analysis code in s-plus. Comput Geosci 19:627–655

Santaló LA (1976) Integral geometry and geometric probability. Encyclopedia of mathematics and its applications, vol 1: Section, Probability. Addison-Wesley, Reading, MA

Schafer J (1997) Analysis of incomplete multivariate data. Chapman & Hall, New York

Schooley RL, Wiens JA (2001) Dispersion of kangaroo rat mounds at multiple scales in New Mexico, USA. Landsc Ecol 16:267–277

Schroff AZ, Lindgren BS, Gillingham MP (2006) Random acts of weevil: a spatial analysis of *Hylobius warreni* attack on *Pinus contorta* var. latifolia in the sub-boreal spruce zone of northern British Columbia. For Ecol Manage 227:42–49

Scott L (1999) The accessible city: employment opportunities in time and space. PhD thesis, Department of Geography, San Diego State University, San Diego

Scott TW, Morrison A, Lorenz LH, Clark GG, Strickman D, Kittayapong P, Zhou H, Edman JD (2000a) Longitudinal studies of *Aedes aegypti* (Diptera: Culicidae) in Thailand and Puerto Rico: population dynamics. J Med Entomol 37:77–88

Scott TW, Morrison AC, Lorenz LH, Clark GG, Strickman D, Kittayapong P, Zhou H, Edman JD (2000b) Longitudinal studies of *Aedes aegypti* (Diptera: Culicidae) in Thailand and Puerto Rico: blood feeding frequency. J Med Entomol 37:89–101

Sen AK, Smith TE (1995) Gravity models of spatial interaction behavior. Springer, Berlin

Shi H, Zhang LJ (2003) Local analysis of tree competition and growth. For Sci 49:938–955

Shi H, Laurent EJ, LeBouton J, Racevskis L, Hall KR, Donovan M, Doepker RV, Walters MB, Lupi F, Liu JG (2006) Local spatial modeling of white-tailed deer distribution. Ecol Model 190:171–189

Skupin A, Hagelman R (2005) Visualizing demographic trajectories with self-organizing maps. GeoInformatica 9:159–179

Smirnov O, Anselin L (2001) Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. Comput Stat Data Anal 35:301–319
Smith A, Roberts G (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J R Stat Soc B 55:3–23
Spooner PG, Lunt ID, Okabe A, Shiode S (2004a) Spatial analysis of roadside acacia populations on a road network using the network k-function. Landsc Ecol 19:491–499
Spooner PG, Lunt ID, Okabe A, Shiode S (2004b) Spatial analysis of roadside *Acacia* populations on a road network using the network k-function. Landsc Ecol 19:491–499
Stamp NE, Lucas JR (1990) Spatial patterns and dispersal distances of explosively dispersing plants in Florida sandhill vegetation. J Ecol 78:589–600
Stein M, Quashnock J, Loh J (2000) Estimating the K function of a point process with an application to cosmology. Ann Stat 28:1503–1532
Stoll P, Bergius E (2005) Pattern and process: competition causes regular spacing of individuals within plant populations. J Ecol 93:395–403
Stone R (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. Stat Med 7:649–660
Strauss D (1975) A model for clustering. Biometrika 62:467–475
Stringer P, Haslett J (1991) The spatial distribution of ill-health and material deprivation: an exploratory analysis using interactive graphics. Working paper, Northern Ireland Regional Research Laboratory, Dublin
Tiefelsdorf M (2000) Modelling spatial processes. Springer, Berlin
Tiefelsdorf M, Griffith DA, Boots BN (1999) A variance-stabilizing coding scheme for spatial link matrices. Environ Plan A 31:165–180
Tirado R, Pugnaire FI (2003) Shrub spatial aggregation and consequences for reproductive success. Oecologia 136:296–301
Tobler WR (1969) The spectrum of US 40. Pap Reg Sci Assoc 23:45–52
Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234–240
Tobler WR (1979a) Cellular geography. In: Gale S, Olsson G (eds) Philosophy in geography. D Reidel, Dordrecht, pp 519–536
Tobler WR (1979b) Smooth pycnophylactic interpolation for geographical regions. J Am Stat Assoc 74:519–530
Tobler WR (1983) An alternate formulation for spatial-interaction modeling. Environ Plan A 15:693–703
Tobler WR (1989) Frame independent spatial analysis. In: Goodchild MF, Gopal S (eds) The accuracy of spatial databases. Taylor & Francis, London, pp 115–122, http://www.loc.gov/catdir/enhancements/fy0744/90113396-d.html
Trevelyan B, Smallman-Raynor M, Cliff AD (2005) The spatial structure of epidemic emergence: geographical aspects of poliomyelitis in north-eastern USA, July–October 1916. J R Stat Soc A 168:701–722
Trpis M, Hausermann W (1986) Dispersal and other population parameters of *Aedes aegypti* in an African village and their possible significance in epidemiology of vector-borne diseases. Am J Trop Med Hyg 35:1263–1279
Tufte E (2001) The visual display of quantitative information. Graphics, Cheshire, CT
Tukey JW (1977) Exploratory data analysis. Addison-Wesley series in behavioral science. Addison-Wesley, Reading, MA
Tun-Lin W, Kay BH, Barnes A (1995) The premise condition index: a tool for streamlining surveys of *Aedes aegypti*. Am J Trop Med Hyg 53:591–594
Ungerer MJ, Goodchild MF (2002) Integrating spatial data analysis and gis: a new implementation using the component object model (com). Int J Geogr Inf Sci 16:41–53
US Census Bureau International Programs Center (n.d.) International data base (IDB). http://www.census.gov/ipc/www/idbnew.html
van der Pijl L (1972) Principles of dispersal in higher plants, 2nd edn. Springer, New York

Wadda N, Ribbens E (1997) Japanese maple (*Acer palmatum* var. Matsumurae, Aceraceae) recruitment patterns: seeds, seedlings, and sapling in relation to conspecific adult neighbors. Am J Bot 84:1294–1300

Walker PA, Moore DM (1988) SIMPLE: an inductive modelling and mapping tool for spatially-oriented data. Int J Geogr Inf Sci 2:347–363

Warren RE (1990) Predictive modelling of archaeological site location: a case study in the Midwest. In: Allen KMS, Green SW, Zubrow EBW (eds) Interpreting space: GIS and archaeology. Taylor & Francis, London, pp 201–215, http://www.loc.gov/catdir/enhancements/fy0744/90034471-d.html

Waterman SH, Novak RJ, Sather GE, Bailey RE, Rios I, Gubler DJ (1985) Dengue transmission in two Puerto Rican communities in 1982. Am J Trop Med Hyg 34:625–632

Watts D, Porter K, Putvatana P, Vasquez B, Calampa C, Hayes C, Halstead SB (1999) Failure of secondary infection with American genotype dengue 2 to cause dengue haemorrhagic fever. Lancet 354:1431–1434

Weeks JR (2004a) The role of spatial analysis in demographic research. In: Goodchild MF, Janelle DG (eds) Spatially integrated social science. Oxford University Press, New York, http://www.loc.gov/catdir/enhancements/fy0613/2002156669-d.html

Weeks JR (2004b) Using remote sensing and geographic information systems to identify the underlying properties of urban environments. In: Champion AG, Hugo G (eds) New forms of urbanization: beyond the urban–rural dichotomy. Ashgate, London

Weeks JR, Gadalla MS, Rashed T, Stanforth J, Hill AG (2000) Spatial variability in fertility in Menoufia, Egypt, assessed through the application of remote-sensing and GIS technologies. Environ Plan A 32:695–714

Weeks JR, Getis A, Hill AG, Gadalla MS, Rashed T (2004) The fertility transition in Egypt: intra-urban patterns in Cairo. Ann Assoc Am Geogr 94:74–93

Weeks JR, Larson D, Fugate D (2005) Patterns of urban land use as assessed by satellite imagery: an application to Cairo, Egypt. In: Entwisle B, Stern PC (eds) Population, land use, and environment: research directions. National Academies Press, Washington, DC, pp 265–286, http://www.loc.gov/catdir/toc/fy0707/2005016858html

Weiner J (1984) Neighbourhood in interference amongst *Pinus rigida* individuals. J Ecol 72: 183–195

Wells ML, Getis A (1999) The spatial characteristics of stand structure in *Pinus torreyana*. Plant Ecol 143:153–170

Whipple SA (1980) Population dispersion patterns of trees in a southern Louisiana hardwood forest. Bull Torrey Bot Club 107:71–76

Whittle P (1954) On stationary processes in the plane. Biometrika 41:434–449

Wiegand T, Moloney KA (2004) Rings, circles, and null-models for point pattern analysis in ecology. Oikos 104:209–229

Wiegand K, Jeltsch F, Ward D (2000) Do spatial effects play a role in the spatial distribution of desert-dwelling *Acacia raddiana*? J Veg Sci 11:473–484

Williams I, Limp WF, Briner FL (1990) Predictive modelling of archaeological site location: a case study in the Midwest. In: Allen KMS, Green SW, Zubrow EBW (eds) Interpreting space: GIS and archaeology. Taylor & Francis, London, pp 239–273, http://www.loc.gov/catdir/enhancements/fy0744/90034471-d.html

Wilson A (1967) A statistical theory of spatial interaction models. Transp Res 1:253–269

Wilson A (1970) Inter-regional commodity flows: entropy maximizing approaches. Geogr Anal 3:255–282

Wolf A (2005) Fifty year record of change in tree spatial patterns within a mixed deciduous forest. For Ecol Manage 215:212–223

Wulder M, Boots BN (1998) Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the Getis statistic. Int J Remote Sens 19:2223–2231

Yamada I, Thill J (2004) Comparison of planar and network K-functions in traffic accident analysis. J Transp Geogr 12:149–158

Yates F (1933) The analysis of replicated experiments when the field results are incomplete. Empir J Exp Agric 1:129–142
Youngblood A, Max T, Coe K (2004) Stand structure in eastside old-growth ponderosa pine forests of Oregon and northern California. For Ecol Manage 199:191–217
Zeiler M (1999) Modeling our world: the ESRI guide to geodatabase design. ESRI, Redlands, CA

Author Index

Abler, R. F. 36, 50, 255
Agterberg, F. 39, 256
Ahuja, N. 45, 255
Aldstadt, J. 6, 16, 178, 255, 261
Almeida-Neto, M. 118, 255
Anselin, L. 3, 4, 6, 11, 24, 35–38, 40–42, 45, 46, 49, 50, 57, 61, 62, 70, 72, 73, 113, 116, 120, 127, 128, 148, 168, 173–175, 179, 195, 241, 255, 261, 268, 269
Antos, J. A. 116, 118, 267
Arbia, G. 44, 127, 255
Arctur, D. 54, 255
Arriaga, E. E. 237, 255
Ashton, M. S. 114, 267
Ashton, P. S. 114, 116, 118, 119, 257, 267
Astete, H. 8, 265
Atalik, G. 197, 255
Atkins, D. E. 55, 256

Bailey, R. E. 204, 223, 270
Bailey, T. C. 36, 256
Bailly, A. 4, 256
Baker, P. 114, 116, 118, 119, 257
Barberis, I. M. 114, 256
Barnes, A. 222, 269
Bartels, C. P. A. 150, 256
Bartlett, M. S. 83, 94, 256
Becker, R. A. 41, 256
Bennett., R. J. 227, 263
Bergius, E. 114, 269
Berglund, S. 68, 76, 256
Berry, B. J. L. 41, 256
Besag, J. E. 84, 85, 115, 176, 256
Bivand, R. 39, 42, 225, 231, 256
Bliss, L. C. 116, 264
Bodziarczyk, J. 117, 118, 267
Bolduc, D. 65, 256
Bonham-Carter, G. 39, 256
Boots, B. N. 4, 45, 86, 104, 108, 109, 116, 119, 120, 149, 150, 166, 256, 261, 266, 267, 269, 270
Bradley, R. 41, 263
Bradshaw, G. A. 114, 116, 118, 257, 258
Brandsma, A. S. 65, 256
Briner, F. L. 39, 41, 270
Brunsdon, C. 6, 37, 39, 41, 57, 148, 260, 266
Brusegard, D. 36, 261
Bryan Jr, A. L. 101, 260
Bunyavejchewin, S. 114, 116, 118, 119, 257
Burrough, P. A. 36, 49, 256
Busing, R. T. 114, 256

Calampa, C. 205, 270
Call, L. J. 117, 118, 256
Camarero, J. J. 120, 257
Casterline, J. B. 234, 259
Chadee, D. D. 204, 221, 224, 259
Chambers, J. M. 41, 256
Chan, K. L. 223, 257
Chan, N. 138, 258
Chapeau-Blondeau, F. 185, 257
Chapilliquen, F. 8, 265
Charlton, M. E. 6, 37, 39, 41, 49, 51, 57, 148, 260, 266
Chen, C. 9, 257
Chen, D. 5, 19, 257
Chen, J. Q. 116, 118, 257
Chicago Area Transportation Study 90, 257
Childress, W. M. 114, 258
Chiu, S.-N. 104, 108, 109, 116, 119, 266
Christensen, N. L. 114, 267
Clark, D. A. 114, 257
Clark, D. B. 114, 257
Clark, G. G. 207, 223, 258, 268

Clark, J. S. 114, 257
Clark, S. C. 117, 118, 262
Cleveland, W. S. 41, 257
Cliff, A. D. 24, 25, 27, 30, 44, 65, 84, 128–130, 133, 166, 170, 174, 180, 195, 257, 269
Coe, K. 116–118, 271
Cohen, D. 117, 118, 264
Cole, R. G. 117, 118, 257
Collinet, F. 118, 259
Condit, R. 114, 116, 118, 119, 257
Connell, J. H. 114, 257
Consoli, G. B. 207, 257
Continental Illinois National Bank, Area Development Division 91, 257
Cooper, C. F. 114, 257
Corless, R. M. 185, 257
Costanzo, C. 24, 25, 29, 33, 44, 133, 263
Costero, A. 223, 258
Couclelis, H. 36, 37, 43, 257
Couteron, P. 118, 258
Cowen, D. J. 50, 258
Cox, D. R. 71, 258
Craft, A. 41, 51, 266
Craig, P. 41, 263
Cressie, N. 16, 44, 45, 61, 128, 138, 166, 229, 258
Crook, D. A. 116, 258
Crosier, S. J. 55, 258
Cross, A. 41, 49, 51, 266
Csillag, F. 36, 258

Dacey, M. F. 149, 258
Dale, M. R. T. 113–116, 119, 120, 258, 259
Daly, F. 229, 263
Daly, K. L. 114, 258
Dangermond, J. 50, 264
Daniels, E. 207, 221, 223, 224, 259
Davies, I. 118, 258
Davies, S. J. 114, 267
Davis, J. C. 45, 128, 258
de Graaff, T. 150, 169, 170, 259
de Oliveira, R. L. 207, 257
Dempster, A. 120, 225, 258, 267
Deng, M. 167, 178, 179, 258
Deutschman, D. H. 114, 258
Diaz, G. 8, 265
Diggle, P. J. xvi, 84, 86, 94, 113, 115, 119, 122, 181, 258, 268
Ding, Y. 37, 39, 41, 49, 258
Dixon, P. M. 101, 258, 260
Doepker, R. V. 117, 268
Donnegan, J. A. 117, 118, 120, 258
Donovan, M. 117, 268
dos Santos, F. A. M. 118, 259
Droegemeier, K. K. 55, 256
Druckenbrod, D. L. 118, 258
Drummy, P. 6, 261
Dufournaud, C. 150, 256
Duncan, R. P. 114, 263
Durbin, J. 65, 258

Edman, J. D. 207, 223, 258, 263, 268
El-Zanaty, F. H. 254, 258
Enright, N. J. 115, 119, 267
Entwisle, B. 234, 259

Fargues, P. 249, 259
Farley, J. A. 39, 41, 259
Feldman, S. I. 55, 256
Fellmann, J. D. 2, 259, 261
Ferguson, M. 4, 261
Fernandez Salas, I. 223, 266
Fischer, M. M. 2, 4, 51, 62, 63, 73, 74, 259
Florax, R. J. G. M. 149, 150, 168–170, 175, 255, 259
Flores-Leal, A. 223, 266
Flowerdew, R. 39, 259
Flury, B. 225, 259
Fochs, D. 8, 265
Focks, D. A. 204, 207, 221, 223, 224, 259
Fonseca, M. G. 118, 259
Forget, P.-M. 118, 259
Fortin, M. J. 113–116, 118–120, 257, 259, 267
Foster, R. B. 114, 116, 118, 119, 257
Foster, S. A. 44, 128, 260
Fotheringham, A. S. 6, 24, 37, 39, 41, 49, 57, 148, 258, 260, 263
Fowler, N. 114, 260
Franklin, J. 5, 13, 94, 96, 101, 103, 109, 112–119, 121, 131, 260, 261
Fugate, D. 240, 267, 270

Gadalla, M. S. 234, 235, 237, 239–242, 245, 251, 260, 267, 270
Gaines, K. F. 101, 260
Gale, N. 24, 263
Garcia-Molina, H. 55, 256
Gartin, J. 6, 261
Gatrell, A. 36, 37, 260
Geary, R. C. 24, 260
Getis, A. 1–8, 11–14, 16, 18, 19, 27, 31, 44, 45, 49, 50, 62, 68, 86, 87, 91, 94, 95, 100, 101, 103, 109, 112, 113, 115–121, 131, 148, 149, 151, 166, 172, 178, 191, 192, 194, 204,

208–210, 212, 223, 228, 231, 234, 240–242, 255–257, 259–262, 265, 266, 270
Getis, J. 2, 3, 259, 261
Getis, V. 2, 261
Gibson, L. J. 4, 256
Gilley, O. 178, 266
Gillingham, M. P. 116, 268
Githeko, A. K. 116, 265
Glass, L. 84, 261
Golledge, R. G. 24, 25, 29, 33, 44, 84, 133, 263, 268
Gonnet, G. H. 185, 257
Good, B. J. 114, 261
Goodchild, M. F. 36, 37, 42, 43, 49, 52, 54–56, 58, 258, 261, 264, 269
Gopal, S. 43, 58, 261
Goreaud, F. 115, 117, 119, 262, 267
Gorr, W. L. 6, 44, 128, 260, 261
Goulard, M. 118, 265
Gould, P. 37, 262
Granger, C. W. J. 171, 262
Gray, K. 7, 8, 18, 261, 265
Green, M. 39, 259
Grieg-Smith, P. 94, 262
Griffin, E. 7, 262
Griffin, P. 7, 262
Griffith, D. A. 7, 35, 42, 45, 61, 62, 87, 127, 149, 150, 175, 227, 228, 231, 241, 242, 255, 261–263, 269
Grunbaum, D. 114, 258
Gu, W. D. 116, 117, 262
Gubler, D. J. 204, 220, 221, 223, 224, 262, 268, 270
Guerra, M. A. 116, 262
Gujarati, D. 191, 262
Gunatilleke, N. 114, 116, 118, 119, 257
Gunatilleke, S. 114, 116, 118, 119, 257
Guo, L. D. 117, 264
Gutierrez, E. 120, 257

Haase, P. 117–119, 262
Hagelman, R. 51, 268
Haile, D. G. 207, 221, 223, 224, 259
Haining, R. P. 23, 27, 36, 37, 41, 42, 44, 45, 49, 227, 261–263
Hall, K. R. 117, 268
Halstead, S. B. 204, 205, 223, 263, 270
Hancock, J. 101, 267
Hand, D. 229, 263
Hanski, I. 116, 117, 262
Hare, D. E. G. 185, 257
Harries, K. 6, 261
Harrington, L. C. 223, 258, 263
Harrison, D. 178, 263
Haslett, J. 41, 263, 269
Hausermann, W. 223, 269
Hawkins, D. 183, 263
Hayes, C. 205, 270
Haynes, K. E. 24, 263
He, F. L. 114, 263
Head, T. H. C. 128, 258
Hepple, L. W. 71, 263
Hill, A. G. 234, 235, 237, 240–242, 245, 251, 267, 270
Hill, L. L. 55, 258
HilleRisLambers, J. 114, 257
HMSO 36, 263
Hoeffding, W. 130, 263
Howe, H. F. 114, 263
Hubbell, S. P. 114, 116, 118, 119, 257
Hubert, L. 24, 25, 29, 33, 44, 133, 263
Humphries, P. 116, 258

Ida, H. 114, 263
Incoll, L. D. 117, 118, 262
Itoh, A. 114, 116, 118, 119, 257

Jaffe, A. B. 73, 263
Janikas, M. V. 268
Jansenberger, E. 62, 73, 74, 259
Janzen, D. H. 114, 263
Jeffrey, D. J. 185, 257
Jeltsch, F. 118, 270
Johnson, N. L. 106, 110, 263

Kadmon, R. 117, 118, 264
Kaneene, J. 7, 8, 266
Karlström, A. 68, 76, 256
Kashian, D. M. 116, 118, 263
Kay, B. H. 222, 269
Keesling, J. E. 221, 259
Kehris, E. 39, 41, 259, 263
Kelejian, H. 172, 264
Kelletaper, R. H. 65, 256
Kemp, A. W. 106, 110, 263
Kenkel, N. C. 114, 116, 118, 264
Kern, R. 114, 257
Kim, Y.-W. 179, 255
King, A. J. 116, 258
Kitron, U. 116, 262
Kittayapong, P. 207, 223, 268
Klein, M. L. 55, 256
Knuth, D. E. 185, 257
Kokou, K. 118, 258

Konttinen, T. 116, 117, 262
Kooijman, S. 149, 150, 177, 264
Kotz, S. 106, 110, 263
Krishnan, T. 225, 265
Kuusinen, M. 116, 117, 262

Laessle, A. M. 114, 264
Laferriere, R. 65, 256
LaFrankie, J. V. 114, 116–119, 257, 264
Laird, N. 225, 258
Larsen, D. R. 116, 264
Larson, B. C. 114, 266
Larson, D. 240, 270
Laurent, E. J. 117, 268
Lawson, A. 181, 264
Lea, A. 4, 261
Leader, R. 7, 8, 266
LeBouton, J. 117, 268
Lee, H. S. 114, 116, 118, 119, 257
Leemans, R. 114, 264
LeSage, J. P. 65, 67, 148, 177, 264
Leung, Y. 51, 259
Lewinsohn, T. M. 118, 255
Lewis, R. 42, 262
Li, B. 42, 262
Liang, Y. 117, 264
Limp, W. F. 39, 41, 259, 270
Lindgren, B. S. 116, 268
Little, R. 226, 264
Liu, J. G. 117, 268
Lloyd, J. 7, 8, 266
Lockhart, J. 39, 41, 259
Loh, J. 101, 269
Longley, P. A. 52, 264
Lookingbill, T. R. 117, 118, 264
Lorenz, L. H. 207, 223, 268
Lorimer, C. G. 116, 118, 263
Losos, E. 114, 116, 118, 119, 257
Lucas, J. R. 117, 269
Lunn, A. 229, 263
Lunt, I. D. 101, 116, 118, 269
Lupi, F. 117, 268

Ma, K. P. 117, 264
MacDougall, E. B. 41, 264
Macklin, E. 114, 257
Maguire, D. J. 50, 52, 56, 264
Makinson, C. 252, 264
Maling, D. H. 52, 264
Malkinson, D. 117, 118, 264
Manokaran, N. 114, 116, 118, 119, 257
Mantel, N. 133, 264
Mapping Science Committee, N. R. C. 53, 264
Marble, D. F. 41, 256
Martin, R. L. 65, 257
Martini, M. Z. 118, 259
Mast, J. N. 114, 116–118, 264, 265
Matern, B. 86, 265
Max, T. 116–118, 271
McCay, D. H. 117, 118, 267
McConway, K. 229, 263
McDonald, J. 148, 177, 265
McDonald, P. T. 223, 265
McDonald, R. I. 114, 118, 265
McGilI, M. E. 41, 257
McKnight, S. 42, 262
McLachlan, G. 225, 265
McMillen, D. 148, 177, 265
Meng, X. 225, 265
Mennis, J. 239, 265
Mercier, F. 118, 259
Messerschmitt, D. G. 55, 256
Messina, P. 55, 256
Michael, B. 56, 264
Michaelsen, J. 94, 96, 116, 117, 119, 260
Miller, B. P. 115, 119, 267
Miller, C. 114, 265
Miller, H. 4, 261
Miller, H. D. 71, 258
Minakawa, N. 116, 265
Mitchell, A. 51, 265
Moellering, H. 94, 265
Moeur, M. 114, 116–118, 265
Moloney, K. A. 113, 119, 270
Monir, A. 185, 257
Montgomery, D. 226, 265
Moore, D. M. 39, 270
Moran, P. A. P. 24, 118, 265
Moravie, M.-A. 118, 265
Morrison, A. 7, 8, 18, 207, 261, 265, 268
Morrison, A. C. 8, 204, 223, 258, 265, 268
Morrison, J. 93, 268
Mosteller, F. 41, 265
Mount, G. A. 207, 223, 224, 259
Muehrcke, P. 93, 268
Mugglestone, M. A. 116, 265
Muller-Landau, H. C. 114, 265
Muncke, C. 101, 267
Munyekenye, O. G. 116, 265
Mur, J. 2, 261
Mushinzimana, E. 116, 265

Nathan, R. 114, 265
Navas, M. L. 118, 265
Nebert, D. 52, 265

Neyman, J. 119, 265
Ng, E. G. xiii, 77, 78, 265
Nicotra, A. B. 117, 265
Nijkamp, P. 36, 266
Nilsen, E. T. 117, 118, 256
Novak, R. J. 204, 223, 270
Nyerges, T. L. 51, 266

O'Brien, D. 7, 8, 266
Okabe, A. 101, 104, 108, 109, 116, 118, 119, 266, 269
Oliver, C. D. 114, 266
Oliver, M. A. 45, 266
Olwell, D. 183, 263
Openshaw, S. 36, 37, 39, 41, 43, 44, 49, 51, 150, 239, 266
Ord, J. K. 5, 7, 8, 16, 23–25, 27, 30, 31, 44, 65, 68, 72, 84, 120, 128–130, 133, 148, 151, 166, 170, 172, 175, 180, 191, 192, 195, 210, 241, 257, 261, 266
Ordonez, J. G. 223, 266
Ostriker, J. P. 55, 256
Ostrowski., E. 229, 263

Pacala, S. 100, 266
Pace, R. 65, 67, 178, 264, 266
PAHO 221, 222, 267
Palmiotto, P. A. 114, 267
Pancer-Koteja, E. 117, 118, 267
Parish, R. 116, 118, 267
Parker, A. J. 117, 118, 267
Parker, K. C. 117, 118, 267
Parton, R. 101, 267
Pascual, M. 114, 258
Peck, E. 226, 265
Peet, R. K. 114, 118, 265, 267
Pelissier, R. 115, 117–119, 262, 267
Peres-Neto, P. R. 7, 262
Perry, G. L. W. 115, 119, 267
Peuquet, D. 37, 42, 267
Peyton, B. W. xiii, 77, 78, 265
Pickles, J. 58, 267
Pielou, E. 45, 93, 113, 119, 267
Pindyck, R. 200, 267
Pinske, J. 169, 267
Pitman, E. J. G. 133, 267
Plane, D. 87, 267
Porter, K. 205, 270
Potvin, F. 120, 267
Powell, R. 240, 267
Powell, R. D. 119, 120, 258
Prior, I. 101, 267
Pugnaire, F. I. 117, 118, 262, 269
Putnam, J. 223, 267
Putvatana, P. 205, 270

Quashnock, J. 101, 269
Quastler, I. 2, 261

R Development Core Team xvi, 121, 122, 267
Racevskis, L. 117, 268
Ramirez-Prada, C. 8, 265
Rashad, H. 239, 267
Rashed, T. 234, 235, 237, 240–242, 245, 251, 267, 270
Rayner, J. N. 44, 84, 94, 267, 268
Rebertus, A. J. 117, 118, 120, 258
Reggiani, A. 63, 259
Reismann, M. 62, 259
Reiter, P. 8, 204, 220, 221, 223, 224, 265, 268
Renshaw, E. 116, 265
Rey, S. J. 114, 118, 149, 168, 175, 241, 255, 259, 260, 268
Rhind, D. W. 52, 264
Ribbens, E. 118, 120, 257, 270
Ridley, H. N. 114, 268
Rietveld, P. 36, 266
Rigau-Perez, J. G. 8, 204, 223, 265, 268
Rios, I. 204, 223, 270
Rip, M. 7, 8, 266
Ripley, B. D. 4, 45, 84, 85, 94, 96, 101, 113, 115, 119, 131, 208, 209, 268
Roberts, A. 118, 265
Roberts, D. 240, 267
Roberts, G. 176, 269
Robertson, A. I. 116, 258
Robinson, A. 93, 268
Robinson, D. 172, 264
Rodhain, F. 222, 268
Rodriguez-Figueroa, L. 223, 268
Rogan, J. 240, 267
Rogerson, P. 6, 49, 260, 261
Romme, W. H. 116, 118, 263
Rosen, L. 222, 268
Rowlingson, B. S. xvi, 122, 181, 258, 268
Rubin, D. 225, 226, 258, 264
Rubinfeld, D. 178, 200, 263, 267

Sale, R. 93, 268
Santaló, L. A. 110, 111, 268
Santarossa, G. 65, 256
Santiago, M. 8, 204, 223, 265
Sather, G. E. 204, 223, 270

Saw, L. G. 117, 118, 264
Sayed, H. A. A. 234, 259
Scanlon, J. E. 204, 223, 263
Schachter, B. J. 45, 255
Schafer, J. 226, 268
Scherngell, T. 62, 73, 74, 259
Schooley, R. L. 116, 117, 268
Schroff, A. Z. 116, 268
Scott, L. 241, 268
Scott, T. W. 7, 8, 18, 207, 223, 258, 261, 263, 265, 267, 268
Sen, A. K. 62, 64, 268
Shi, H. 117, 120, 268
Shihuincha, M. 8, 265
Shiode, S. 101, 116, 118, 269
Shugart, H. H. 118, 258
Shumaker, N. H. 114, 258
Silander, J., Jr 100, 266
Silman, M. 114, 257
Skupin, A. 51, 268
Smallman-Raynor, M. 174, 269
Smirnov, O. 72, 73, 269
Smith, A. 176, 269
Smith, T. E. 62, 64, 268
Smith, T. R. 55, 258
Spooner, P. G. 101, 116, 118, 269
Stamp, N. E. 117, 269
Stanforth, J. 234, 235, 237, 245, 251, 270
Stein, M. 101, 269
Stoe, D. 6, 261
Stoll, P. 114, 269
Stone, R. 181, 269
Stow, D. 240, 267
Strahler, A. H. 94, 96, 116, 117, 119, 260
Strauss, D. 84, 86, 87, 269
Strickman, D. 207, 223, 268
Stringer, P. 41, 269
Suarez, E. 223, 268
Sugihara, K. 104, 108, 109, 116, 119, 266
Sukumar, R. 114, 116, 118, 119, 257
Syms, C. 117, 118, 257
Syrabi, I. 179, 255
Szwagrzyk, J. 117, 118, 267

Tanner, E. V. J. 114, 256
Taylor, P. 44, 266
Thill, J. 101, 270
Tiefelsdorf, M. 149, 150, 256, 269
Tirado, R. 117, 118, 269
Tobler, W. R. 24, 40, 44, 57, 84, 94, 167, 261, 265, 269
Tracey, J. G. 114, 257
Trajtenberg, M. 73, 263
Trevelyan, B. 174, 269
Trpis, M. 223, 269
Tufte, E. 168, 269
Tukey, J. W. 41, 45, 265, 269
Tun-Lin, W. 222, 269
Turner, M. G. 116, 118, 263

Udomsakdi, S. 204, 223, 263
Umpaivit, P. 204, 223, 263
Ungerer, M. J. 54, 269
Unwin, A. 41, 263
Urban, D. L. 114, 118, 265
U.S. Census Bureau International Programs Center 252, 269

van der Pijl, L. 114, 269
Vasiliev, I. 42, 262
Vasquez, B. 205, 270
Veblen, T. T. 116, 118, 264
Vogt, D. J. 114, 267
Vogt, K. A. 114, 267

Wadda, N. 118, 270
Walker, E. D. 116, 262
Walker, P. A. 39, 270
Walters, M. B. 117, 268
Ward, D. 118, 270
Warren, R. E. 39, 270
Waterman, S. H. 204, 223, 270
Watts, D. 8, 205, 265, 270
Way, A. A. 254, 258
Webb, L. J. 114, 257
Webster, R. 45, 266
Weeks, J. R. 234, 235, 237, 240–242, 245, 251, 267, 270
Weiner, J. 100, 270
Wells, M. L. 117, 118, 120, 121, 270
Whipple, S. A. 114, 261, 270
Whittle, P. 24, 270
Wiegand, K. 118, 270
Wiegand, T. 113, 119, 270
Wiens, J. A. 116, 117, 268
Wilks, A. R. 41, 256
Williams, I. 39, 41, 270
Wills, G. 41, 263
Wilson, A. 24, 62, 270
Wise, S. 36, 37, 49, 261
Wolf, A. 117, 118, 270
Wolf, J. J. 114, 116–118, 265
Wright, D. 39, 256
Wright, M. H. 55, 256

Wright, R. 6, 261
Wu, J. 114, 258
Wulder, M. 120, 270
Wymer, C. 41, 51, 266

Yakamura, T. 114, 116, 118, 119, 257
Yamada, I. 101, 266, 270
Yan, G. Y. 116, 265
Yang, X. 42, 262
Yates, F. 226, 271
Youngblood, A. 116–118, 271

Zavala, M. A. 117, 118, 264
Zeiler, M. 54, 255, 271
Zhang, L. J. 120, 268
Zhou, G. F. 116, 265
Zhou, H. 207, 223, 268
Zoller, H. G. 2, 261
Zoppè, A. 225, 259

Index

A
aggregation, 37, 44, 113, 114, 118, 171, 239
algorithm, 44–46, 62, 77, 78, 225
ANOVA, 45
asymmetry, 18, 168
asymptotic, 6, 26, 69, 169, 170
autocorrelated, 28, 65, 70, 149, 158, 191, 192, 195–198, 200, 201, 241
autocorrelation, 3–20, 27, 66, 122, 127, 153, 156–158, 160, 162, 163, 191
autoregressive parameter, 71, 77, 78

B
bias, 74, 84, 85, 89, 144, 192, 206
binary, 67, 68, 129, 150, 166, 169, 228
bivariate, 41, 118, 120
bootstrap, 46
border, 67, 68, 74, 75, 78, 83, 85, 89, 99, 100, 148, 208
boundary, 89, 94, 95, 115, 167, 172, 235, 244

C
calibration, 29
case control, 18, 183, 187
Chicago, 5, 13, 14, 85, 87–92
Cholesky decomposition, 73
clustering, 5, 14, 17, 19, 41, 45, 84–93, 96, 98, 100, 101, 114, 120, 140, 147, 148, 151, 156, 158, 193, 194, 203, 204, 208–213, 215, 216, 219, 220, 223, 235, 240, 241, 253
clusters, 5, 6, 8, 13, 15, 16, 19, 20, 86, 87, 91, 96, 98–100, 114, 118–120, 123, 131, 139, 152–154, 158, 162, 163, 174, 203–205, 210–213, 215–220, 222–224, 240, 243, 246, 249, 253, 287
complete spatial randomness, 209
confirmatory analysis, 37, 38, 40, 42, 43, 51
connectedness, 169
constant variance, 64
contiguity, 17, 29, 121, 122, 147–150, 154, 155, 162
  first order, 67
  queen criterion, 17, 132, 133, 149, 159–162, 169–171
  rook criterion, 17, 131, 135, 162
continuous, 74, 94, 104, 108, 109, 120
convergence, 10, 188
correlation, 23, 44, 74, 132, 134, 156, 158, 204, 205, 218, 219
  coefficient, 26
covariance, 4, 26, 28, 44, 70, 134, 136, 163, 171, 227
crime, 6, 7, 14, 18, 19, 102, 104, 109, 112, 178, 179, 195, 196
cross-product statistic, 3, 5, 11, 12, 16, 23, 25–30, 32, 33, 44, 133
cross-section, 46, 61, 66
cumulative sum statistic, 18, 183

D
dasymetric mapping, 241
data
  collection, 179, 221
  modeling, 54
  set, 4, 16, 19, 44, 53, 152, 153, 156, 160, 161, 165, 176, 178, 200, 225, 230, 235
Dengue fever, 7, 8, 18, 19, 203–205, 221–224
density function, 103, 119, 123
dependence, 7–18, 27, 40, 44, 45, 68, 70, 71, 77, 78, 127, 128, 136, 148, 151, 152, 166, 170, 178, 179
diagonal, 66, 67, 71, 73, 89, 91, 95, 169, 170, 180
distance
  critical, 17, 147, 148, 151, 153
  decay, 148, 162

distributions, 113, 184
inverse, 147, 148, 155, 156, 158
nearest neighbor, 83, 95, 96, 121, 148, 151
statistics, 5, 16
distribution, 5, 8, 14–16, 18, 19, 29, 63, 83–88, 91, 94–97, 104–112, 114, 119, 121, 130–132, 144, 148, 153, 156, 162, 166, 168, 169, 172, 177, 184, 203, 204, 208, 220, 223, 238, 239, 243, 245, 246, 286
dummy variable, 17, 74, 151, 152, 163, 178
dynamic, 19, 41, 42, 204, 224, 286

E
econometric, 40, 61, 62, 70, 73, 77, 79, 128, 150, 168, 200
models, 77
economic geography, 3, 6, 285, 287
economics, 3, 17, 47, 79, 165, 285
edge, 2, 10, 15, 46, 88, 89, 102–104, 117, 119, 123, 150, 169
efficiency, 17, 170, 176
eigenvalue, 18, 71, 72, 175
eigenvector, 7, 20, 225, 229–231
EM algorithm, 225–227, 230, 231
endogenous, 15, 114, 250
epidemiology, 6–8, 16, 224, 287
equifinality, 15, 114
equilibrium, 62, 224
error
term, 28, 63, 64, 70, 192
ESDA, 41, 42, 45, 46
estimator
maximum likelihood, 72, 77–79, 137, 177
ordinary least squares, 64, 65, 70, 77, 192, 194, 195, 197, 199–201, 231, 245, 248
SUR, 3, 54, 235, 237, 238, 243, 252, 253
experimental, 153
exploratory spatial data analysis, 123
exponential, 157, 159–161, 182, 230

F
fertility, 19, 20, 233–240, 242–254
filter, 7, 19, 20, 149, 191–194, 196, 197, 200, 201, 225, 229–231, 241, 242, 245, 246, 248
filtering, 7, 20, 44, 128, 194, 196, 225, 241
filtering procedure, 19, 191–193, 235
forecasting, 287
FORTRAN, 241

G
GAUSS, 148, 162, 230
geo-referenced data, 20, 152, 156, 228
geocoding, 53, 230
GeoDa, 57, 59, 285
geographer, 1, 3–5, 13, 41, 128, 286
geographic information systems, 3, 4, 12, 13, 35–47, 49–59, 68, 179, 207, 285–287
geographically weighted regression, 6, 57
geography, 1–4, 6, 17, 50, 58, 73, 84, 116, 165–167, 285–288
geospatial, 20, 53, 54, 56–59, 233, 285, 286, 288
geostatistical, 6, 27, 57, 157, 158, 163, 228
Getis–Ord statistic, 68, 78, 191
GLM, 231
goodness-of-fit, 157
GRASS, 39, 52, 56, 58
gravity models, 24, 33

H
heterogeneity, 5–20, 40, 64, 76, 88, 93, 96, 100, 114
homogeneity, 40, 42, 84
homogeneous, 15, 88, 119, 123, 150, 233
hot-spots, 172
hypergeometric, 31

I
identification, 5, 15, 45, 49, 165, 174, 222, 240
inference, 5, 16, 40, 42, 43, 46, 70, 92, 123, 168, 177
inhibition, 4, 14, 83, 85–91, 96, 114, 119, 123
initial model, 179
instrumental variables, 2, 46, 127
intercept, 70, 226
isotropy, 25, 84

J
journey to work, 87, 90, 91

K
K function, 120, 208, 211, 215, 221, 223
global Voronoi cross, 106
local cross, 102
local Voronoi cross, 106
Ripley's, 113, 116–118
kernel, 120

L
land use, 3, 44, 87
large sample, 132, 149
lattice, 17, 27, 83, 131, 132, 134, 175, 230
linkages, 3, 169
LISA, 57, 173, 174
local
  estimation, 175, 178
  Moran coefficient, 18, 120
  spatial autocorrelation, 5, 16, 158
  statistic, 5–8, 10, 11, 14–16, 19, 112, 113, 116, 120, 147, 148, 151, 163, 172, 208, 210
  variation, 175
location, 3–6, 11–14, 16–19, 24, 29, 30, 37, 42, 43, 53, 61, 62, 84, 93, 96–98, 101, 104, 109, 113, 116, 120, 121, 130, 131, 133, 140, 156, 168, 181–184, 196, 204, 206, 209, 210, 216, 219, 220, 222–225, 230, 234–236, 239, 249, 285
log-additive, 13, 61, 62, 70, 75, 77
log-likelihood, 176, 187
log-likelihood function, 18, 71, 72, 175
logit, 46

M
Markov
  model, 150
  Chain Monte Carlo (MCMC), 176, 179
Matlab, 54
matrix, 6–17
meta-analysis, 169
missing data, 20, 225–228, 230
missing values, 226, 228–230
misspecification, 17, 42, 149, 192
model specification, 71, 77, 165, 167, 179, 228
modifiable areal unit problem, 42, 44, 51, 239
moments, 5, 6, 16, 25, 44, 45, 180
Monte Carlo
  simulation, 110, 115
Moran coefficient, 44, 118, 122, 147, 158, 229
multiple comparisons, 6
multiregional models, 39
multivariate
  normality, 227

N
neighbors, 13, 96, 120, 130–132, 140, 144, 148, 149, 152, 154–156, 162, 173, 175, 212, 213, 223
network, 9–11, 52, 53, 62, 101, 244, 287
non-stationary, 71
normal distribution, 26, 69, 104, 106, 107, 109, 134, 152, 174
normality, 6, 31, 130, 226

O
one-dimensional, 72
origin-destination, 13, 61, 62, 65–68, 70, 73, 76–79
overspecification, 149

P
parameter
  spatial, 151, 176
parametric, 45, 112, 176
planar, 149
plant ecology, 13, 15, 113, 115–119, 121
point pattern, 3–5, 13–15, 57, 83, 84, 88, 101, 113, 115–121, 123, 208, 239
Poisson, 14, 83, 85–87, 89–91, 93–97, 99, 103, 115, 119, 209
population
  change, 287
  density, 5, 14, 87, 91, 92, 204, 219–221
  dynamics, 100
  growth, 240
positive definite, 71
production
  attraction, 63
  constrained, 24, 26, 30
pycnophylactic, 44
PySAL, 287
Python, 54

Q
quadrat, 3, 83, 93, 94, 100, 178

R
random
  expectation, 193, 211
  location, 14, 171
  set, 14
randomization, 16, 29
REGIO, 197, 224, 287
regional science, 1–4, 6, 12, 17, 35–37, 46, 47, 49–51, 166, 285, 287
regression, 7, 15, 17, 19, 20, 24, 39, 42, 46, 61, 62, 65, 66, 70, 79, 119, 120, 128, 137, 147, 157, 167, 172, 173, 178, 179, 191–194, 196, 199, 201, 202, 226, 228, 229, 240–242, 244, 246, 250, 251, 253
regression coefficients, 226, 229
remote sensing, 20, 94, 128, 240, 286

residual, 19, 45, 61, 62, 65, 68, 69, 76, 77, 79, 120, 147, 157, 158, 161, 162, 167, 168, 172, 179, 194–201, 226, 242, 244–248, 250
row-standardization, 6, 149, 150, 153, 228

S
sample size, 173, 204, 221, 222, 252
SANET, 287
second order, 12, 14, 94, 115, 116
second order statistic, 3, 4, 116, 191
semi-parametric, 179
semi-variance, 17, 128, 147, 148
semivariogram, 156, 228–231
simulated data, 149
simulation, 17, 18, 38, 45, 54–56, 96, 147, 148, 152, 157, 158, 163, 169, 178, 210, 221
singular, 165
singularity, 152
software, 5, 6, 12, 13, 19, 39, 42, 49–56, 58, 59, 207, 285, 286
space–time, 8, 46, 178, 287
SpaceStat, 6, 42, 285
sparseness, 150, 169
spatial
  association, 5, 16, 23, 39, 41, 61, 62, 79, 113, 120–122, 128, 131, 136, 138, 140, 144, 148, 149, 151, 163, 173, 194–196, 200, 202
  autocorrelation, 3, 5, 11, 13, 16, 19, 23–25, 27–30, 32, 33, 44, 61, 62, 64–66, 69, 77, 79, 109, 127, 128, 136, 138, 140, 142, 144, 149, 150, 165–167, 173, 191, 194, 195, 197, 198, 201, 202, 228, 235, 240–244, 246, 248, 285
  autoregressive, 28, 65, 147
  correlation, 7, 23, 127, 151, 224
  data, 2–4, 6, 12, 16, 17, 24, 28, 35–45, 51, 53, 57, 93, 166, 168, 172, 216, 285, 286
  dependence, 17, 28, 35, 40–42, 44, 46, 47, 51, 57, 59, 61, 62, 67, 71, 77–79, 118, 148, 150, 163, 165, 167–169, 172, 173, 192–196, 199, 201, 202, 241, 242
  econometrics, 2, 46, 168, 287
  effects, 7, 19, 40, 46, 149, 150, 194, 196, 200, 241
  error, 70, 71, 77
  filtering, 6, 18–20, 163, 194, 225, 231, 233, 242, 248
  heterogeneity, 6, 12, 15, 35, 40–42, 46, 57, 59, 114, 150
  interaction, 3, 5, 11–13, 16, 23, 24, 26, 29, 30, 32, 33, 37, 61–66, 70, 71, 75, 77–79, 120, 165, 169, 239, 253
  lag, 18, 27, 152, 157, 167, 178
  models, 24, 26, 149, 178
  process, 17, 35, 38, 42, 46, 47, 70, 96, 150, 171, 179, 241
  scale, 19, 20, 35, 46, 93, 127, 234, 243
  structure, 17, 47, 147, 148, 150, 152, 157, 162, 204, 221, 222, 285
  unit, 40, 42, 44, 66, 67, 137, 149, 150, 166, 167, 221
  weights matrix, 6, 11, 13, 16, 17, 25–28, 30, 61, 66–68, 70–73, 75, 78, 79, 128, 147, 148, 157, 191, 192, 210
spectral, 24, 28, 84, 94, 116
spherical, 160, 161, 192, 230
spillover, 18, 74
standard deviation, 91, 152, 162, 210
STARS, 287
stationarity, 25, 68, 70, 71, 84, 88, 116, 119, 123, 156, 168, 171
stationary, 17, 28, 150, 156, 168, 208
surface, 37, 44, 45, 57, 68, 84, 114, 119, 149
surveillance, 16, 18, 19, 204, 205, 221, 222, 224

T
temporal, 8, 18, 181, 205, 206, 221, 223
tesellation, 42, 45
time series, 168, 169, 287
transformation, 3, 50, 59, 74, 88, 115, 130, 137, 191
trend surface, 44–46
Type I error, 173

U
univariate, 15, 39, 41, 106, 110, 112, 152
urban planning, 285, 287

V
variance, 12, 16, 27, 44, 69, 70, 78, 85, 87, 94, 100, 104–112, 115, 130, 172, 177, 187, 202, 209
variance-covariance matrix, 65, 71
variogram, 24, 27, 44, 156, 157
virus, 203, 204, 222–224
Voronoi, 15, 68, 107–109, 111, 112, 119, 285, 287

W
weight, 13–17

Z
zones, 171

Contributors

Luc Anselin is Director and Foundation Professor, School of Geographical Sciences and Urban Planning, and Director, GeoDa Center for Geospatial Analysis and Computation at Arizona State University. He was elected Fellow of the Regional Science Association International in 2004 and was awarded their Walter Isard Prize in 2005 and William Alonso Memorial Prize in 2006. He was elected to the National Academy of Sciences in 2008. His research interests include the analysis of spatial data (i.e., data containing a specific location) ranging from exploration to visualization and modeling, and the development of appropriate methods, their implementation in software and application in empirical studies. Anselin is the developer of the SpaceStat and GeoDa software packages for spatial data analysis. Specific application fields include environmental and natural resource economics, real estate economics, economics of innovation, criminology, public health, electoral studies and international relations.

Barry Boots retired from Wilfrid Laurier University in 2007 and is currently an Adjunct Professor of Geography at the University of Victoria. His interests are in modeling spatial structure and spatial autocorrelation, local spatial statistics, and Voronoi diagrams. He was an undergraduate student of Art Getis at the University of Bristol in 1966–1967 and Art was his MA and PhD advisor at Rutgers University 1968–1972. He also holds a DSc from Bristol and is the co-author of two books with Art.

Manfred M. Fischer is Professor of Economic Geography and Director of Institute for Economic Geography and GIScience at Vienna University of Economics and Business (WU-Vienna). His research spans a broad array of fields including regional and urban economics, housing and labour market research, transportation systems analysis, innovation economics, spatial behaviour and decision processes, spatial analysis and spatial statistics, and GIS. He is one of the leading scholars in the field of GeoComputation. Based on the expertise and the scientific impact Dr. Fischer has gained a high reputation both nationally and internationally. In 1995 he was elected as a member of the International Eurasian Academy of Sciences, in 1996 as a corresponding member of the Austrian Academy of Sciences and in 1999 as a foreign member of the Royal Netherlands Academy of Arts and Sciences. Dr. Fischer

has published over 250 scientific publications, including about 30 monographs and edited books.

Janet Franklin received a Bachelors degree on Environmental Biology (1979), Master of Arts (1983), and Ph.D. (1988) in Geography, all from the University of California at Santa Barbara. She has been the Editor of The Professional Geographer (1997–2000), Associate Editor of Journal of Vegetation Science (1999–2006), and Board Member of Ecology, Diversity and Distributions,and Landscape Ecology. She was a professor of Geography at San Diego State University, from 1988–2002, and then joined the faculty of Biology at SDSU (2002–2009). where she was Associate Chair (2006–2009). She joins the School of Geographical Sciences, Arizona State University, Fall 2009. She has published more than 80 refereed book chapters and articles. Research emphases include landscape ecology, plant community ecology, biogeography, biophysical remote sensing and geographic information science. She is interested in the dynamics and spatio-temporal patterns of plant communities.

Michael F. Goodchild is Professor of Geography at the University of California, Santa Barbara, and Director of spatial@ucsb. He received his BA degree from Cambridge University in Physics in 1965 and his PhD in Geography from McMaster University in 1969. After 19 years at the University of Western Ontario, he moved to Santa Barbara in 1988. He was Director of the National Center for Geographic Information and Analysis from 1991 to 1997. He was elected member of the National Academy of Sciences and Foreign Fellow of the Royal Society of Canada in 2002, and member of the American Academy of Arts and Sciences in 2006. He was Editor of Geographical Analysis between 1987 and 1990 and Editor of the Methods, Models, and Geographic Information Sciences section of the Annals of the Association of American Geographers from 2000 to 2006. He serves on the editorial boards of ten other journals and book series. His published books include Accuracy of Spatial Databases; Geographical Information Systems: Principles and Applications; Environmental Modeling with GIS; Scale in Remote Sensing and GIS; Interoperating Geographic Information Systems; Geographic Information Systems and Science; Uncertainty in Geographical Information; Foundations of Geographic Information Science; Spatially Integrated Social Science; GIS, Spatial Analysis, and Modeling; and Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools. In addition he is author of some 350 scientific papers. He was Chair of the National Research Council's Mapping Science Committee from 1997 to 1999, and currently chairs the Advisory Committee on Social, Behavioral, and Economic Sciences of the National Science Foundation. His current research interests center on geographic information science, spatial analysis, and uncertainty in geographic data.

Daniel A. Griffith is an Ashbel Smith professor, University of Texas/Dallas, and earned degrees in mathematics, statistics and geography. He holds awards from the Fulbright and John Simon Guggenheim Memorial Foundations, American Statistical Association, Leverhulme Trust, and Pennsylvania Geographical Society, is an

elected Fellow of the New York Academy of Sciences, the Spatial Econometrics Association, and Fitzwilliam College (University of Cambridge), a past president of the North American Regional Science Council and the Syracuse Chapter of Sigma Xi, and received a Doctor of Science, *honoris causa*, degree from Indiana University of Pennsylvania. He has published 15 books/monographs and more than 200 papers, given 200+ invited talks. His research includes: spatial statistics, urban public health, economic geography, and Puerto Rican agriculture.

Atsuyuki Okabe received his Ph.D. from the University of Pennsylvania in 1975 and the degree of Doctor of Engineering from the University of Tokyo in 1977. Previously he has held the position of Associate Professor at the Institute of Socio-Economic Planning, University of Tsukuba. He is currently Professor of the University of Tokyo, and a member of the Science Council of Japan. He was Director of the Center for Spatial Information Science from 1998–2005. He is specialized in spatial analysis and geographic information science. Recently, he with his colleagues is developing spatial analysis on networks and its toolbox, SANET. One of his co-authored books is: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams (2nd edition; John Wiley 2000).

Keith Ord is a professor in the Operations and Information Management group at the McDonough School of Business at Georgetown University. His research interests include time series forecasting, spatial modeling and the statistical modeling of business processes. He is a Fellow of the American Statistical Association and of the International Institute of Forecasters.

Sergio J. Rey is Professor of Geography and Urban Planning at Arizona State University. His research interests include open source spatial analysis, spatial econometrics, exploratory space–time data analysis and regional science. He is the creator of the package STARS: Space–Time Analysis of Regional Systems and lead developer of PySAL: A Python Library for Spatial Analysis. He is a Fellow of the Spatial Econometrics Association and Editor of the International Regional Science Review.

Peter A. Rogerson is Professor of Geography and Biostatistics at the University at Buffalo. His research interests are in the area of demography and population change, epidemiology, spatial statistics, and spatial analysis. He has authored or co-authored four books, the most recent being Statistical Detection and Monitoring of Geographic Clusters (2008; coauthored with Ikuho Yamada). He is co-editor (with Stewart Fotheringham) of The SAGE Handbook of Spatial Analysis (2009).

Toshiaki Satoh is currently a researcher in Research & Development Center of PASCO Corporation, a surveying and GIS consulting company. He received a Bachelor's degree from Tohoku University in 1992, a Master's degree from the Tokyo Institute of Technology in 1994 and Ph.D. from the University of Tokyo in 2007, respectively. His main interests of research are network spatial analysis and computer visualization.

John R. Weeks is Professor of Geography and Director of the International Population Center at San Diego State University, and Clinical Professor of Family and Preventive Medicine at the University of California, San Diego, School of Medicine. His research focuses on the application of geospatial techniques to demographic research and he is currently the project director and principal investigator of an NIH-funded study of the spatial inequalities in health in Accra, Ghana.