

2001

Analysis of aggregated spatial social data

Gandhi Pawitan

University of Wollongong

Recommended Citation

Pawitan, Gandhi, Analysis of aggregated spatial social data, Doctor of Philosophy thesis, School of Mathematics and Applied Statistics, University of Wollongong, 2001. <http://ro.uow.edu.au/theses/2041>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

ANALYSIS OF AGGREGATED SPATIAL SOCIAL DATA

A thesis submitted in fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

from

THE UNIVERSITY OF WOLLONGONG

by

Gandhi Pawitan, BSc., MSc.

SCHOOL OF MATHEMATICS AND APPLIED STATISTICS

2001

Declaration

In accordance with the regulation of the University of Wollongong, I hereby state that the work described here is my own original work, except where due references are made, and has not been submitted for a degree in any university or institution.

Gandhi Pawitan

Acknowledgements

There was a long path for this obsession to become a reality. I realize that many people have given a support before, during, and after my study at University of Wollongong.

Professor David G. Steel is the man behind this research. Your comments and suggestions are enormous in terms of quantity and quality, and have deepened the philosophy of my research and understanding. No other words except my sincere thanks for his patience and availability throughout my study. To Dr. Pamela Davy, I extend great thanks and appreciation for comments and suggestion on the final thesis.

Ulbert Silalahi, MA., Dean of the Faculty of Social and Political Science, *Universitas Katholik Parahyan-gan* for his support of my study here. Also, for the Australian Government, who through the AusAid scheme, have financially supported my study here.

Eric Beh, Riccardo Biondini, Craig McLaren, and Sifa Mvoi for being such good friends. Kerry Gamble, Carolyn Sylveri, and all staff of the School of Mathematics and Applied Statistics for their kindness and help during my study. Ava Davies of Campus East for being helpful during my stay at Kurumul Bldg. of Graduate House. All Indonesian individuals or families, who cannot be listed here, for their support. The "fishing gang", thanks for the extra curricular activity at Port Kembla, Coniston, North Beach, or Windang.

Thanks for all big Pawitan's and Hassan's families for their moral support. At last, the days are knitted with the colorful emotion with my wife "Honey" and the "bees" Arif and Sadikin. This is the most wonderful time of our life. *Pengertian dan Dukungannya tidak terukur . . .* Their understanding and supports are un-measurable (no statistics can be used to explain this) . . .

Abstract

There is an increasing tendency to take a spatial perspective in analysing census or sample data. This thesis contributes to the development of spatial analysis and concentrates on methods for analysing data on social characteristics. The important case of aggregated census data will be considered.

If we are interested in spatial relationships, then we must consider how to analyse social data that have been obtained by methods of sampling or aggregation. There may not be a direct interest in spatial relationships, but the presence of spatial interdependence may still need to be taken into account in the analysis. There may be spatial trends in means and variances, and the correlation between the characteristics of different individuals that depend on their relative locations.

The main outputs of the thesis have contributed in the development of the analysis of aggregate social data from a spatial perspective, in particular using semivariogram analysis. Some outputs are outlined here. The role of the semivariogram and cross-semivariogram of the aggregate data is to explore and explain the covariance structure and spatial dependency in the population. This includes the relationship between spatial autocorrelation and the variogram. The connection between the variogram of the aggregated data and the variogram of the unit level data, which leads to the development of a non-linear model of the group level semivariogram to provide estimates of the individual level semivariogram model parameters. The extension into the bivariate case involving cross-semivariogram analysis is discussed. The MAUP as an analysis tool to explore spatial dependence of aggregate social data is discussed. Simulation results and empirical analysis of actual aggregate data are implemented to confirm the methods. The empirical work is based on analysis of the 1991 Australian Census of Population and Housing.

The thesis shows the role of semivariogram analysis (univariate case) and cross-semivariogram analysis (bivariate case) in understanding the aggregation effect for social data. The aggregation effect, which includes the two main aspects of the scaling effect and zoning effect, is mainly determined by the presence of dependency within the data. Another factor is existence of the relationship between group size and within group variation.

Contents

| | |
|--|-----------|
| Declaration | ii |
| Acknowledgements | iii |
| Abstract | iv |
| Contents | ix |
| List of Tables | xi |
| List of Figures | xiv |
| Dedication | xv |
| | |
| 1 Introduction | 1 |
| 1.1 Background : data, targets of inference, and analysis | 1 |
| 1.2 Problems and motivation | 2 |
| 1.2.1 The problems | 2 |
| 1.2.2 Motivation : what and why spatial analysis | 3 |
| 1.3 Objectives | 5 |
| 1.4 Scope | 5 |
| | |
| 2 Definition and Problem Identification | 7 |
| 2.1 Definition of population | 7 |
| 2.1.1 Superpopulation approach | 7 |
| 2.1.2 Spatial series of social data | 8 |
| 2.1.3 Structure and sources of spatial data | 9 |
| 2.2 Problem identification | 9 |
| 2.2.1 Aggregation problems | 9 |
| 2.2.2 Sampling design problems | 12 |
| 2.3 Some spatial assumptions and implementations | 13 |
| | |
| 3 Literature Review | 16 |
| 3.1 Background | 16 |
| 3.2 Aggregation | 17 |
| 3.2.1 Group formation : zoning and scaling | 17 |
| 3.2.2 Aggregation of census data | 18 |
| 3.3 Non-spatial solutions of aggregation problems | 19 |
| 3.3.1 The MAUP solutions | 19 |
| 3.3.2 The ecological fallacy solutions | 20 |
| 3.4 Spatial analysis | 20 |
| 3.4.1 Target of inference : spatial variability | 21 |
| 3.4.2 Spatial analysis : analysis of spatial variability | 22 |
| 3.4.3 Themes of spatial analysis | 23 |
| 3.4.4 Application of spatial analysis and GIS | 25 |
| 3.5 Spatial perspective in analysing aggregated data | 27 |
| 3.5.1 The MAUP and ecological fallacy from a spatial perspective | 27 |

| | | |
|----------|--|------------|
| 3.5.2 | Spatial interpolation | 30 |
| 3.5.3 | Boundary problem issues | 30 |
| 3.5.4 | Variogram | 31 |
| 3.6 | Simulation | 32 |
| 3.7 | Summary | 32 |
| 4 | Some Theory of Spatial Aggregation | 34 |
| 4.1 | Aggregation and simple statistics | 35 |
| 4.2 | Aggregation effect on variance | 41 |
| 4.2.1 | Case 1 : constant group size | 47 |
| 4.2.2 | Case 2 : Different group size but allowing a constant $\bar{\Sigma}_g$ and $\bar{\Delta}_g$ | 49 |
| 4.3 | Empirical perspective of squared differences of pairs of observations | 50 |
| 4.3.1 | Relationship between $\hat{\Gamma}_{gh}$ and $\hat{\gamma}_{ij}$ | 52 |
| 4.3.2 | Mean square error within the group – $S_{yy}^{<W>}$ | 54 |
| 4.3.3 | Empirical aggregation effect | 55 |
| 4.4 | Aggregation effects in term of variance of differences | 56 |
| 4.5 | Summary | 61 |
| 5 | The Role of Semivariogram In Aggregation Effect | 63 |
| 5.1 | Variogram and semivariogram | 63 |
| 5.1.1 | Definition and assumptions | 64 |
| 5.1.2 | Semivariogram model and its parameter | 66 |
| 5.1.3 | Estimation of the semivariogram | 68 |
| 5.1.4 | Illustration of semivariogram from Illawarra dataset | 70 |
| 5.1.5 | Relationship between semivariogram and spatial autocorrelation | 72 |
| 5.1.6 | Nugget effect and spatial autocorrelation at zero distance | 74 |
| 5.1.7 | Illustration of spatial autocorrelation from Illawarra dataset | 75 |
| 5.1.8 | Generating random observations based on semivariogram model | 77 |
| 5.2 | Group level variogram ($\Gamma(d_{gh})$) | 78 |
| 5.2.1 | Fitting a group level semivariogram model | 81 |
| 5.3 | Deriving group level semivariogram in terms of individual level variogram by Taylor series expansion | 81 |
| 5.3.1 | Approximation of $\Gamma(d_{gh})$ | 85 |
| 5.3.2 | Approximation of moment structure of random distance within and between groups. | 88 |
| 5.3.3 | Evaluating $\hat{\Gamma}_{gh}$ for exponential model | 103 |
| 5.4 | Estimation of individual level semivariogram parameters from the group level semivariogram | 105 |
| 5.4.1 | Development of the method | 105 |
| 5.4.2 | Illustration from simulated data | 112 |
| 5.4.3 | Output of the estimation process using group level data | 114 |
| 5.5 | The weighted version of group level semivariogram | 121 |
| 5.5.1 | Illustration from simulated data | 124 |
| 5.5.2 | Discussion | 132 |
| 5.6 | Summary | 132 |
| 6 | Cross-Semivariogram | 134 |
| 6.1 | Introduction | 135 |
| 6.2 | Basic theorems of covariance | 137 |
| 6.2.1 | Individual level covariances | 137 |
| 6.2.2 | Group level covariance | 139 |
| 6.2.3 | Relationship between individual and group level covariance | 143 |
| 6.2.4 | Aggregation effect | 144 |
| 6.3 | Cross-semivariogram | 145 |
| 6.3.1 | Assumptions and definitions | 145 |
| 6.3.2 | Relationship between cross-semivariogram and spatial correlation | 147 |

6.3.3 The empirical cross-semivariogram 148

6.3.4 Relationship between cross-semivariogram and semivariogram 149

6.3.5 Generating random observations based of a cross-semivariogram model 151

6.4 Group level cross-semivariogram 152

6.5 Relationship between sample covariances and cross-semivariogram 153

6.5.1 Relative covariance of N_g and $\bar{\gamma}_{ab_g}$ 155

6.5.2 Expectation of the $S_{ab}^{<W>}$, $|\bar{S}_{ab}$, and ${}_N\bar{S}_{ab}$ 155

6.6 Aggregation effect in terms of cross-semivariogram 157

6.7 The weighting factors of the group level cross-semivariogram 158

6.8 Estimation of individual level cross-semivariogram parameters
from the group level cross-semivariogram 162

6.9 Summary 167

7 The MAUP as a tool in Semivariogram Analysis 169

7.1 Introduction 169

7.2 Development of the study on the MAUP 169

7.3 Evidence of the MAUP 170

7.3.1 Theoretical evidence 170

7.3.2 The empirical evidence 172

7.4 Empirical illustration of the MAUP 175

7.4.1 Simulation process 175

7.4.2 The theoretical background 178

7.4.3 Some illustration from the simulation 179

7.5 The use of the MAUP as an analysis tool : spatial analysis development 188

7.5.1 Exponential model of semivariogram 189

7.5.2 The scale effect 192

7.5.3 The zoning effect 193

7.6 Discussion 194

7.7 Summary 195

8 Empirical Analysis 196

8.1 Geographical aspects of Adelaide region 196

8.2 Description of the characteristics 198

8.3 Spatial graphical description of the characteristics 201

8.4 Semivariogram analysis of the Adelaide data 206

8.4.1 Summary 212

8.5 The micro sample of the Adelaide data 216

8.6 Estimation of individual level semivariogram parameters by non-linear model 217

8.7 Cross-semivariogram analysis 220

8.7.1 Estimation of individual level cross-semivariogram parameter using non-linear re-
gression methods 223

8.8 Summary 225

9 Summary and Discussion 228

9.1 Basic aggregation effect 229

9.2 Semivariogram approach to aggregation effect 230

9.2.1 Empirical perspective on the semivariogram and aggregation effect 231

9.2.2 Theoretical perspective on the semivariogram and aggregation effect 231

9.3 Relationship of group level semivariogram and individual level semivariogram 234

9.4 Group level semivariogram adjustment 234

10 Conclusion 236

10.1 Contributions, recommendations, and limitations 236

10.1.1 Some contributions 236

| | |
|--|------------|
| 10.1.2 Recommendations | 239 |
| 10.1.3 Limitations | 240 |
| 10.2 Further research | 240 |
| A Results of analysing Adelaide CD data | 243 |
| A.1 Graphical view of spatial perspective of the characteristics | 243 |
| A.2 Semivariogram results | 252 |
| A.3 Tabulation of micro-sample data of the Adelaide region | 261 |
| B Longitude & Latitude Conversion | 262 |
| C Evaluation of the Probability density function of the random distance within the region | 265 |
| D SAS codes for semivariogram analysis of Illawarra data | 267 |
| E Non-linear procedure for estimating n, s, and r | 269 |
| E.1 Fortran program | 269 |
| E.2 SAS procedures | 276 |
| F Description of Fortran codes and SAS procedure | 278 |
| F.1 Simulation of semivariogram and cross-semivariogram | 278 |
| F.2 The grouping process | 278 |
| F.3 Empirical semivariogram computation | 279 |
| F.4 Exploring the MAUP | 279 |
| F.5 Model fitting procedures | 279 |
| F.6 Miscellaneous subroutine | 280 |
| F.7 Some notes of Fortran | 280 |
| G Data sets | 281 |
| G.1 Illawarra data set | 281 |
| G.2 Adelaide data set | 281 |
| G.3 Simulated data set | 282 |
| References | 284 |

List of Tables

| | | |
|------|---|-----|
| 5.1 | The SAS output of the estimation procedure | 72 |
| 5.2 | Moran coefficient of the labor participation rate of the Illawarra data at difference neighborhood distance | 76 |
| 5.3 | The mean and variances of random distances within a unit area | 91 |
| 5.4 | The mean and variance of the distance within groups. Comparison between the simulated value and its approximation | 102 |
| 5.5 | The description of the mean and variance of the distance between groups. Comparison between the simulated values and its approximation (5.87) and (5.89) respectively | 103 |
| 5.6 | Descriptive values of Figure (5.18-b) | 113 |
| 5.7 | Descriptive values of difference between the estimated parameters of individual level and group level semivariogram | 113 |
| 5.8 | The mean of difference between the estimated parameters based on two different initial values of the nugget, equal the sill and zero | 115 |
| 5.9 | The estimated parameters based on two different initial values of the nugget, equal the sill and zero | 116 |
| 5.10 | Descriptive values of the Fig. 5.20 | 117 |
| 5.11 | Correlation between parameters estimated of approach 1a and approach 1b | 118 |
| 5.12 | Description of distribution of the estimated parameters by approach 2 and approach 3 | 119 |
| 5.13 | Correlation between parameters estimated of approach 2 and approach 3 | 120 |
| 5.14 | The estimated parameter of the model (5.123) of individual and unweighted group level semivariogram | 125 |
| 5.15 | The weighting factor | 125 |
| 5.16 | Description of the nugget, sill, and range distribution, of the individual and unweighted group level semivariogram model. | 128 |
| 5.17 | Description of the nugget distribution | 128 |
| 5.18 | Description of the sill distribution | 128 |
| 5.19 | Description of the range distribution | 132 |
| 6.1 | The weighting factor | 158 |
| 6.2 | The description of the estimated parameter of the individual population | 159 |
| 6.3 | The description of the estimated parameter of the group level cross-semivariogram, un-weighted and weighted | 162 |
| 6.4 | Estimated individual level parameters of cross semivariogram as shown in figure (6.8), (6.9), and (6.10) | 167 |
| 7.1 | Number of groups for each grouping factor | 172 |
| 7.2 | Some statistics of the variables at different grouping factors | 173 |
| 7.3 | Moran autocorrelation coefficient of Adelaide CD data at different neighborhood distances | 175 |
| 7.4 | Twenty different number of groups (scales) | 177 |
| 8.1 | Description of the characteristics (rate %) | 199 |

| | | |
|------|--|-----|
| 8.2 | The unweighted and weighted mean rate and variance | 200 |
| 8.3 | Estimated parameters of the exponential semivariogram model of the employment rate . . . | 207 |
| 8.4 | Estimated parameters of the exponential semivariogram model of the unemployment rate . . | 208 |
| 8.5 | Estimated parameters of the exponential semivariogram model of the labor participation rate | 210 |
| 8.6 | Summary of labor force characteristics | 213 |
| 8.7 | Summary of the income characteristics | 214 |
| 8.8 | Summary of the nature of the employment characteristics | 215 |
| 8.9 | Summary of the qualification achievement characteristics | 215 |
| 8.10 | The rate and variance of the characteristics from the micro sample of Adelaide | 217 |
| 8.11 | Estimated n , s , and r of the employment rate semivariogram model by non-linear model at different grouping factor levels, when individual and group level variance is known | 219 |
| 8.12 | Non-spatial correlation coefficient of employment rate versus its component at the CD level data | 220 |
| 8.13 | Estimated parameters of the exponential cross-semivariogram model of the formal qualifi- cation rate and rate of income below 20,000 at CD level | 221 |
| 8.14 | Estimated parameters of the exponential cross-semivariogram model of the formal qualifi- cation rate and rate of income over 40,000 | 222 |
| 8.15 | Spatial correlation and codispersion coefficient at distance 0 and r_{ab} | 223 |
| 8.16 | Non-spatial correlation coefficient of employment rate versus its component | 224 |
| 8.17 | Estimation of n , s , and r of the employment rate versus income below 20,000 and employ- ment rate versus income over 40,000 | 224 |
| 8.18 | Comparison of \hat{s}_{ms}^2 , S_{yy} , and estimated sill from the non-linear estimation methods of the employment rate of Adelaide data | 226 |
| 8.19 | The comparison of the sill of the group level semivariogram parameter estimates and the variance of the micro sample | 227 |
| 9.1 | The weighting factors and the expectation of $(\bar{Y}_g - \bar{Y}_h)^2$ | 235 |
| A.1 | Estimated parameters of the exponential semivariogram model of rate of the income below 20000 | 253 |
| A.2 | Estimated parameters of the exponential semivariogram model of rate of the income be- tween 20000 and 40000 | 253 |
| A.3 | Estimated parameters of the exponential semivariogram model of rate of the income over 40000 | 254 |
| A.4 | Estimated parameters of the exponential semivariogram model of rate of the wage or salary earner | 255 |
| A.5 | Estimated parameters of the exponential semivariogram model of rate of the self employed person | 256 |
| A.6 | Estimated parameters of the exponential semivariogram model of rate of the employer . . . | 257 |
| A.7 | Estimated parameters of the exponential semivariogram model of rate of the formal quali- fication | 258 |
| A.8 | Estimated parameters of the exponential semivariogram model of rate of the informal qual- ification | 259 |
| A.9 | The tabulation of the labor force status, income, nature of income, and qualification level . | 261 |

List of Figures

| | | |
|------|---|-----|
| 2.1 | Aggregation process ignores the location of individuals. Variation within groups is lost. . . | 10 |
| 3.1 | Aggregation process of unit level into CD level | 19 |
| 3.2 | Relationship of variables; (a) relationship between variables ignoring spatial inter-dependence, (b) spatial inter-dependence in variable, and (c) relationship between variables considering spatial inter-dependence. | 21 |
| 5.1 | Graph of the theoretical semivariogram model | 66 |
| 5.2 | Distance classes | 69 |
| 5.3 | The CD distribution of the labor participation rate (%) in Illawarra NSW, mean=58.55, median=58.68, min.=29.64, max.=80.34, variance=73.38 | 71 |
| 5.4 | Empirical semivariogram and empirical exponential model with nugget 44.49, sill 75.98, and range 9.81 | 72 |
| 5.5 | The positive side of spatial autocorrelation model based on parameters of semivariogram model | 74 |
| 5.6 | Moran coefficient and exponential model of spatial autocorrelation for the labor participation rate of the Illawarra data | 76 |
| 5.7 | The likely relative value of γ' and γ'' compared with the average distance between group and group area size, the symbol $\circ = \gamma'$, and $\Delta = \gamma''$ | 86 |
| 5.8 | Group level semivariogram and its individual level semivariogram | 88 |
| 5.9 | Relationship of the mean and variance of the random distance with perimeter of the unit area | 91 |
| 5.10 | Relationship of the mean and variance of the random distance within the region with area of different shapes of the region | 92 |
| 5.11 | Approximation of the mean and variance of random distance within the region by regression method | 93 |
| 5.12 | The random distance between points in two groups | 95 |
| 5.13 | The shapes considered in the simulation, (a) square, (b) rectangle, (c) circle, (d) equilateral triangle, (e) "L-shape". | 99 |
| 5.14 | Relationship between average distance within group with the area of the group at different shapes of the groups, note : the \blacktriangle indicates the mean point. | 100 |
| 5.15 | Relationship between variance of distance within group with area of the group at different shapes of the groups, note : the \blacktriangle indicates the mean point. | 101 |
| 5.16 | The plot between the simulated and approximated of the mean and variance distance between groups | 102 |
| 5.17 | (a) Relationship between $(\hat{\Gamma}_{gh} - \Gamma_{gh}^0)$ and \bar{d}_{gh} , (b) boxplot of $(\hat{\Gamma}_{gh} - \Gamma_{gh}^0)$ | 106 |
| 5.18 | (a) The categorized version of the individual and group level semivariogram; (b) distribution of estimated parameters of individual level semivariogram model from 1000 simulations | 112 |
| 5.19 | Distribution of the group level semivariogram estimated parameters nugget, sill, and range, (a) 1000 simulations (b) 950 simulation | 114 |
| 5.20 | Distribution of parameters estimated by approach 1a and approach 1b | 117 |
| 5.21 | Distribution of parameters estimated by approach 2 and approach 3 | 119 |

| | | |
|------|---|-----|
| 5.22 | Exponential model of $\gamma(d_{ij}) = 15 + 15(1 - \exp(-3\frac{d_{ij}}{20}))$ | 124 |
| 5.23 | The first two simulation results. Note : \circ = individual level and \bullet = group level | 125 |
| 5.24 | The first two simulation results of the weighting factors. Note : \blacksquare = individual level and \bullet = unweighted group level, $\square = N\hat{\Gamma}_{gh}$, $\Delta = \hat{\Gamma}_{gh}^{w2}$ | 126 |
| 5.25 | Semivariogram model parameter estimator distribution | 127 |
| 5.26 | Nugget Distribution of individual level, unweighted and weighted group level | 129 |
| 5.27 | Sill Distribution of the individual level, unweighted and weighted group level | 130 |
| 5.28 | Range Distribution of the individual level, unweighted and weighted group level | 131 |
| 6.1 | A diagram of relationship between income and employment status | 136 |
| 6.2 | Cross-semivariogram and their corresponding semivariogram | 150 |
| 6.3 | Simulation result of the cross-semivariogram with ($n_{ab} = 5$, $s_{ab} = 8$, $r_{ab} = 10$). | 152 |
| 6.4 | The distribution of the estimated individual population parameter from the 100 simulations | 159 |
| 6.5 | Distribution of the estimated nugget from the individual level, group level, and the two different weighted group level cross-semivariogram | 160 |
| 6.6 | Distribution of the estimated sill from the individual level, group level, and the two different weighted group level cross-semivariogram | 161 |
| 6.7 | Distribution of the estimated range from the individual level, group level, and the two different weighted group level cross-semivariogram | 161 |
| 6.8 | Distribution of the estimated individual level parameter nugget, sill, and range, when \hat{S}_{aa} , \hat{S}_{bb} , and \hat{S}_{ab} are available. | 165 |
| 6.9 | Distribution of the estimated individual level parameter nugget, sill, and range, when only \hat{S}_{aa} and \hat{S}_{bb} are available. | 166 |
| 6.10 | Distribution of the estimated individual level parameter nugget, sill, and range, when individual sample was not available | 166 |
| 7.1 | Boxplot at different grouping factors, ranging from CD level to LGA level. | 174 |
| 7.2 | The grouping process of the region with 7×8 grids. | 176 |
| 7.3 | The semivariogram plot of the individual population | 177 |
| 7.4 | Coefficient of variation of the N_g and the relationship of \bar{N} (scale) with the $N\bar{S}_{yy}$ | 179 |
| 7.5 | Relationship between \bar{N} and each factor of the $N\bar{S}_{yy}^{(1)}$ (\circ), $N\bar{S}_{yy}^{(2)}$ (Δ), and $N\bar{S}_{yy}^{(3)}$ ($+$). The \blacklozenge is the $N\bar{S}_{yy}$ | 180 |
| 7.6 | Separate plot between $N\bar{S}_{yy}^{(2)}$ and $N\bar{S}_{yy}^{(3)}$ versus \bar{N} | 181 |
| 7.7 | Relationship between \bar{N} and $\tilde{\gamma}_w$. The curve line indicates the relationship with the model $\tilde{\gamma}_w = 4.988 + \bar{N}^{0.4687}$, with $MSE = 0.0209$ | 182 |
| 7.8 | The scatter plot of $\bar{C}_{N\tilde{\gamma}}$ and $\bar{R}_{N\tilde{\gamma}}$ | 182 |
| 7.9 | Relationship between \bar{N} with the aggregation effect, (a) in term of difference $N\bar{S}_{yy} - S_{yy}$ and (b) in term of ratio $\frac{N\bar{S}_{yy}}{S_{yy}}$ | 183 |
| 7.10 | (a) Boxplot of the $N\bar{S}_{yy}$ at each scale, (b) the scatter plot of the variation of $N\bar{S}_{yy}$ each scale versus \bar{N} , (c) boxplot of $\bar{R}_{N\tilde{\gamma}}$ at each scale. | 184 |
| 7.11 | (a) Relationship between $\tilde{\gamma}_w$ with $N\bar{S}_{yy}$, (b) relationship of variation of $\tilde{\gamma}_w$ at each scale with the average of $N\bar{S}_{yy}$ at each scale, (c) relationship between $\bar{C}_{N\tilde{\gamma}}$ with $N\bar{S}_{yy}$, and (d) relationship of variation of $\bar{C}_{N\tilde{\gamma}}$ at each scale with the average of $N\bar{S}_{yy}$ at each scale. | 186 |
| 7.12 | Relationship between \bar{N} with the variance of the aggregation effect in term of difference and ratio at a particular \bar{N} | 186 |
| 7.13 | Coefficient of variation of the \mathcal{A}_g over the whole simulation | 187 |
| 7.14 | Relationship between \bar{N} with the average group area, (a) for whole scale and (b) only the last scale | 187 |
| 7.15 | Some examples of the $\tilde{\mathcal{A}}_w^*$ values from different zoning at the region of area 4.0, (a) $\tilde{\mathcal{A}}_w^* = 1.0$, (b) $\tilde{\mathcal{A}}_w^* = 0.9659$, and (c) $\tilde{\mathcal{A}}_w^* = 0.9830$ | 191 |
| 7.16 | The weighted group level variance at different scale | 193 |
| 7.17 | The weighted group level variance at different zoning scheme | 194 |

7.18 The scale and zoning effect 195

8.1 Adelaide region, (a) CD boundaries – based on Australian Census 1991, (b) CD’s centroid.
The region is approximately 33 km wide and 55 km long. 197

8.2 Scatter plot, contour and surface plot of the employment rate 202

8.3 Scatter plot, contour and surface plot of the unemployment rate 204

8.4 Scatter plot, contour and surface plot of the labor participation rate 205

8.5 Semivariogram model fitting for the employment rate 207

8.6 Exponential semivariogram model fitting for the unemployment rate 209

8.7 Exponential semivariogram model fitting for the labor participation rate 211

8.8 Cross-semivariogram model fitting for the formal qualification rate with the rate of income
below 20,000 221

8.9 Cross-semivariogram model fitting for the formal qualification rate with the rate of income
over 40,000 222

A.1 Scatter plot, contour and surface plot of rate of the income less than 20000 244

A.2 Scatter plot, contour and surface plot of rate of the income between 20000 to 40000 245

A.3 Scatter plot, contour and surface plot of rate of the income greater than 40000 246

A.4 Scatter plot, contour and surface plot of rate of the wage or salary earner 247

A.5 Scatter plot, contour and surface plot of rate of the self employed persons 248

A.6 Scatter plot, contour and surface plot of rate of the employer 249

A.7 Scatter plot, contour and surface plot of rate of the formal qualification 250

A.8 Scatter plot, contour and surface plot of rate of the informal qualification 251

A.9 Exponential semivariogram model fitting for rate of income below 20000 252

A.10 Exponential semivariogram model fitting for rate of income 20000-40000 254

A.11 Exponential semivariogram model fitting for rate of income over 40000 255

A.12 Exponential semivariogram model fitting for rate of wage or salary earner 256

A.13 Exponential semivariogram model fitting for rate of self employed 257

A.14 Exponential semivariogram model fitting for rate of employer 258

A.15 Exponential semivariogram model fitting for formal qualification rate 259

A.16 Exponential semivariogram model fitting for the informal qualification rate 260

Dedication

*Siti Hanifah,
Arif and,
Sadikin*

... you are my sunshine.

Chapter 1

Introduction

There is an increasing tendency to take a spatial perspective in analysing census or sample data. Many physical and social phenomena exhibit strong spatial aspects. This research contributes to the development of spatial analysis and concentrates on methods for analysing data on social characteristics. The important case of aggregated census data will be considered.

This chapter introduces a spatial perspective in analysing aggregated census data. The main points considered are the data, targets of inference and approach to analysis. General problems and motivation of the research are cited and the objectives and outputs are outlined.

1.1 Background : data, targets of inference, and analysis

Social data mainly come from observations of persons or households. Each observation may contain social and spatial characteristics. These observations constitute individual level social data. There are often reasons why aggregation is applied to individual level data. Aggregation will create aggregate level social data. The aggregation process is based on groups in the population (section 3.2.1).

If we are interested in spatial relationships, then we must consider how to analyse social data that have been obtained by methods of sampling or aggregation. There may not be a direct interest in spatial relationships, but the presence of spatial interdependence may still need to be taken into account in the analysis. There may be spatial trends in means and variances, and the correlation between the characteristics of different individuals that depend on their relative locations.

We will focus on spatial approaches based on analysis of variograms. The variogram is a statistic which refers to the square of the difference between two observations at different locations. Methods are developed to analyse variogram to infer spatial relationships at the aggregated and individual level (chapter 4).

1.2 Problems and motivation

1.2.1 The problems

Statistical analysis of social data is often done using aggregated data, because of the availability of such data, and constraint on funding and time (Langbein & Lichtman, 1978). Problems arise when the data are available in aggregated form but inferences concerning individual level relationships are required. This involves ecological inference (King, 1997). The problems will be detailed in chapter (2.2).

Two factors should be taken into account when performing analysis of aggregated data; aggregation bias and the statistical methods problem (King, 1997). Aggregation bias arises because variation at the individual level data is lost as a result of the transformation from individual to aggregate data. The second factor is that many basic statistical methods are not suitable for the analysis of aggregated data. These factors will affect the results of such analysis. Two key problems associated with analysing aggregate data are the modifiable areal unit problem (MAUP) and the ecological fallacy (section 2.2.1).

Some solutions to the MAUP and ecological fallacy have been suggested, such as in King (1997) and Steel and Holt (1996a). For example, Steel and Holt (1996a) proposed a model and developed methods to analyse aggregated data and offered a way to adjust the aggregation effects to provide less biased estimates of unit level parameters.

In this research, we will show that the MAUP is due to the spatial relationships at the individual level. If we can estimate the spatial relationships at the individual level then the MAUP is resolved. In fact, once we recognize that the MAUP is due to spatial relationships at the individual level, then the problem becomes how to estimate these relationships from the available data. Chapter (4) will discuss this question in more detail.

1.2.2 Motivation : what and why spatial analysis

The analysis of social data often ignores the spatial characteristics and usually assumes that the observations are distributed independently and identically (IID). This may lead to incorrect results. Observations from different individual units may exhibit inter-relationships. One individual may be influenced by others at nearby locations. As a result, social data may show dependence between observations, and hence the IID assumptions will not be appropriate. The data may be obtained in ways that depend on location, through the sampling method and/or aggregation process to give aggregate data for a particular set of areal units. For these reasons a spatial perspective provides an appropriate way to treat social data. In the spatial perspective, we can account for the presence of spatial interdependence within the social data.

What is spatial statistics analysis ?

Cressie (1989) noted that most data have a space and time label associated with them. For example, in the Australian 1991 Census of Population and Housing, the space label is represented as the centroid location of each collection district (CD).

Data analyses are defined to be spatial if the locations are relevant for interpretation of the data, that is if spatial variability is important. Fotheringham, Charlton, and Brunsdon (1996) noted a distinction between the analysis of spatial data and spatial data analysis. Both analyses are done on spatial series data, where the physical locations are recorded for each observation. The former analysis ignores the location aspect and treats the data as if they were non-spatial. The latter analysis uses the location information extensively to examine spatial variability in the data.

Analysis on spatial variability leads to the development of statistics and models which can discriminate between different configuration of observations in a two dimensional surface. In chapter (2), some spatial statistics and models will be considered.

Why is spatial analysis used ?

This question will be considered by looking at spatial analysis in general and more specifically variogram analysis. Considering census data as spatial series may lead to alternative ways of analysing it. So, what is the significance of using spatial analysis?

In general, spatial analysis can disclose spatial variation, which will show whether the geographic locations of observations are important or not. Cressie (1991) noted that spatial analysis offers a more general way of analysing data than the classical approach, since it makes less assumptions, in particular no IID assumptions are made.

More specifically, the spatial analysis approach allows us to develop a model for describing the spatial correlation structure in a region. Once the spatial model is developed, interpolation may be used to estimate the value of a characteristic at a specific location. Morissette (1997) noted that a spatial model can allow for different spatial correlation structures at different spatial scales. Spatial analysis can be extended to spatial multi level modeling, which allows us to evaluate a correlation structure at different levels.

For example, we may estimate a variogram model at the aggregate level. The developed variogram model can summarize the spatial data by providing a measure of the spatial dependence between observations. A variogram model can be used with kriging to provide estimates of unmeasured observations at a specified location.

The latter point raises the possibility of looking at smaller spatial scales. Analysing aggregated level data using variogram methods may permit us to explore aggregation bias and the possibility of developing a method of estimating the individual level variogram parameters. These issues will be discussed further in chapter (4).

The spatial approach has a minor disadvantage in term of the computations involved, requiring powerful software and extensive computer time. In the case of non-spatial data, the IID assumptions will eliminate the off-diagonal elements of the variance-covariance matrix. But for spatial analysis, all elements of the variance-covariance matrix are relevant, resulting in more complex and intensive computation.

1.3 Objectives

The main objective of this thesis is to develop methods of analysing aggregated social data. This study will explore the covariance structure when it is difficult to accept the IID assumption and spatial dependence is more realistic. Analysis of variogram (or semivariogram) will be performed to examine the covariance structure of the aggregated data.

There are several outputs expected from this research. First, we will examine how the variogram of the aggregate data can be used to explore covariance structure and how the variogram can explain spatial dependency. Second, connection between the variogram of the aggregated data and the variogram of the unit level data will be derived. Can analysis based on aggregate data provide an adjustment to infer unit level spatial relationships ? Also what is the effect of clustering people in households ? Third, the extension of the variogram analysis into the bivariate case involving cross-semivariogram analysis will be considered. Fourth, the relationship between spatial autocorrelation and the variogram will be investigated. Fifth, a methodology of using the MAUP as a tool to explore spatial dependence of aggregate social data will be developed. Simulation results and empirical analysis of actual aggregate data will be used to examine these issues. The empirical work will be based on analysis of the 1991 Australian Census of Population and Housing.

1.4 Scope

This thesis is divided into ten chapters. Some background of the problems and objectives of the research will be briefly discussed in the first chapter. The second chapter will look at some relevant definitions and provide a more detail discussion of the problems. The third chapter will discuss some previous results that have been presented in the literature. This literature study will review some important points in understanding the nature of the problems and survey some applied methods available to tackle the problems. The fourth chapter will develop the theoretical basics needed to develop analysis methods. The fifth chapter will discuss some methodologies for estimating individual level spatial relationships from aggregate data. The sixth chapter extends the methods to the cross-variogram. The seventh chapter will discuss the use of the MAUP as a tool in semivariogram analysis. The eighth chapter will present empirical work

based on the proposed approach based on analysis of aggregated data obtained from the 1991 Australian Census of Population and Housing for the Adelaide region. Finally, some discussion and conclusions will be presented in chapter nine and chapter ten, respectively.

Chapter 2

Definition and Problem Identification

This chapter defines the features of the population under study, and also identifies the problems to be tackled. The approaches and assumptions necessary are also defined.

2.1 Definition of population

Suppose there is a finite population of individual units $\mathcal{U} = \{1, \dots, N\}$ in a particular region \mathcal{D} . The region \mathcal{D} has an explicit boundary. Individual units could be people or households and are distributed within the boundaries of region \mathcal{D} . Associated with each individual are several social characteristics, which are observable and measurable, for example sex, income, employment status, etc. Each observation also has a space label, such as the geographical location of the individual. The social characteristics will be denoted by a vector \mathbf{Y} , where in the univariate case $Y_i \in \mathbf{Y}$ for $i \in \mathcal{U}$, and the spatial label will be ℓ_i . In a population consisting of people, one or more individuals may be at the same location.

2.1.1 Superpopulation approach

Analytical statistical methods consider the finite population values as a realization of random variables generated by a stochastic process. Theoretically the stochastic process will generate a hypothetical population, which is called the superpopulation (Skinner, Holt, & Smith, 1989).

The stochastic process is usually described by a model involving random variables with a specific probability distribution function. The model could be used to summarize superpopulation characteristics, leading to model-based inference. The objective of the inference is not limited to only describing a particular population but into a study of relationship between characteristics.

Spatial analysis may use analytical methods based on the superpopulation approach. Cressie (1991) notes that spatial analysis will treat the population as a random process and the observations are a realization of the random process.

2.1.2 Spatial series of social data

Spatial data arise in many disciplines, such as geology, forestry, geography, meteorology, remote sensing, ecology, economics, sociology, and other fields of study that have a spatial concern. Methods for the analysis of spatial data have been developed in many fields of study, as shown by the work by Anselin (1988), Haining (1990), Griffith and Amrhein (1991), Martin (1996), Burrough (1986), Arbia (1989a), Cliff and Ord (1981), Hagget, Cliff, and Frey (1977).

In some cases the spatial data follows the Gaussian distributional assumptions (Diggle, Tawn, & Moyeed, 1998). For a single characteristic the Gaussian spatial stochastic process considers the observations as a realization of the stationary Gaussian process ($S(l)$), that is

$$Y_i = \mu + S(\ell_i) + \epsilon_i \quad (2.1)$$

It is assumed that the μ is constant, $S(\ell)$ is a stationary Gaussian process with zero mean and covariance between Y_i and Y_j equal to $\sigma^2 \rho_{ij}$, and the ϵ_i are mutually independent $N(0, \tau^2)$.

Arbia (1989b) differentiated between continuous and discontinuous spatial data. The former is common in observing natural phenomena, such as rainfall, temperature, or other geological or environmental variables. Discontinuous data are common in observing social phenomena, where observations may have a discontinuity in space, e.g. where there are no individuals at a particular location. For example, the amount of income is measured on individuals where they are located and so there is not a value at every spatial location.

The discontinuous data may be represented as points, lines, or areas (Arbia, 1989b). For example, point data may lead to a study of point pattern, such as the occurrence of disease, crime, location of industrial plants, etc. Line data may indicate network like features, such as transportation networks, travel to work patterns, shortest distance problems. Areal data may be used to represent discontinuous data. In this case, aggregation may be involved, for example, points of data within a particular area or group are aggregated

such that it will represent the area's values. This set of representation may be effective in reporting data, but raises issues regarding the analysis of this type of data, which will be discussed in section (2.2.1).

2.1.3 Structure and sources of spatial data

Spatial data have two components of measurements. The first component consists of measurements describing the location of the object in space; we will refer to this component as the spatial characteristics. The second component consists of measurements that identify other properties of the object that we will call attribute characteristics. The attribute characteristics may be measured in nominal, ordinal, interval or ratio scales.

Arbia (1993) discussed sources of spatial data, which he categorized into census and administrative records, sample surveys, and satellite photographs. A major source of social data is the population census. The census collects data about units, such as persons or households, but the statistics released from the census and administrative sources are usually in aggregated or tabular form. Aggregate data provide a level of confidentiality protection.

2.2 Problem identification

Analysing aggregated data with conventional statistical methods raises a range of problems, which have been identified. Here we introduce some assumptions and the framework needed to develop some solutions.

2.2.1 Aggregation problems

What are the consequences of aggregation on social data ? Tranmer and Steel (1998) noted that appropriately weighted means are not affected by aggregation, but aggregation does affect their variances. The aggregate data will usually consist of the means for areal units which contain individuals that are close to each other, and will have some degree of similarity in their social characteristics. This phenomenon is called within-group homogeneity (Tranmer & Steel, 1998).

Aggregation bias can be illustrated in Figure (2.1). This figure illustrates a process of aggregation of individual level data into group level data. The region is divided into five subregions/groups and the resulting aggregate data consist of the mean value of each subregion/group. All groups have the same mean. It shows that information about the individual locations and variation within groups are lost in the aggregation process. Also, the assumption that the observations within a group are independent and identically distributed is unlikely to be reasonable.

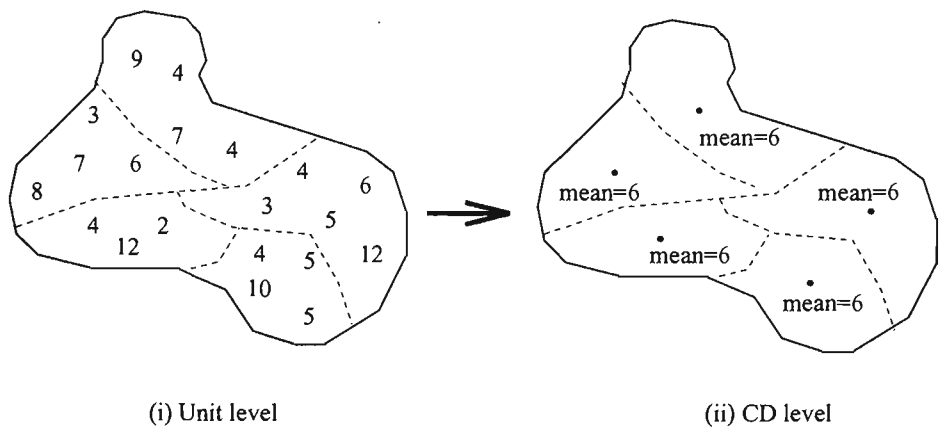


Figure 2.1. Aggregation process ignores the location of individuals. Variation within groups is lost.

From Figure 2.1 we can identify at least two kinds of problems. First, the aggregated data do not contain all the information about the unit level values. Variation at individual level data is lost by the process of aggregation. Hence, analysis of the aggregated data to infer individual relationship will often result in incorrect inferences. This is the ecological fallacy.

Second, the unit level data are often aggregated into artificial areal units. The boundaries of the groups are imposed rather than natural (Harvey, 1969). The arrangement shown in Figure 2.1 is only one example from many other possibilities for forming the groups. Results of the analysis will change as different groups are used. The problems of analysing spatially aggregated data are referred to the Modifiable Areal Unit Problem or MAUP. The MAUP consists of two main aspects, referred to as the scale effect and zoning effect (Openshaw, 1984). The zoning scheme refers to how the region is partitioned and arranged for a particular number of zones. The spatial scaling is related with how many zones are formed.

Aggregation effect

Steel and Holt (1996b) used the term of aggregation effect to cover the effects of allocating individual units into spatial groups and combining spatial groups at one level into higher level groups. Tranmer and Steel (1998) argued that the aggregation effect arises because the individuals who live close to one another tend to have a degree of similarity for particular social characteristics. The small spatial groups within the region therefore contain units which have similar characteristics.

Holt, Steel, and Tranmer (1996) noted that aggregation effects are caused by the non-random allocation of individuals to areas. Individuals in the same area (neighborhoods) generally tend to be more alike or show positive association for some social characteristics.

Spatial scale effect

Spatial scale effect refers to the difference of result derived from two or more different aggregation processes based on different spatial scales on the same region.

Anselin (1988) viewed spatial scale effects as consequence of measurement errors of observations in contiguous geographic regions. The aggregation caused the measurement errors to spill over across the boundaries of the spatial units. Therefore the error of one aggregated value is likely to be associated with the errors of a neighboring unit. He mentioned that the appearance of spatial scale effect may be due to the fundamental importance of space as an element in human behaviour.

Anselin's definition of the spatial scale effect indicates that aggregated data might be affected by the spatial or geographic location of the observations. It follows that aggregation effects may be studied through spatial analysis, using the spatial characteristics of the aggregated data. Implicitly he defined a connection between the aggregation effect and spatial scale effect.

Ecological fallacy

In some situations researchers may be interested in relationships between variables at the individual level, but the data available are group level aggregated data. This situation involves a mismatch between data availability and the target of inference. Applying analysis on aggregated data to infer the relationship at

the individual level may lead to the ecological fallacy. The ecological fallacy refers to the inconsistency in the results of analysing aggregated data compared with analysing individual level data (Robinson, 1950). The aggregation process transforms the individual value into aggregate value, at a particular spatial scale. Hence, the ecological fallacy is one case of the spatial scale effect.

2.2.2 Sampling design problems

Consider a simple situation in which the individuals are distributed spatially within the geographical region of interest. Their locations are not needed to be assumed as uniformly distributed. The analysis is conditional on the spatial locations. A grouping is then superimposed into the region creating groups of individuals.

Assume that spatial dependency is present among individuals within the region of interest, then the individuals within the group are not independent. Arbia (1993) discussed two issues that arise from this spatial dependency. The first, spatial dependency has been recognized in data analysis or modeling but almost neglected in the context of sampling design. The second, is that some techniques of spatial statistics have been developed based on regular shaped regions, but in practice we often deal with irregular shaped regions.

Given a sampling method, the questions of how large the sample should be and how much precision will be achieved need to be considered. The precision of the sample is usually measured by the standard error of the estimate. For a given standard error of the estimate we may determine the required sample size.

The spatial perspective on the use of aggregate social data may be employed to derive implications for determination of standard error of the estimate. Recall that the mean is not affected by aggregation (Tranmer & Steel, 1998). A problem may arise in estimating standard error of the estimate when the observed characteristics are spatially correlated. Griffith, Haining, and Arbia (1994) discussed a solution to the problem of estimating the standard error of spatially correlated characteristics.

The spatial location of the objects may affect how samples are selected. Fotheringham and Rogerson (1993) noted that cluster samples may be applied on observations which exhibit low spatial variation and

lead to lower collection cost but a higher estimate of variance. For observations which exhibit greater spatial variation, systematic sampling may be a desirable simple method.

Cressie and Aldworth (1997) noted that the best spatial sampling plan indicates the selection of locations at which to sample the phenomena in order to achieve optimality according to a given criterion, for example, minimize the average mean squared prediction error. Furthermore, they compared the performances of several sampling designs, such as systematic random sampling, stratified random sampling, simple random sampling, and cluster sampling. They considered the estimation of the spatial mean and spatial cumulative distribution function. The result showed that cluster sampling should not be used when estimating the spatial mean and spatial cumulative distribution function, where as systematic, stratified, and simple random sampling may be used. They also concluded that good spatial analysis gives superior results regardless of the design employed.

Griffith et al. (1994) discussed the estimation of the standard error of the estimate of the mean from a spatially correlated variable in the case where data are obtained by a process of random sampling. An appropriate sampling design in a spatial context will have two different sources of error, these being attribute sampling error and location error. Location error arises, for example, from the process of representation of the areal unit, e.g. boundaries definition, group formation. The attribute sampling error arises from estimation of the population parameters, e.g. mean and variance.

Pettit and McBratney (1993) carried out a study of sampling designs and estimation procedures of the variogram for regionalized variables. The regionalized variables can be considered in the same way as the aggregated data since the values of regionalized variables represent the values at a particular areal unit. The study suggested that highly unbalanced staggered design may be more efficient in terms of sampling effort than balanced nested design.

2.3 Some spatial assumptions and implementations

Consider a single social characteristic and let $Y_i[\ell_i]$ represent the value of the characteristic of the i th unit which is located at ℓ_i , and the j th unit may have a location $\ell_j = \ell_i$. It is reasonable to assume that

there may be dependence between values of characteristics at difference locations. We initially assume the following moment structure for $i, j \in \mathcal{U}$;

$$\begin{aligned}
 & \text{(i)} \quad E(Y_i[\ell_i]) = \mu_i(\ell_i); \quad \ell_i \in \mathcal{D} \subset \mathcal{R}^2 \\
 & \text{(ii)} \quad V(Y_i[\ell_i]) = \Sigma_i(\ell_i); \quad \ell_i \in \mathcal{D} \subset \mathcal{R}^2 \\
 & \text{(iii)} \quad \text{Cov}(Y_i[\ell_i]; Y_j[\ell_j]) = \Delta_{ij}(\ell_i; \ell_j); \quad \ell_i, \ell_j \in \mathcal{D} \subset \mathcal{R}^2; \quad \ell_i \neq \ell_j \\
 & \text{(iv)} \quad \text{Cov}(Y_i[\ell_i]; Y_j[\ell_i]) = \Delta_{ij}(\ell_i; \ell_i); \quad \ell_i \in \mathcal{D} \subset \mathcal{R}^2; \quad i \neq j \in \mathcal{U}
 \end{aligned} \tag{2.2}$$

The condition (2.2-iii) reflects a relation between two different individuals at two different locations. The condition (2.2-iv) shows a relation of two different individuals at the same location. Note that $\Delta_{ij}(\ell_i, \ell_i)$ is not $\Sigma(\ell_i)$. We will not allow the same individual at two different locations, hence $\text{Cov}(Y_i[\ell_i]; Y_i[\ell_j])$ is not defined.

Define

$$\rho_{ij}(\ell_i, \ell_j) = \frac{\Delta_{ij}(\ell_i; \ell_j)}{\sqrt{\Sigma(\ell_i) \cdot \Sigma(\ell_j)}} \quad \ell_i, \ell_j \in \mathcal{D} \tag{2.3}$$

The correlation between individuals will affect some inferences. For example consider the population mean $\bar{Y} = \frac{1}{N} \sum_{i \in \mathcal{U}} Y_i$, $V(\bar{Y})$ can be expressed as;

$$V(\bar{Y}) = \frac{\bar{\Sigma}}{N} (1 + (N-1)\bar{\rho}) \tag{2.4}$$

where,

$$\bar{\Sigma} = \frac{1}{N} \sum_{i \in \mathcal{U}} \Sigma_i(\ell_i) \quad \text{and} \quad \bar{\rho} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \rho_{ij}(\ell_i, \ell_j)$$

The usual IID assumption would imply $\bar{\rho} = 0$ and $V(\bar{Y}) = \bar{\Sigma}/N$. But when there is spatial interdependency, then we cannot ignore the $\bar{\rho}$ term. Using ordinary statistical analysis methods would give incorrect inferences.

The spatial assumptions start from a view point that there is a spatially dependence within the region \mathcal{D} , depending on the location of individuals. Often it is assumed that spatial dependence may be formulated as the correlation being a function of distance between individuals by assuming the process is second order stationary and intrinsically stationary (Cressie, 1991). Further discussion of these assumptions will be given later in chapter (5).

Suppose that the region \mathcal{D} is partitioned into M groups denoted by $\mathcal{D}_1, \dots, \mathcal{D}_M$, and consider the group means

$$\bar{Y}_g = \frac{1}{N_g} \sum_{\{i; \ell_i \in \mathcal{D}_g\}} Y_i; \quad g = \{1, \dots, h, \dots, M\} \quad (2.5)$$

where N_g is the number of individuals in \mathcal{D}_g . For convenience we will write $\sum_{i \in \mathcal{U}_g}$ to denote $\sum_{\{i; \ell_i \in \mathcal{D}_g\}}$, where $\mathcal{U}_g = \{1, \dots, N_g\}$.

The group level means have the following properties,

$$\begin{aligned} \text{(i)} \quad E(\bar{Y}_g) &= \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} \mu_i(\ell_i) = \bar{\mu}_g; \\ \text{(ii)} \quad V(\bar{Y}_g) &= \frac{1}{N_g} (\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g); \\ \text{(iii)} \quad \text{Cov}(\bar{Y}_g, \bar{Y}_h) &= \bar{\Delta}_{gh}; \quad g \neq h \end{aligned} \quad (2.6)$$

where,

$$\bar{\Sigma}_g = \frac{\sum_{i \in \mathcal{U}_g} \Sigma_i(\ell_i)}{N_g}; \quad \bar{\Delta}_g = \frac{\sum_{i \neq j \in \mathcal{U}_g} \Delta_{ij}(\ell_i, \ell_j)}{N_g(N_g - 1)}; \quad \bar{\Delta}_{gh} = \frac{\sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \Delta_{ij}(\ell_i, \ell_j)}{N_g N_h} \quad (2.7)$$

The $\bar{\Sigma}_g$ is an average of variance within the g th group. The $\bar{\Delta}_g$ is an average of the covariance of units within the g th group and $\bar{\Delta}_{gh}$ is an average of the covariances between units in the g th and h th group.

These results indicate that analysing group level data will be affected by $\bar{\Sigma}_g$, $\bar{\Delta}_g$, and $\bar{\Delta}_{gh}$. Using the superpopulation approach allows us to explore $\bar{\Sigma}_g$, $\bar{\Delta}_g$, and $\bar{\Delta}_{gh}$ from a spatial perspective (see Ripley, 1981). Variogram (or semivariogram) analysis is a common way to explore the covariance structure in spatial data. The spatial dependence is examined by seeing how close the g th group's value relates to another group's value in terms of the distance between the groups. How and how far the spatial properties of the aggregate data can be used to infer unit level relationships are the main topics in this study.

Chapter 3

Literature Review

In the last two decades, there have been many fruitful developments in spatial analysis as reported by Unwin (1998) and Griffith (1996). In this chapter some key developments for dealing with spatial data are discussed. The analysis of spatial data is considered in general, with special relevance to analysing aggregated spatial data.

3.1 Background

Analysing data can be done using descriptive methods and analytical methods. The difference between the two is in the definition of the target of inference. Descriptive methods only attempt to describe the particular population under study. While analytical methods aim not only to describe relationships in the particular population under study, but also to analyse relationships between characteristics that may apply in other populations. Analytical methods assume the values in the population are the realization of some stochastic process, which can be represented by a statistical model. A more detailed discussion of these issues can be found in Skinner et al. (1989).

Census data are usually presented in aggregated form and viewed as non-spatial data. Analysing relationships using aggregated data leads to problems, such as the ecological fallacy and the MAUP. Some methods have been introduced to tackle these problems, such as discussed in Tranmer and Steel (1998), King (1997), Steel, Holt, and Tranmer (1996), Steel and Holt (1996a, 1996b), and Holt et al. (1996).

Relaxing the IID assumptions on which standard analytical methods are based and considering spatial relationships may open an alternative perspective to handling census and often aggregate social data. This approach to data analysis will take into account spatial variability among observations, and how this variation may be related to distances between individuals. Various articles have provided an overview of the theoretical and methodological aspects of the spatial perspective, such as Cliff and Ord (1981), Cressie (1991), Ord and Getis (1995), Getis and Ord (1992), and Diggle et al. (1998). Applications of this approach may be found in Griffith et al. (1994), McCracken (1983), Müller, Stadtmüller, and Tabnak (1997), Clifford, Richardson, and Hémon (1989). These articles will be discussed later in this chapter.

3.2 Aggregation

Aggregation involves summing data at one level to produce data at another higher level. For social data the lowest level data that is potentially available usually refers to people. Aggregation then involves summary the data from groups of people. The groups may be essentially non-spatial, such as schools or hospitals, but we will focus on situation in which groups are defined spatially. Aggregation also occurs when data from low level groups are summed to produce data at a higher level.

3.2.1 Group formation : zoning and scaling

Group formation is the first step of an aggregation process. Groups are defined by drawing boundaries across the population region so that the region is partitioned into smaller area. The region \mathcal{D} is partitioned into M smaller areas $\mathcal{D}_g, g = \{1, \dots, h, \dots, M\}$, that is

$$\mathcal{D} = \bigcup_{g=1}^M \mathcal{D}_g \quad (3.1)$$

The g th group consists of N_g individuals then, the population of the region is

$$N = \sum_{g=1}^M N_g \quad (3.2)$$

The spatial scale is determined by the number of groups. At a particular spatial scale, the region could be partitioned into areas in various ways. This process of forming groups at a particular scale is defined as zoning and scaling. The groups formed are not unique, hence the resulting aggregated data are not unique.

Consider the example of the Australian census data. The data collected refers to individuals. The aggregation of the data was based on the geographical classification provided by the Australian Standard Geographical Classification (ASGC), (McLennan, 1995). The ASGC divides Australia into Collection Districts (CDs), Statistical Local Areas (SLAs), Statistical Subdivisions (SSDs), Statistical Divisions (SDs), and States or Territories. These are examples of different scales. A comprehensive discussion of the ASGC can be found in McLennan (1995).

3.2.2 Aggregation of census data

The census information is presented in aggregate data form, where the aggregation is done at a particular scale and for a particular zoning arrangement. For example, the lowest scale and zoning arrangement in the Australian census consists of Collection Districts (CDs) which contain approximately two hundred households (Castles, 1991). This aggregate data could be considered as a representation of the area's value.

Based on the superpopulation approach, Cressie (1991, 1997) defined aggregation as an integral of a random process over the areal unit \mathcal{D}_g , which has area $|\mathcal{D}_g|$, for all location ℓ which are defined in \mathcal{D}_g . This definition is appropriate for a continuous spatial process. Ripley's (1981) definition for a discontinuous spatial process of the group mean is,

$$\bar{Y}_g = \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} Y_i \quad (3.3)$$

In the Australian census, the lowest level spatial unit for which data are made available is a Collection District (CD). Figure (3.1) illustrates the aggregation process of the unit level data into CD level data for a region that is divided into five CDs. The statistics produced for each CD are a result of a transformation from unit level data. The statistics can be considered as spatial data at the CD level, and involve aggregation of both the attribute and spatial characteristics. The individual locations are represented by a representative point, that is a centroid of the CD. The location of the centroid is determined by an interpolation procedure using the CD boundary coordinates, which is described in Griffith and Amrhein (1991) and is not necessarily the average of the spatial locations of the individuals in the CD.

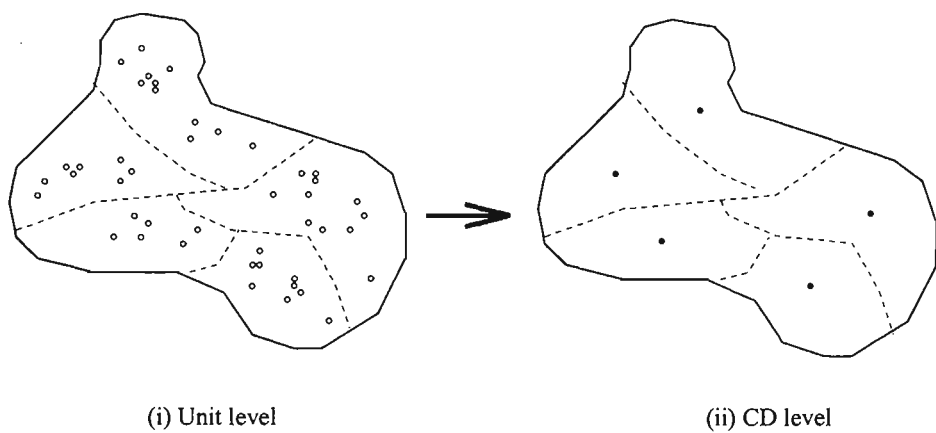


Figure 3.1. Aggregation process of unit level into CD level

3.3 Non-spatial solutions of aggregation problems

Understanding of the MAUP and ecological fallacy are key points in considering the analysis of aggregated data. This section provide an overview of some suggested solutions of these aggregation problems. The suggested solutions have come from a non-spatial perspective.

3.3.1 The MAUP solutions

Steel and Holt (1996b) derived the effect of aggregation on some common statistics, such as means, variances, regression, and correlation coefficients when groups are randomly formed. Aggregation can affect the expectation or variance of these statistics, or both. They showed that in general both the expectation and variance of the statistics will depend on four factors; the number of groups, group sizes, spatial pattern of the association between individuals, and boundary definition of the groups. The scale and zoning effects, may be equated with the terms expectation and variance of the statistics, respectively.

Amrhein (1995) performed a simulation to capture the aggregation effect. He found that the aggregation effect could be identified by the standard deviation of a statistic over repeated trials of the groups formation process at the same spatial scale. This empirical result is in agreement with the finding of Steel and Holt (1996b) that variances of statistics may contain a key indicator of the aggregation effect.

3.3.2 The ecological fallacy solutions

Ecological inference is the process of using aggregate data to infer individual level relationships when the individual level data are not available. The basic problem of ecological inference is the ecological fallacy, as defined by Robinson (1950). King (1997) proposed a method of ecological inference which is based on an extension of linear regression. He claimed that the proposed method could effectively minimize the bias associated with ecological inference.

Tranmer and Steel (1998) noted that within-area homogeneity is a key factor in the ecological fallacy. Holt et al. (1996) defined this phenomenon as positive intra-cluster correlation. Steel et al. (1996) proposed a solution of this problem by incorporating population structure into the statistical model underpinning the analysis. The population structure could be determined by identifying the *grouping variables* that characterize the process to explain the within group homogeneity. The authors noted that if appropriate grouping variables can be identified then the methods can provide unbiased estimates of individual level parameters from aggregated data, hence the ecological fallacy could be avoided. This method requires individual level data for the grouping variables.

3.4 Spatial analysis

Spatial analysis developed in several steps, those are point processes, geostatistics, and then spatial autocorrelation (Griffith, 1988). The point process approach is concerned with the study of point/location of the observations in the study area, and features such as density, and nearest neighbor distance. Geostatistics initially developed as an attempt to use classical univariate and bivariate statistical methods as a model for spatial analysis. Spatial autocorrelation reflected the existence of systematic spatial variation in the characteristic under study. Cliff and Ord (1981) formulated the spatial autocorrelation in the model (3.4). Griffith (1988) described three type of statistics for measuring spatial autocorrelation, these are the Moran coefficient, Geary ratio, and Cliff-Ord statistics.

3.4.1 Target of inference : spatial variability

How can we investigate spatial variability in the data ? This question leads to the idea of spatial autocorrelation. For example, a positive spatial autocorrelation indicates a small spatial variability among observations closer together and larger variability when observations are farther apart (Littel, Milliken, Stroup, & Wolfinger, 1996).

In many cases some sort of relationship might exist between measurements taken on objects geographically distributed over a surface and the underlying configuration of these objects (Griffith, 1988). This type of relationship arises in many physical phenomena, such as weather conditions, contents of ore body, contents of oil, etc. Social phenomena may also exhibit spatial autocorrelation, for instance in the study of epidemics of diseases, study of housing prices, study of consumer preference (marketing research), pollution, mortality rates, etc.

The existence of spatial variability in social characteristics could be viewed as a consequence of the inter-relationship mechanism among individuals within the population region. Consider two points are located at ℓ_i and ℓ_j with the observed values are (x_i, y_i) and (x_j, y_j) , respectively. Figure (3.2) shows a diagram of the possible relationship within and between social characteristics. The diagram indicates that the value y_i is not only affected by the presence of x_i and y_j , but also by x_j (Whittaker, 1990). The presence and the effect of (x_j, y_j) can be observed by introducing spatial perspective into the analysis.

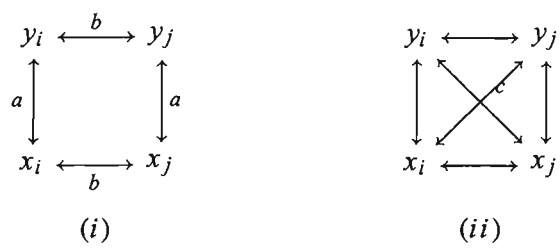


Figure 3.2. Relationship of variables; (a) relationship between variables ignoring spatial inter-dependence, (b) spatial inter-dependence in variable, and (c) relationship between variables considering spatial inter-dependence.

The first situation (a) indicates a relationship between variables (e.g. dependent and explanatory variables). The second situation (b) shows inter-dependence for a particular variable at different locations. The third situation (c), illustrates inter-dependence between different variables for different individuals.

The independent and identical distribution (IID) assumptions will ignore the second and third situations. Hence analyzing social data with classical statistical methods could give biased or inefficient estimation. Applying a spatial data analysis procedure is an approach to account for these relationships.

3.4.2 Spatial analysis : analysis of spatial variability

There are several approaches to the spatial analysis of spatial data. These approaches introduce a model and/or statistic that will take account of the presence of spatial dependence.

Spatial variability affects analyses, such as correlation, regression, or linear models. Measuring the spatial variability can give a description of the population in a spatial dimension. The description of spatial variability, may be for analytical and descriptive usages. This thesis will focus on the analytical usages.

Statistics have been developed to measure spatial dependence. Those are categorized into global and local spatial statistics. The first category measures spatial dependence based on simultaneous measurement from all observed locations. The latter measures spatial dependence using only a portion of the study area or up to a specified distance, such as distance statistics, $G_i(d)$ (Getis & Ord, 1992).

Global statistics include joint-count, Moran (I), and Geary (C) (Cliff & Ord, 1981). Extension of Moran and Geary ratios are the correlogram and variogram, respectively (Getis & Ord, 1996). The variogram is a statistic based upon the squared differences $(Y_i - Y_j)^2$, which can represent interdependence between observations as a function of distance.

O'Brien (1990) identified three important aspects in the analysis of spatial data; spatial unit problems, survey analysis problems, and measurement problems. The spatial unit problems are that a different result is obtained from data for different spatial units, leading to scale or aggregation effect (Wong, 1996). The survey analysis problems relate to the sampling design and analysis of the data. The design problem includes coverage and non-response problems, which can be minimized by good sampling design. The analysis problem is related to the violation of the IID assumption, on which classical inference methods are based. The measurement problems refer to the validity and quality of the data being taken.

3.4.3 Themes of spatial analysis

Guidelines for applying spatial analysis for social data in general can be found in Haining (1990). The variogram has been used extensively in geological science (see Cressie, 1991), but has also been applied for analysing social data, for instance Griffith et al. (1994) and Haslett (1997). Meanwhile Clifford et al. (1989) used the variogram when they explored the effect of spatial variation on bivariate correlation. Finally, spatial modeling was discussed in Wackernagel (1988) and Ver Hoef and Cressie (1993). Another important issue is the implications for sampling design, which has been mentioned in section (2.2.2).

Haining (1994) discussed some directions of spatial analysis, namely to accurately describe events in geographical space, to systematically explore patterns of events and associations between events in space, and lastly to improve the ability to predict events occurring in geographical space. These directions categorize spatial analysis into three main study areas, statistical spatial data analysis, map based analysis, and mathematical modeling.

There are several themes in the current research in spatial analysis of social data. Broadly they may be categorized into spatial autocorrelation, exploratory spatial data analysis, spatial multivariate modeling, and geographical information systems.

Spatial autocorrelation

Spatial autocorrelation reveals a feature of the association among individuals within the study region. It can have detrimental effect on the analysis of the data if it is ignored. It can often be represented as a function of the distance that separates the locations and the directions involved. Cliff and Ord (1981) introduced a spatial autocorrelation model,

$$Y_i = \rho \sum_{j \in \mathcal{U}} w_{ij} Y_j + \epsilon_i; \quad \text{for } i \in \mathcal{U} \quad (3.4)$$

The parameter ρ reflects the level of spatial autocorrelation among the (Y_i, Y_j) pairs for which $w_{ij} > 0$. The weight, w_{ij} , indicates the spatial connection between the pair of points (i, j) . The w_{ij} , could be expressed as a function of distance between the pair of points. The error terms of the model (ϵ_i) are independent and identically distributed. This model is also identified as a simultaneous autoregressive model and often used in spatial econometrics (Anselin, 1988). Another model is the conditional autoregressive

model, which consists of a linear relationship between the conditional expectation of the dependent variable and its values in the whole model (see Anselin, 1988, page 33). Hence the ordinary least squares method can be applied as an estimation techniques.

Spatial autocorrelation is represented by some statistics such as the Moran and Geary coefficient, distance statistics $G_i(d)$ (see, Cliff and Ord, 1981; and Getis and Ord, 1992). Initially spatial autocorrelation is used to measure spatial association globally across the whole study region, but it has been developed into a localized spatial association (Getis & Ord, 1996). Anselin (1995) refers to these statistics as members of a class of local indicators of spatial association (LISA). Ord and Getis (1995) discussed an extension of distance statistics $G_i(d)$ as local spatial autocorrelation statistics and focused on the distributional issues associated with these statistics. Ord and Getis (1995) discussed the issue of spatial autocorrelation and defined what they call local spatial autocorrelation statistics to describe a local pattern in spatial data. The method was implemented to study the AIDS epidemic centering on San Francisco.

Hagget et al. (1977) gave some examples of the effect of spatial autocorrelation on the results of t -test and regression analysis. They concluded that the presence of spatially autocorrelated error terms within a regression equation will affect the variances and produce a misleading significance test.

Exploratory spatial data analysis

Exploratory spatial data analysis techniques are popular techniques of handling spatial data in the fast growing development of computers and information technologies, i.e. geographical information systems. Wilhelm and Steck (1998) noted that exploratory spatial data analysis has several purposes including determining spatial structure, describing and visualizing geographical distribution, exploring spatial dependencies, measuring heterogeneity, and identifying outliers. *The Statistician*, Vol. 47, part 3, 1998 reported some proposed methods, such as interactive graphics procedures, or applying local spatial autocorrelation statistics (Wilhelm and Steck, 1998; Haining, Wise, and Ma, 1998).

Spatial multivariate modeling

The theme of spatial multivariate modeling has mainly appeared in physical applications, for example Wackernagel (1988). There have been some efforts to translate the methodology to spatial linear models, such as in Hepple (1996). Hepple (1996) discussed an application of spatial modeling in spatial econometrics. Spatial econometrics refers to a mix of spatial modeling and spatial analysis that is particularly concerned with modeling socio-economic relationship, using data for counties, states, or other region based data.

Wackernagel (1988) discussed an implementation of spatial analysis for interpreting multivariate spatial information. He presented some techniques of spatial analysis, which are based on a combination of variogram modeling, principal component analysis, and cokriging. Similar issues were also discussed in Ver Hoef and Cressie (1993). Multivariate spatial data could also be analysed by defining spatial linear models. These issues are discussed in Hepple (1996) and Christensen, Johnson, and Pearson (1993).

Anselin (1988) developed a framework for dealing with spatial variations in economics studies. Following Jean Paelinck's work, he popularized spatial econometrics as the appropriate framework (Anselin, 1992). He defined a framework, consisting of methods that appropriately deal with the special properties of spatial data and spatial models. This theme is a combination of spatial autocorrelation and spatial modeling, as discussed by Kelejian and Robinson (1992). They introduced a spatial autocorrelation test of a per capita county police expenditures, based on the error of regression model. Their finding showed how omitted explanatory variables may cause spatial autocorrelation in the error terms.

3.4.4 Application of spatial analysis and GIS

The increased interest in spatial statistics and their analysis is due to the revolution in computing and information technology. Developments in information technology are a driving force in spatial analysis, and include automatic data capture, image processing, and geographical information systems (Haslett, 1992).

The developments in information technology have resulted in a merging of spatial analysis and geographic information systems (Ding & Fotheringham, 1992). A geographic information systems (GIS)

may contain comprehensive tools for doing spatial analysis, containing inputting system, analysis, and outputting system (Burrough, 1986). The interface between GIS and spatial analysis has become a main topic of discussion in recent years, see, for example Anselin and Getis (1992).

Geographic information systems have been widely applied in physical and social sciences. Burrough (1986) discussed the use of GIS on the assessment of land use, while Martin (1996) mentioned some applications of GIS in socio-economics studies. Arbia (1993) and Bond and Devine (1991) discussed the role of GIS in statistical surveys and survey analysis. Pawitan (1993) did empirical work to construct an area frame, which is useful to draw a random sample in agriculture economics studies. The area frame was composed from different map boundaries: administrative boundaries and land use boundary.

Spatial analysis is one of the basic functions of GIS. There are efforts to include some spatial analysis techniques into GIS functionality. These efforts are very closely related to where a GIS is implemented. Related discussion of this matter can be found in Griffith (1993), Bailey (1994), or Anselin and Getis (1992).

Griffith (1993) mentioned the need to convert spatial statistics analysis into GIS functions and he also identified some gateways for spatial statistical tools to be introduced into a GIS. The same themes are found in Haining (1994). Bailey (1994) reviewed the potential for statistical spatial analysis in relation to a GIS and also discussed progress and benefits related to this analysis.

Flowerdew and Green (1992) noted that the key purpose of analysis in GIS is the integration of different data sets, and a problem for the integration of different data sets is diversity of spatial scales. Hence they emphasized the development of interpolation methods within the GIS functions.

Birkin, Clarke, Clarke, and Wilson (1990) discussed an application of model-based GIS for the evaluation of urban policy. They defined two comprehensive models of the urban and regional economy. The urban model was defined as a micro-simulation model, which is generated synthetically from a known aggregate distribution. The urban model is based on the interdependence between four different characteristics, those are population change, housing change, services, and employment change. This application is an example of geodemographic analysis.

Brown (1991) defined geodemographics as a method of the analysis of spatial aspects of the socio-economic structure of region (towns and cities). Flowerdew and Goldstein (1989) discussed a commercial application of geodemographic data and associated procedures of market analysis for companies in North America.

3.5 Spatial perspective in analysing aggregated data

This section will focus on some issues affecting spatial analysis of aggregated data. The zoning and spatial scale issues are the most cited issues. But recently, multivariate and multilevel spatial modeling issues appear in the discussion as well as spatial autocorrelation. This section highlights some issues in spatial analysis of aggregated data, that have been mentioned by some researchers.

The work of Steel et al. in exploring the effects of aggregation (see for example Steel and Holt, 1996a). Most of their work did not take account of spatial component in the analysis. A spatial perspective on analysing aggregated data has been mentioned in Amrhein and Reynolds (1996), Wong (1996), and Cressie (1996). One significance aspect of spatial perspective in the analysis is the presence of spatial variability in the data.

Fotheringham and Rogerson (1993) listed eight issues that arise in spatial analysis. The list included the MAUP, boundary problems, spatial interpolation, spatial sampling procedures, spatial autocorrelation, goodness-of-fit in spatial modeling, context-dependent results and non stationarity, aggregate versus disaggregate models. The issues relevant to this thesis are discussed here.

3.5.1 The MAUP and ecological fallacy from a spatial perspective

Openshaw (1978) noted that initially most of the analysis of spatial data was done in a non-spatial manner, but there is a significant effect of space in the analysis of spatially aggregated data. He defined the effect of space as the zoning and spatial scale effects. He argued that these problems can be controlled by defining an appropriated zone design procedure, and proposed an optimal zone design approach. He proposed developing the zone design procedure as an optimization problem in which model performance is traded

off against various zone design constraint. This approach is implemented in solving the automatic zoning problem.

Tobler (1989) argued that analysis of geographical data should not depend on the spatial aspect, which implies that the result showed be frame independent. He gave an argument that the problems, such as the MAUP, arise because the method of analysis used was inappropriate. An example, he cited the correlation coefficient, which is an inappropriate measure of association amongst spatial units, rather than the cross-coherence function. He suggested that the use of correct analysis procedures may eliminate the MAUP.

McCracken (1983) looked at the impact of the spatial frame on the result of statistical analysis of regional social well-being. Implicitly he defined the aggregated data by some spatial frame, his term for the zoning. He investigated the extent to which multivariate dimensions of regional social well-being are dependent on the spatial data frame employed and the degree to which postulated contributory processes of well-being levels are frame specific. He did empirical work to analyse aggregated data, which were generated by some spatial frame. He found that no one frame has any compelling a priori relevance to a particular topic. Hence he concluded that the issue of partitioning the region might be a main consideration in analysing aggregated data.

Curtis and MacPherson (1996) investigated the similar issues of zone definition problem in survey research. They stated that the results of a spatially structured survey of private companies can change significantly depending on the manner in which the region of the analysis is defined. They looked at the two aspects of spatial scale and spatial zoning at any one scale, which causes variation in the composition of the study region.

Arbia (1989a) discussed regional economics studies from a spatial perspective. He mentioned that in this case the observations are non-randomly generated and constituted by aggregation of the characteristics of individuals within a particular subregion. Furthermore, he observed aggregation effects on some different spatial scales.

The main point is that the MAUP and ecological fallacy could be overcome by incorporating spatial aspect into the analysis, as argued in Cressie (1996). Wong (1996) and Amrhein and Reynolds (1996) have presented empirical work that support this idea. The MAUP issues can be discouraging and encouraging.

They are discouraging in the sense that they make the result of any aggregate level analysis suspect and potentially unreliable. They are encouraging in that they present the challenge of reporting on the reliability of parameters estimates in particular scale and zoning systems.

Cressie (1996) noted the analogous term of the MAUP in Geology is the change of support problem. The available data are defined at a particular level of spatial unit support, and inference is limited to the level of aggregation obtained. Cressie (1996) argued that MAUP cannot be resolved until the spatial aspect is incorporated into the problem formulation, as in the change of support problem. He identified three important geographic operations, sub-setting, stratification, and aggregation. The available data may be obtained by a combination of the three operations. He suggested a model to overcome the change of support problem and the MAUP as well. The model is started by defining a conventional regression model, then the spatial characteristic is introduced into the model. Based on the model, the error term would exhibit spatial dependence. He concluded that the model could serve as a starting point to illustrate the modeling of stratification and aggregation operation. But the model does not provide a complete solution for all cases.

A relation between aggregation effects and spatial effects was observed by Amrhein and Reynolds (1996). They argued that the aggregation effect can be viewed as a contrast of two spatial processes within a data set in two different spatial scales, for example between individual level and group/aggregated level. They claimed that aggregation effect can be defined as a difference between variance of the aggregated data and variance of the original data or individual data.

Wong (1996) considered the aggregation effect in spatial data. He described the problems by comparing a variance covariance matrix of socio-economic characteristics calculated at different levels of spatial scale (county, town, census tract, and block group). The variance covariance matrices exhibit a different value for the different spatial scales.

Furthermore, Wong (1996) illustrated the impact of aggregation on bivariate and multivariate cases. In the bivariate case, an estimator of a bivariate regression tends to vary across different spatial scales. In the multivariate case, he argues that the effect of MAUP is unpredictable. That is, given the level of scale, a wide range of results can be developed from different spatial zoning schemes. He listed some

approaches to solve the MAUP, they include data manipulation approach, methods-oriented approach, and error modeling approach.

Arbia (1989b) discussed the process of aggregation as a spatial data transformation and the effect of this transformation on the statistical analysis. He cited that the MAUP might be the main problem and proposed a general framework by defining group-process probability distribution in terms of the individual-process probability distribution. He claimed that this framework may give an alternative approach to understand the MAUP.

3.5.2 Spatial interpolation

Spatial interpolation indicates a method to predict unknown values from the observed values. Flowerdew and Green (1992, 1994) discussed developments and methods of areal interpolation. They introduced areal interpolation method based on the EM algorithm. Spatial interpolation may have an important role in data integration of different sources or different zonal systems. Flowerdew and Green (1989) discussed the role of areal interpolation in data integration.

Recent application of spatial analysis of aggregated data can be found in Müller et al. (1997). They discussed a construction of incidence maps of AIDS data in San Francisco by spatial smoothing of geographically aggregated data, which involved the issue of interpolation using aggregated data.

The MAUP causes problems in spatial data integration. Spatial data integration is a fundamental function when merging spatial analysis and geographical information system. Data integration might provide a systematic transformation procedure of data at different spatial scale (point, line, areal, or surface). Data integration is usually defined in terms of areal interpolation, see, Flowerdew and Green (1989,1991,1992).

3.5.3 Boundary problem issues

Boundary problems are related to the shapes of the region and are particularly prominent in studies of spatial point patterns. The shape of the region may influence statistical measurement, such as nearest-neighbor statistics. Ripley (1979,1981) gave an approach to correct the boundary problems. Wong and

Fotheringham (1990) introduced a buffer zone around the study region to eliminate the edge or boundary effect.

3.5.4 Variogram

Spatial variability can be represented in terms of spatial autocorrelation. The variogram has been introduced to measure spatial autocorrelation, beside other statistics such as distance statistics, Moran coefficient, and Geary ratio. These statistics may have an important role in studying the correlation structure of the observations, since it may affect bivariate or multivariate correlation and spatial modeling.

The variogram is an important tool representing spatial variability. Cressie (1989) gave a brief discussion of geostatistics, in which he considered the linear methods of spatial prediction (kriging). This approach is defined on the basis of a variance of difference of the observations as a function of distance, that is a variogram model, from which the predictions are made.

Robinson (1990) discussed the role of the variogram in time series analysis. He showed a connection between the variogram and autocovariance function, the latter being more common in time series analysis. He noted that when the variance of the process is not well known then the variogram can be estimated with moderate precision but the autocovariance function cannot. With a simulation he showed empirically the situation when the variogram is more suitable than the autocovariance function. Haslett (1997) looked at the sample variogram for non-stationary process, while Watson (1997) evaluated variogram estimators under normal conditions when outliers are present in data.

Griffith et al. (1994) used the variogram in analysing social data. They considered that the data are a result of a random process in two dimensional space. The characteristics are observed for individuals who were distributed over a geographic region. The standard error of the estimate of the mean of the household income is examined when spatial autocorrelation was present. The data was considered at the census tract scale of Syracuse, New York. They implied that sample mean may not be the best estimator for either the tract mean or population mean in this situation. Instead they suggested that looking at the spatial autocorrelation structure of the characteristic will give a better estimator. They compared the estimators by

looking at their standard error and proposed an approach to use the variogram function to give an estimate of the standard error of the estimate of the mean.

Clifford et al. (1989) looked at the effect of spatial variability using the relationship between lung cancers, smoking, and industrial factors. They investigated the effect of spatial autocorrelation on the significance test of the correlation coefficient. The covariance structure of each variable was examined, and they developed an adjustment for a significance test of a correlation coefficient. The covariance structure was developed by taking account of spatial autocorrelation, and a variogram model was used to derive the autocorrelation function for each variable.

3.6 Simulation

Some sources of spatial data have been mentioned in section (2.1.3). But sometimes, we need to clarify the properties of proposed methods with the data, for which all parameters are known and this can be done using simulated data. Some simulation procedures have been discussed in Haining, Griffith, and Bennet (1983) and Goodchild (1980).

Haining et al. (1983) presented a framework for the generation of surfaces that possess the property of spatial autocorrelation. They stated two objectives of this simulation, the first is to generate spatial data with known, specific, and limited characteristics in order to investigate the properties of estimators and hypothesis test for spatial data. The second is to obtain realization of a spatial process in order to identify properties of the process.

Goodchild (1980) proposed an algorithm to generate data to be considered at aggregated level, which took account of the spatial autocorrelation factor. He recognized the importance of the spatial autocorrelation of parameters between groups in controlling the severity of aggregation effects, and noted that if neighboring groups have a similarity then the aggregation effect will be relatively weak.

3.7 Summary

The spatial perspective in the analysis of social data has been developed almost more than two decades ago. Cliff, Hagget, Ord, Basset, and Davies (1975) and Cliff and Ord (1981) gave a foundation of spatial analy-

sis, in particular for the spatial autocorrelation of social characteristics. Anselin (1988) developed a Spatial Econometrics, which took a spatial perspective in the modeling of social and economics phenomena. Then Haining (1990) gave a further introduction of spatial analysis for general social data.

The aggregation effect was initially recognized in a non-spatial perspective, such as reported and shown by Openshaw (1978). Many researcher tried to give a solution, such as proposed by Steel and Holt (1996a). The spatial perspective for this problem was discussed empirically by McCracken (1983), and a simulation study by Amrhein and Reynolds (1996). Theoretically it may be found in Arbia (1989a), Cressie (1996).

Chapter 4

Some Theory of Spatial Aggregation

It is often thought that aggregated data analysis will allow inferences at the same level as the aggregate data are available. When an objective is to make inference at a different level, then the aggregated data has to be used with extreme care (Tranmer & Steel, 1998; Holt et al., 1996; Openshaw & Taylor, 1979). The well-known problems associated with the aggregated data analysis are ecological fallacy and MAUP (see section 2.2.1) . We will show that even for analysis and inference at the aggregate level, the analysis has to be tested with caution, as the results are affected by relationships at several levels and between different individuals.

Holt et al. (1996) discussed non-spatial models to understand the problems and showed how the effect of aggregation can be related to within group homogeneity or correlation. However, this result is based on a statistical model that assumes no correlation between individuals in different groups and constant correlation between individuals in the same groups. In this chapter we will look at the problem with a more explicitly spatial approach and consider the implications of aggregation on the basic statistics of variances and covariances from which other common statistics, such as regression and correlation coefficients are calculated.

In this chapter it is shown how the more general patterns of correlation between individuals can be used to consider the effect of aggregation. The difference between the statistics calculated at group level and individual level will be examined, which will indicate the aggregation effect.

4.1 Aggregation and simple statistics

Section (2.1) described the population \mathcal{U} , that contains N individual units within a particular geographic region, \mathcal{D} , with boundary \mathcal{B} . Then the population consist of the random variables

$$(\mathbf{Y}, \mathbf{L}) \quad (4.1)$$

where $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_p]^\top$ and $\mathbf{L} = [\ell_1, \dots, \ell_N]^\top$ contain the population values of the attributes and locations respectively. The vector \mathbf{Y}_k represents p social attributes, such as income, sex, age, marital status, job status, nature of occupancy, etc. The ℓ_i denotes the location of the i 'th individual. In general $\mathbf{Y}_k \in \mathcal{R}^p$ and $\ell_i \in \mathcal{D} \subseteq \mathcal{R}^2$ or \mathcal{R}^3 . Notice that the two different individuals may be located at the same location, so that physically the distance between the two is zero, for example two people in the same household. This case correspond to $(\mathbf{Y}_k, \ell_i), (\mathbf{Y}_k, \ell_j)$ where $i \neq j$ but $\ell_i = \ell_j$. Another case may also represents the same locations, such observations from a group of households.

In general we will assume that \mathbf{Y}_k and \mathbf{L} are realization of a joint random process. This concept is recognized as the superpopulation approach, that the population is a random set (Cressie,1991; and Ripley,1981). We will mainly be interested in analysis conditional on the locations \mathbf{L} . In some cases there may be interest in the process that generated the locations, however we confine ourselves to analyse that are focused on the distribution and relationship between attributes.

Commonly used statistics such as correlation and regression coefficients are calculated from the variance and covariance of different variables. In this chapter we consider the univariate case, i.e. $p = 1$ such that $\mathbf{Y}_1 \in \mathbf{Y} = \{Y_1, \dots, Y_N\}$. For simplicity, the vector \mathbf{Y}_1 will be denoted by the scalar Y_i for $i = 1, \dots, N$. We assume that there exist a first and second moment structure for the conditional distribution of $Y_i|\mathbf{L}$ of the form;

$$\begin{aligned} (i) \quad E(Y_i|\mathbf{L}) &= \mu_i(\mathbf{L}) \\ (ii) \quad Cov(Y_i; Y_j|\mathbf{L}) &= \Delta_{ij}(\mathbf{L}) \end{aligned} \quad (4.2)$$

In case of $i = j$, we denote Δ_{ii} as Σ_i . In this section, the spatial location is only used to define which group an individual is in. More extensive use of the spatial locations in the modeling and analysis is considered in chapter (5).

At this stage the model allows the mean and covariances to depend on the locations of all individuals in the population. The model may be simplified to assume that $\mu_i(\mathbf{L})$ depends only on ℓ_i and the individual concerned and also that $\Delta_{ij}(\mathbf{L})$ depends only on ℓ_i and ℓ_j and the two individuals concerned. This implies $\mu_i(\mathbf{L}) = \mu_i(\ell_i)$ and $\Delta_{ij}(\mathbf{L}) = \Delta_{ij}(\ell_i, \ell_j)$. A further assumption is that $\Delta_{ij}(\ell_i, \ell_j)$ depends only on the distance $d_{ij} = \|\ell_i - \ell_j\|$ between the two individuals. But this assumption should aware with some restriction, such as physical barrier (coastal area, river, mountains, etc). We do not make such assumptions at this stage. For convenience we will denote $\mu_i(\mathbf{L})$ and $\Delta_{ij}(\mathbf{L})$ by μ_i and Δ_{ij} . At this point, we have not stated any assumptions regarding the spatial location, but these assumptions will be stated later in chapter 5 and the rest of the thesis. In census data, the individuals' locations are not recorded, although the location of the group's centroid may be.

The population mean of the attribute can be used to give a description or summary of the population. We can define the population mean as;

$$\bar{Y} = \frac{1}{N} \sum_{i \in \mathcal{U}} Y_i \quad (4.3)$$

Theorem 4.1.1. *The first and second moment of the $\bar{Y}|\mathbf{L}$ are*

$$\begin{aligned} (i) \quad E(\bar{Y}|\mathbf{L}) &= \bar{\mu} \\ (ii) \quad V(\bar{Y}|\mathbf{L}) &= \frac{1}{N}(\bar{\Sigma} + (N-1)\bar{\Delta}) \end{aligned} \quad (4.4)$$

where

$$\bar{\mu} = \frac{1}{N} \sum_{i \in \mathcal{U}} \mu_i \quad ; \quad \bar{\Sigma} = \frac{1}{N} \sum_{i \in \mathcal{U}} \Sigma_i \quad ; \quad \text{and} \quad \bar{\Delta} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \Delta_{ij} \quad (4.5)$$

Proof.

$$\begin{aligned}
 (i) \quad E(\bar{Y}|\mathbf{L}) &= E\left(\frac{1}{N} \sum_{i \in \mathcal{U}} Y_i | \mathbf{L}\right) = \frac{1}{N} \sum_{i \in \mathcal{U}} E(Y_i | \mathbf{L}) = \frac{1}{N} \sum_{i \in \mathcal{U}} \mu_i = \bar{\mu} \\
 (ii) \quad V(\bar{Y}|\mathbf{L}) &= V\left(\frac{1}{N} \sum_{i \in \mathcal{U}} Y_i | \mathbf{L}\right) \\
 &= \frac{1}{N^2} \left(\sum_{i \in \mathcal{U}} V(Y_i | \mathbf{L}) + \sum_{i \neq j \in \mathcal{U}} Cov(Y_i; Y_j | \mathbf{L}) \right) \\
 &= \frac{1}{N^2} \left(\sum_{i \in \mathcal{U}} \Sigma_i + \sum_{i \neq j \in \mathcal{U}} \Delta_{ij} \right) \\
 &= \frac{1}{N} (\bar{\Sigma} + (N-1)\bar{\Delta})
 \end{aligned}$$

□

The parameter $\bar{\Sigma}$ is the average of the population variances of the individuals and $\bar{\Delta}$ is the average of the population covariances between different individuals.

Suppose that the geographic region \mathcal{D} is partitioned into M non-overlapping subareas, says \mathcal{D}_g for $g = \{1, \dots, h, \dots, M\}$, and \mathcal{D}_g has boundary \mathcal{B}_g , such that ;

$$\bigcup_g \mathcal{D}_g = \mathcal{D} \subseteq \mathcal{R}^d \quad (4.6)$$

where \mathcal{R}^d is a random variables space of d dimension. In this thesis, it will be considered $d = 2$.

The areal unit \mathcal{D}_g can be defined following administrative boundaries or natural features, e.g. rivers, mountains, etc. In the Australian census we will focus on the case when the \mathcal{D}_g are collection districts.

Partitioning the region corresponds to grouping individuals into M groups. An individual belongs to one and only one group as determined by its location. The individual i is an element of \mathcal{D}_g if and only if $\ell_i \in \mathcal{D}_g$.

The data available from the census, and other sources often consist of summary data for each group or subarea. This process will give group level statistics. A common summary statistic is the group level mean, defined as;

$$\bar{Y}_g = \frac{1}{|\mathcal{D}_g|} \sum_{i \in \mathcal{U}_g} Y_i \quad (4.7)$$

here $|\mathcal{D}_g|$ is number of individuals within subarea \mathcal{D}_g , that is N_g .

Theorem 4.1.2. *The first and second moment of the $\bar{Y}_g|\mathbf{L}$ are*

$$\begin{aligned} (i) \quad E(\bar{Y}_g|\mathbf{L}) &= \bar{\mu}_g \\ (ii) \quad V(\bar{Y}_g|\mathbf{L}) &= \frac{1}{N_g}(\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g) \\ (iii) \quad \text{Cov}(\bar{Y}_g; \bar{Y}_h|\mathbf{L}) &= \bar{\Delta}_{gh} \end{aligned} \quad (4.8)$$

where :

$$\begin{aligned} \bar{\mu}_g &= \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} \mu_i & \bar{\Sigma}_g &= \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} \Sigma_i \\ \bar{\Delta}_g &= \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \Delta_{ij} & \bar{\Delta}_{gh} &= \frac{1}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \Delta_{ij} \end{aligned}$$

Proof. Apply theorem (4.1.1) within group g . □

The parameter $\bar{\Sigma}_g$ is the average of population variances of individuals in group g and $\bar{\Delta}_g$ is the average of the population covariances between different individuals within group g . The parameter $\bar{\Delta}_{gh}$ is the average of population covariances between pairs of individuals, with one member of the pair in group g and the other in group h . From now on all expectation and variances will be conditional on the locations, unless indicated otherwise. From the group level data, an estimator of the variance can be written in a quadratic form, that is;

$$\bar{S}_{yy} = \bar{\mathbf{Y}}^T \mathbf{A} \bar{\mathbf{Y}} = \sum_{g,h} \bar{Y}_g a_{gh} \bar{Y}_h \quad (4.9)$$

where $\bar{\mathbf{Y}}^T = [\bar{Y}_1, \dots, \bar{Y}_M]$ is the vector of group means and \mathbf{A} is a symmetric matrix with M by M dimension with element a_{gh} . An unweighted variance is defined by putting the element of \mathbf{A} as;

$$a_{gg} = \frac{1}{M}; \text{ and } a_{gh} = \frac{-1}{M(M-1)} \quad (4.10)$$

giving,

$${}_1\bar{S}_{yy} = \frac{1}{M-1} \sum_g (\bar{Y}_g - {}_1\bar{Y})^2 \quad (4.11)$$

where

$${}_1\bar{Y} = \frac{1}{M} \sum_g \bar{Y}_g$$

is the unweighted average of the group means.

A weighted variance is determined by defining the element of \mathbf{A} as ;

$$a_{gg} = \frac{N_g(1 - \frac{N_g}{N})}{M-1}; \text{ and } a_{gh} = \frac{-N_g N_h}{N(M-1)} \quad (4.12)$$

giving,

$${}_N\bar{S}_{yy} = \frac{1}{M-1} \sum_g N_g (\bar{Y}_g - \bar{Y})^2 \quad (4.13)$$

Notice that the individual level population mean,

$$\bar{Y} = \frac{1}{N} \sum_{i \in \mathcal{U}} Y_i = \frac{1}{N} \sum_g N_g \bar{Y}_g$$

is also a weighted average of the group means, with weights equal to the group population sizes.

Lemma 4.1.3. *Expectation of the quadratic form of the variance (equation 4.9) is*

$$E(\bar{S}_{yy}) = \sum_g a_{gg} V(\bar{Y}_g | \mathbf{L}) + \sum_{g \neq h} a_{gh} \text{Cov}(\bar{Y}_g; \bar{Y}_h | \mathbf{L}) + \bar{S}_{\mu\mu} \quad (4.14)$$

where $\bar{S}_{\mu\mu}$ is defined as;

$$\bar{S}_{\mu\mu} = \sum_g a_{gg} \bar{\mu}_g^2 + \sum_{g \neq h} a_{gh} \bar{\mu}_g \bar{\mu}_h = \bar{\mu}^\top \mathbf{A} \bar{\mu}$$

with $\bar{\mu}^\top = [\bar{\mu}_1, \dots, \bar{\mu}_M]$.

Proof.

$$\begin{aligned} E(\bar{S}_{yy}) &= E(\bar{\mathbf{Y}}^\top \mathbf{A} \bar{\mathbf{Y}}) = E\left(\sum_g a_{gg} (\bar{Y}_g)^2 + \sum_{g \neq h} a_{gh} \bar{Y}_g \bar{Y}_h\right) \\ &= \sum_g a_{gg} E(\bar{Y}_g^2) + \sum_{g \neq h} a_{gh} E[\bar{Y}_g \bar{Y}_h] \\ &= \sum_g a_{gg} [V(\bar{Y}_g) + \bar{\mu}_g^2] + \sum_{g \neq h} a_{gh} [\text{Cov}(\bar{Y}_g; \bar{Y}_h) + \bar{\mu}_g \bar{\mu}_h] \\ &= \sum_g a_{gg} V(\bar{Y}_g) + \sum_{g \neq h} a_{gh} \text{Cov}(\bar{Y}_g; \bar{Y}_h) + \bar{S}_{\mu\mu} \end{aligned}$$

□

Putting appropriate value of a_{gg} and a_{gh} into lemma (4.1.3) gives;

Theorem 4.1.4. *Expectation of the unweighted group level variance is*

$$E(\bar{S}_{yy}) = \frac{1}{M} \left(\sum_g \frac{1}{N_g} (\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g) - \sum_{g \neq h} \frac{\bar{\Delta}_{gh}}{M-1} \right) + \bar{S}_{\mu\mu} \quad (4.15)$$

Theorem 4.1.5. *Expectation of the weighted group level variance is*

$$E({}_N\bar{S}_{yy}) = \frac{1}{M-1} \left(\sum_g (1 - \frac{N_g}{N}) (\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g) - \sum_{g \neq h} \frac{N_g N_h}{N} \bar{\Delta}_{gh} \right) + {}_N\bar{S}_{\mu\mu} \quad (4.16)$$

In the same way, the variance of the individual level data can be written as a quadratic form, that is

$$S_{yy} = \frac{1}{N-1} \sum_{i \in \mathcal{U}} (Y_i - \bar{Y})^2 = \mathbf{Y}^T \mathbf{A} \mathbf{Y} \quad (4.17)$$

where $\mathbf{Y}^T = [Y_1, \dots, Y_N]$ and \mathbf{A} is an $N \times N$ matrix with element $a_{ii} = \frac{1}{N}$ and $a_{ij} = \frac{-1}{N(N-1)}$.

Theorem 4.1.6. *Expectation of the individual level variance is*

$$E(S_{yy}) = \bar{\Sigma} - \bar{\Delta} + S_{\mu\mu} \quad (4.18)$$

Proof. Applying lemma 4.1.3 to the individual level data gives

$$\begin{aligned} E(S_{yy}) &= \sum_{i \in \mathcal{U}} a_{ii} V(Y_i) + \sum_{i \neq j \in \mathcal{U}} a_{ij} \text{Cov}(Y_i; Y_j) + S_{\mu\mu} \\ &= \sum_{i \in \mathcal{U}} \frac{1}{N} \Sigma_i + \sum_{i \neq j \in \mathcal{U}} \frac{-1}{N(N-1)} \Delta_{ij} + S_{\mu\mu} \\ &= \bar{\Sigma} - \bar{\Delta} + S_{\mu\mu} \end{aligned}$$

□

The expectation of the individual level variance has three components, $\bar{\Sigma}$, $\bar{\Delta}$, and $S_{\mu\mu}$, coming from the variances, covariances, and means of the individuals. The IID assumptions imply that the $\bar{\Delta}$ and $S_{\mu\mu}$ are equal to zero, and then S_{yy} is unbiased estimator of $\bar{\Sigma}$. In this case $\Sigma_i = \Sigma$ for all $i \in \mathcal{U}$ and thus $\bar{\Sigma}_g = \bar{\Sigma} = \Sigma$ and ${}_N\bar{S}_{yy}$ is also unbiased for $\bar{\Sigma}$. But in practice, observations are often not independently and identically distributed. The first condition indicates a correlation among observations and the second may indicate that mean is not constant. A non-constant mean will be exhibited by existence of a spatial trend.

Theorem (4.1.6) can be applied to determine the individual level variance for group (g).

Corollary 4.1.7. *The expectation of the individual level variance within the g 'th group is*

$$E(S_{yy}^{<g>}) = \bar{\Sigma}_g - \bar{\Delta}_g + S_{\mu\mu}^{<g>} \quad (4.19)$$

where $S_{yy}^{<g>}$ is

$$S_{yy}^{<g>} = \frac{1}{N_g - 1} \sum_{i \in \mathcal{U}_g} (Y_i - \bar{Y}_g)^2 \quad (4.20)$$

4.2 Aggregation effect on variance

The population mean is $\bar{Y} = \frac{1}{N} \sum_g N_g \bar{Y}_g$ and so, provided appropriate weights are used, aggregation does not affect the mean. However, in general there will be an effect on the variance.

The aggregation effect can be examined in terms of the difference between statistics calculated from a data set in two different scales, for example between individual level and group level (Steel et al., 1996). Steel and Holt (1996b) showed that the variance is a key indicator of the aggregation effect. Openshaw and Taylor (1979), Amrhein (1995) also examined the effect of aggregation by looking at variances. Another way of looking at the effect of aggregation is the ratio between the weighted group level variance and the individual level variance (Steel & Holt, 1996a). The first approach will be discussed in this section, and the second will be discussed in section (4.4) and (4.3.3) in this chapter.

In this section we investigate how the effect of aggregation can be related to the spatial structure of the population as represented by Δ_{ij} .

Theorem 4.2.1. *The individual level variance can be partitioned into*

$$(N - 1)S_{yy} = (M - 1)N\bar{S}_{yy} + \sum_g (N_g - 1)S_{yy}^{<g>} \quad (4.21)$$

Proof. This is the standard identity in the analysis of variance. The total sum of squares of the individual level data can be partitioned into

$$\sum_{i \in \mathcal{U}} (Y_i - \bar{Y})^2 = \sum_g N_g (\bar{Y}_g - \bar{Y})^2 + \sum_g \sum_{i \in \mathcal{U}_g} (Y_i - \bar{Y}_g)^2 \quad (4.22)$$

Substituting in equation (4.13), (4.17) and (4.20) complete the proof. \square

Theorem 4.2.2. *Expectation of the weighted variance is*

$$E(N\bar{S}_{yy}) = \frac{1}{M-1} \left((N-1)E(S_{yy}) - \sum_g (N_g - 1)(\bar{\Sigma}_g - \bar{\Delta}_g) \right) + N\bar{S}_{\mu\mu} - \frac{N-1}{M-1} S_{\mu\mu} \quad (4.23)$$

Proof. Take expectation in (4.21) and use (4.19) and (4.21) applied to μ_i . \square

Equation (4.21) can be rewritten into

$$(N-1)S_{yy} = (M-1)N\bar{S}_{yy} + (N-M)S_{yy}^{<W>} \quad (4.24)$$

where

$$S_{yy}^{<W>} = \frac{1}{N-M} \sum_g (N_g - 1)S_{yy}^{<g>} \quad (4.25)$$

is the average within group individual level variance. Rearranging gives

$$N\bar{S}_{yy} = \frac{(N-1)S_{yy} - (N-M)S_{yy}^{<W>}}{M-1} \quad (4.26)$$

Holt et al. (1996) noted that the value of the $N\bar{S}_{yy}$ is affected by the value of the $S_{yy}^{<W>}$. Two special cases are noted. The first is a situation when the groups are perfectly homogeneous such that $S_{yy}^{<W>} = 0$ in which case

$$N\bar{S}_{yy} = \frac{N-1}{M-1} S_{yy} \approx \bar{N} S_{yy}$$

The second situation is if area membership is determined randomly, in which case $S_{yy}^{<g>} \approx S_{yy}$ for all g and so $\bar{S}_{yy} \approx S_{yy}$. For a particular population of individuals, the weighted group level variance increases as the within group variance decreases. Thus the more homogeneous the groups are the larger is the effect of aggregation.

Define the aggregation effect as the difference between the group level variance and the individual level variance. In general the group level variance can be the unweighted or the weighted group level variance,

$$\text{Aggregation effect} = ({}_A\bar{S}_{yy} - S_{yy}) \quad (4.27)$$

where ${}_A\bar{S}_{yy}$ is the group level variance, with $A = 1$ if unweighted and $A = N$ if weighted. The expectation of (4.27) will be considered for a general spatial population.

To explore the aggregation effect, we will consider three components of covariance. These are average of individual covariance ($\bar{\Delta}$), average of between group covariance ($\bar{\Delta}_B$), and average of within group covariance ($\bar{\Delta}_W$). The parameters $\bar{\Delta}_B$ and $\bar{\Delta}_W$ are defined as follows,

$$\bar{\Delta}_B = \frac{\sum_{g \neq h} N_g N_h \bar{\Delta}_{gh}}{\sum_{g \neq h} N_g N_h} ; \text{ and } \bar{\Delta}_W = \frac{\sum_g N_g (N_g - 1) \bar{\Delta}_g}{\sum_g N_g (N_g - 1)} \quad (4.28)$$

We may also use the unweighted versions

$$\tilde{\Delta}_B = \frac{1}{M(M-1)} \sum_{g \neq h} \bar{\Delta}_{gh} ; \text{ and } \tilde{\Delta}_W = \frac{1}{M} \sum_g \bar{\Delta}_g \quad (4.29)$$

Theorem 4.2.3. *The average of the individual level covariances, $\bar{\Delta}$, can be decomposed into component of $\bar{\Delta}_W$ and $\bar{\Delta}_B$, that is*

$$\bar{\Delta} = \frac{1}{N-1} \left\{ [\bar{N}(1+C^2) - 1] \bar{\Delta}_W + [N - \bar{N}(1+C^2)] \bar{\Delta}_B \right\} \quad (4.30)$$

where C^2 is square of the coefficient of variation of the group size (N_g),

$$C^2 = \frac{1}{M} \frac{\sum_g (N_g - \bar{N})^2}{\bar{N}^2} , \text{ and } \bar{N} = \frac{N}{M} \quad (4.31)$$

where \bar{N} is the average group population size.

Proof. By definition of $\bar{\Delta}$,

$$\begin{aligned} N(N-1)\bar{\Delta} &= \sum_{i \neq j \in \mathcal{U}} \Delta_{ij} = \sum_g \sum_{i \neq j \in \mathcal{U}_g} \Delta_{ij} + \sum_{g \neq h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \Delta_{ij} \\ &= \sum_g N_g(N_g - 1) \bar{\Delta}_g + \sum_{g \neq h} N_g N_h \bar{\Delta}_{gh} \\ &= \sum_g N_g(N_g - 1) \bar{\Delta}_W + \sum_{g \neq h} N_g N_h \bar{\Delta}_B \end{aligned}$$

By definition of $\bar{\Delta}_W$, $\bar{\Delta}_B$. Moreover,

$$\sum_{g \neq h} N_g N_h = \left(\sum_g N_g \right)^2 - \sum_g N_g^2 = N^2 - \sum_g N_g^2 \quad (4.32)$$

and by definition of C^2 ,

$$\sum_g N_g^2 = \bar{N}^2 M(1 + C^2) \quad (4.33)$$

Therefore

$$N(N-1)\bar{\Delta} = \left(\sum_g N_g^2 - N \right) \bar{\Delta}_W + \left(N^2 - \sum_g N_g^2 \right) \bar{\Delta}_B$$

$$\bar{\Delta} = \frac{1}{N-1} \left\{ [\bar{N}(1+C^2) - 1] \bar{\Delta}_W + [N - \bar{N}(1+C^2)] \bar{\Delta}_B \right\}$$

□

Corollary 4.2.4. *If the N_g are constant at \bar{N} , then $\bar{\Delta}_W = \tilde{\Delta}_W$ and $\bar{\Delta}_B = \tilde{\Delta}_B$ and (4.30) can be expressed as*

$$\bar{\Delta} = \frac{\bar{N}-1}{N-1} \tilde{\Delta}_W + \frac{\bar{N}(M-1)}{N-1} \tilde{\Delta}_B \quad (4.34)$$

Proof. If $N_g = \bar{N}$ then $C = 0$ giving the result. □

Theorem 4.2.5. *Expectation of the unweighted group level variance is*

$$E({}_1\bar{S}_{yy}) = \frac{1}{M} \sum_g \frac{1}{N_g} (\bar{\Sigma}_g - \bar{\Delta}_g) + \tilde{\Delta}_W - \tilde{\Delta}_B + {}_1\bar{S}_{\mu\mu} \quad (4.35)$$

Proof. From (4.15)

$$\begin{aligned} E({}_1\bar{S}_{yy}) &= \frac{1}{M} \left(\sum_g \frac{1}{N_g} [\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g] - \sum_{g \neq h} \frac{\bar{\Delta}_{gh}}{M-1} \right) + {}_1\bar{S}_{\mu\mu} \\ &= \frac{1}{M} \sum_g \frac{1}{N_g} \bar{\Sigma}_g + \frac{1}{M} \sum_g \bar{\Delta}_g \left(1 - \frac{1}{N_g} \right) - \tilde{\Delta}_B + {}_1\bar{S}_{\mu\mu} \\ &= \frac{1}{M} \sum_g \frac{1}{N_g} \bar{\Sigma}_g - \frac{1}{M} \sum_g \frac{1}{N_g} \bar{\Delta}_g + \tilde{\Delta}_W - \tilde{\Delta}_B + {}_1\bar{S}_{\mu\mu} \end{aligned}$$

□

Theorem 4.2.6. *Expectation of the weighted group level variance is*

$$E({}_N\bar{S}_{yy}) = \tilde{\Sigma} \left(1 - \frac{1}{M} \bar{C}_{N\bar{\Sigma}} \right) - \frac{N-1}{M-1} \bar{\Delta} - \tilde{\Delta}_W \bar{N} \left(\frac{M}{M-1} \frac{\bar{N}-1}{\bar{N}} + \bar{C}_{N\bar{\Delta}} \right) \quad (4.36)$$

where $\bar{C}_{N\bar{\Sigma}}$ is a relative covariance of N_g and $\bar{\Sigma}_g$, and $\bar{C}_{N\bar{\Delta}}$ is a relative covariance of N_g and $\bar{\Delta}_g$.

Proof. Modify (4.16) gives

$$\begin{aligned}
 E(N\bar{S}_{yy}|\mathbf{L}) &= \frac{1}{M-1} \left(\sum_g \left[1 - \frac{N_g}{N} \right] (\bar{\Sigma}_g + (N_g - 1)\bar{\Delta}_g) - \sum_{g \neq h} \frac{N_g N_h}{N} \bar{\Delta}_{gh} \right) + N\bar{S}_{\mu\mu} \\
 &= \frac{1}{M-1} \sum_g \bar{\Sigma}_g - \frac{1}{M-1} \sum_g \frac{N_g \bar{\Sigma}_g}{N} + \sum_g \frac{(N_g - 1)\bar{\Delta}_g}{M-1} \\
 &\quad - \sum_g \frac{\bar{\Delta}_g (N_g - 1)N_g}{N(M-1)} - \sum_{g \neq h} \frac{N_g N_h}{N(M-1)} \bar{\Delta}_{gh} + N\bar{S}_{\mu\mu} \\
 &= \frac{1}{M-1} (M\tilde{\Sigma} - \bar{\Sigma}) + \frac{1}{M-1} \sum_g (N_g - 1)\bar{\Delta}_g \\
 &\quad - \frac{1}{M-1} \left(\sum_g \frac{N_g (N_g - 1)}{N} \bar{\Delta}_g + \sum_{g \neq h} \frac{N_g N_h}{N} \bar{\Delta}_{gh} \right) + N\bar{S}_{\mu\mu} \\
 &= \frac{1}{M-1} (M\tilde{\Sigma} - \bar{\Sigma}) + \frac{1}{M-1} \sum_g N_g \bar{\Delta}_g - \frac{M\tilde{\Delta}_W}{M-1} - \frac{N-1}{M-1} \bar{\Delta} + N\bar{S}_{\mu\mu}
 \end{aligned}$$

Define $\bar{S}_{N\bar{\Delta}}$ as the group level covariance between N_g and $\bar{\Delta}_g$,

$$\bar{S}_{N\bar{\Delta}} = \frac{1}{M-1} \sum_g (N_g - \bar{N}) \cdot (\bar{\Delta}_g - \bar{\Delta}_W) \quad (4.37)$$

which can be rearranged to give

$$\sum_g N_g \bar{\Delta}_g = (M-1)\bar{S}_{N\bar{\Delta}} + M\bar{N}\bar{\Delta}_W \quad (4.38)$$

In the same way, we define $\bar{S}_{N\bar{\Sigma}}$ as the group level covariance between N_g and $\bar{\Sigma}_g$,

$$\begin{aligned}
 \bar{S}_{N\bar{\Sigma}} &= \frac{1}{M-1} \sum_g (N_g - \bar{N}) \cdot (\bar{\Sigma}_g - \bar{\Sigma}) \\
 &= \frac{N}{M-1} (\bar{\Sigma} - \tilde{\Sigma})
 \end{aligned} \quad (4.39)$$

and it can be rearranged to give

$$M\tilde{\Sigma} = M\bar{\Sigma} - \frac{M(M-1)}{N} \bar{S}_{N\bar{\Sigma}} \quad (4.40)$$

Define the relative covariance between N_g and $\bar{\Sigma}_g$,

$$\bar{C}_{N\bar{\Sigma}} = \frac{\bar{S}_{N\bar{\Sigma}}}{\bar{N}\bar{\Sigma}} \quad (4.41)$$

or the $\bar{S}_{N\bar{\Sigma}}$ can be denoted as

$$\bar{S}_{N\bar{\Sigma}} = \bar{N} \cdot \bar{\Sigma} \cdot \bar{C}_{N\bar{\Sigma}} \quad (4.42)$$

and the relative covariance between N_g and $\bar{\Delta}_g$,

$$\bar{C}_{N\bar{\Delta}} = \frac{\bar{S}_{N\bar{\Delta}}}{\bar{N} \cdot \bar{\Delta}_W} \quad (4.43)$$

Substituting the (4.38) and (4.40) into the $E(N\bar{S}_{yy})$ gives

$$\begin{aligned} E(N\bar{S}_{yy}) &= \frac{(M\bar{\Sigma} - \bar{\Sigma})}{M-1} + \frac{1}{M-1} \left((M-1)\bar{S}_{N\bar{\Delta}} + M\bar{N}\bar{\Delta}_W \right) - \frac{M}{M-1}\bar{\Delta}_W - \frac{N-1}{M-1}\bar{\Delta} + N\bar{S}_{\mu\mu} \\ &= \bar{\Sigma} + \frac{1}{M-1}(\bar{\Sigma} - \bar{\Sigma}) - \frac{N-1}{M-1}\bar{\Delta} + \frac{M(\bar{N}-1)}{M-1}\bar{\Delta}_W + \bar{S}_{N\bar{\Delta}} + N\bar{S}_{\mu\mu} \end{aligned}$$

and substituting (4.39), (4.42) and (4.43) in and rearranging it completes the proof. \square

The $\bar{\Delta}$ is a parameter that does not depend on any grouping structure in the population, but $\bar{\Delta}_W$ depends on the grouping scheme. These result express the expectation of the unweighted or weighted group level variance in terms of components that can be used to investigate the effect of scaling and zoning. For example equation (4.36) represents the expectation of the weighted group level variance in terms of four components, which depend on the scaling and zoning, these are $\bar{\Sigma}$, \bar{N} , $\bar{\Delta}_W$, and $\bar{S}_{N\bar{\Delta}}$. We expect that a key term will be $\bar{\Delta}_W$, which is an average of the within group covariances and will be different for different groupings. The average number of individuals per group, \bar{N} , will also be a key factor and reflects the scale.

We can derive expectation of the aggregation effect of the weighted group level variances.

Theorem 4.2.7. *Expectation of the aggregation effect of the weighted group level variance is*

$$E(N\bar{S}_{yy} - S_{yy}) = -\bar{\Sigma}\bar{C}_{N\bar{\Sigma}} - \frac{M(\bar{N}-1)}{M-1}\bar{\Delta} + \bar{\Delta}_W \left(\frac{M(\bar{N}-1)}{M-1} + \bar{N}\bar{C}_{N\bar{\Delta}} \right) + N\bar{S}_{\mu\mu} - S_{\mu\mu} \quad (4.44)$$

Proof. Applying theorem 4.2.6, we have

$$E(N\bar{S}_{yy} - S_{yy}) = \bar{\Sigma} \left(1 - \frac{\bar{C}_{N\bar{\Sigma}}}{M} \right) - \bar{\Delta} - S_{\mu\mu} - \frac{N-1}{M-1}\bar{\Delta} - \bar{\Delta}_W\bar{N} \left(\frac{M}{M-1} \frac{\bar{N}-1}{\bar{N}} + \bar{C}_{N\bar{\Delta}} \right) + N\bar{S}_{\mu\mu}$$

Simplifying this completes the proof. \square

Corollary 4.2.8. *The aggregation effect can be formulated into*

$$\begin{aligned} E(N\bar{S}_{yy} - S_{yy}) &= \frac{M(\bar{N}-1)}{M-1}(\bar{\Delta}_W - \bar{\Delta}) + \bar{C}_N \left(\bar{\Delta}_W\bar{R}_{N\bar{\Delta}}\bar{C}_{\bar{\Delta}} - \bar{\Sigma}\bar{R}_{N\bar{\Sigma}}\bar{C}_{\bar{\Sigma}} \right) \\ &\quad + N\bar{S}_{\mu\mu} - S_{\mu\mu} \end{aligned} \quad (4.45)$$

where

$$\bar{R}_{N\bar{\Sigma}} = \frac{\bar{S}_{N\bar{\Sigma}}}{\bar{S}_N \bar{S}_{\bar{\Sigma}}}; \quad \text{and} \quad \bar{R}_{N\bar{\Delta}} = \frac{\bar{S}_{N\bar{\Delta}}}{\bar{S}_N \bar{S}_{\bar{\Delta}}}$$

are the correlation between N_g and $\bar{\Sigma}_g$ and $\bar{\Delta}_g$ respectively.

Proof. We have

$$\bar{C}_{N\bar{\Sigma}} = \bar{R}_{N\bar{\Sigma}} \bar{C}_N \bar{C}_{\bar{\Sigma}}; \quad \text{and} \quad \bar{C}_{N\bar{\Delta}} = \bar{R}_{N\bar{\Delta}} \bar{C}_N \bar{C}_{\bar{\Delta}}$$

where

$$\bar{C}_N = \frac{\bar{S}_N}{\bar{N}}; \quad \bar{C}_{\bar{\Sigma}} = \frac{\bar{S}_{\bar{\Sigma}}}{\bar{\Sigma}}; \quad \text{and} \quad \bar{C}_{\bar{\Delta}} = \frac{\bar{S}_{\bar{\Delta}}}{\bar{\Delta}_W}$$

Putting these equality into (4.44) complete the proof \square

This corollary may be useful in obtaining an idea of the size of the second term in (4.45) as \bar{C}_N can usually be calculated and will often be much less than 1. It may also be possible to obtain an idea of the other components factors. In many cases this term will be small. This corollary clearly shows the role the factors \bar{N} and $(\bar{\Delta}_W - \bar{\Delta})$ play in explaining the difference between ${}_N\bar{S}_{yy}$ and S_{yy} . In general we could expect there to be greater average correlation within groups than in the population as a whole, i.e. $\bar{\Delta}_W > \bar{\Delta}$. Hence the first term in (4.45) could usually contribute positively to the bias of ${}_N\bar{S}_{yy}$ as an estimator of S_{yy} .

We will consider how the aggregation effect on variance behaves in two different special cases. The first case is when the group size is constant, i.e. $N_g = \bar{N} = \frac{N}{M}$. The second case is when the group sizes are different but the average within group variance and covariance are constant.

4.2.1 Case 1 : constant group size

In this case, there is a simple relation between unweighted and weighted group level variance

$${}_N\bar{S}_{yy} = \frac{1}{M-1} \sum_g \bar{N} (\bar{Y}_g - \bar{Y})^2 = \bar{N} {}_1\bar{S}_{yy} \quad (4.46)$$

In this case the coefficient of variation of the group size (C^2) is zero, $\bar{S}_{N\bar{\Sigma}}$ and $\bar{S}_{N\bar{\Delta}}$ are zero, $\bar{\Sigma} = \bar{\Sigma}$, and

$$\bar{\Delta} = \frac{\bar{N}-1}{N-1} \bar{\Delta}_W + \frac{\bar{N}(M-1)}{N-1} \bar{\Delta}_B \quad (4.47)$$

Evaluating theorem (4.2.5), (4.2.6), and (4.2.7) with N_g constant gives the following two corollaries.

Corollary 4.2.9. *Expectation of the unweighted variance is*

$$E({}_1\bar{S}_{yy}) = \frac{1}{\bar{N}} \left\{ \bar{\Sigma} - \left(\frac{N-1}{M-1} \right) \bar{\Delta} + \left(\frac{N-M}{M-1} \right) \tilde{\Delta}_w \right\} + {}_1\bar{S}_{\mu\mu} \quad (4.48)$$

Corollary 4.2.10. *Expectation of the weighted variance is*

$$E({}_N\bar{S}_{yy}) = \bar{\Sigma} - \left(\frac{N-1}{M-1} \right) \bar{\Delta} + \left(\frac{N-M}{M-1} \right) \tilde{\Delta}_w + {}_N\bar{S}_{\mu\mu} \quad (4.49)$$

The only terms in corollary 4.2.10 that depend on the grouping are $\tilde{\Delta}_w$ and ${}_N\bar{S}_{\mu\mu}$.

Corollary 4.2.11. *Expectation of the aggregation effect of the unweighted group level variance is*

$$E({}_1\bar{S}_{yy} - S_{yy}) = -\frac{\bar{N}-1}{\bar{N}} \bar{\Sigma} - \left(\frac{\bar{N}-1}{\bar{N}} \right) \left(\frac{M}{M-1} \right) \left(\frac{\bar{\Delta}}{M} - \tilde{\Delta}_w \right) + ({}_1\bar{S}_{\mu\mu} - S_{\mu\mu}) \quad (4.50)$$

Equation (4.50) shows that the difference between the unweighted group level variance and the individual level variance depends on the parameters $\bar{\Sigma}$ and $(\frac{\bar{\Delta}}{M} - \tilde{\Delta}_w)$ with factors $\frac{\bar{N}-1}{\bar{N}}$ and $\frac{\bar{N}-1}{\bar{N}} \cdot \frac{M}{M-1}$ respectively. The IID assumption implies the $(\frac{\bar{\Delta}}{M} - \tilde{\Delta}_w)$ component is zero in which case the expectation of the difference is $-\bar{\Sigma}$ factor by $\frac{\bar{N}-1}{\bar{N}} \approx 1$ if \bar{N} is large. The fact that this estimator has a bias even when the IID assumption is valid, is a reason not to use it.

Corollary 4.2.12. *Expectation of the aggregation effect of the weighted group level variance is*

$$E({}_N\bar{S}_{yy} - S_{yy}) = -(\bar{N}-1) \frac{M}{M-1} (\bar{\Delta} - \tilde{\Delta}_w) + ({}_N\bar{S}_{\mu\mu} - S_{\mu\mu}) \quad (4.51)$$

Equation (4.51) shows that expectation of the difference between weighted group level variance and the individual level variance depends on the $(\bar{\Delta} - \tilde{\Delta}_w)$ component with a negative factor $(\bar{N}-1) \frac{M}{M-1} \approx \bar{N}-1$. The $\bar{\Delta}$ components is not affected by either the scale or zoning, but $\tilde{\Delta}_w$ will be. The case of IID data will give zero expectation, since the $(\bar{\Delta} - \tilde{\Delta}_w)$ component is zero.

4.2.2 Case 2 : Different group size but allowing a constant $\bar{\Sigma}_g$ and $\bar{\Delta}_g$

In this case we assume that $\bar{\Sigma}_g$ and $\bar{\Delta}_g$ are constant over the population, with values denoted by $\bar{\Sigma}'$ and $\bar{\Delta}'$ respectively. This condition implies,

$$\begin{aligned}\bar{\Sigma} &= \sum_g \frac{N_g}{N} \bar{\Sigma}_g = \bar{\Sigma}'; \quad \bar{\Sigma} = \sum_g \frac{\bar{\Sigma}_g}{M} = \bar{\Sigma}'; \\ \bar{\Delta}_W &= \sum_g \frac{\bar{\Delta}_g}{M} = \bar{\Delta}'; \quad \bar{S}_{N\bar{\Delta}} = 0; \quad \text{and} \quad \bar{S}_{N\bar{\Sigma}} = 0\end{aligned}\quad (4.52)$$

There is no assumption for the $\bar{\Delta}_{gh}$, therefore $\bar{\Delta}_B$ is defined as in equation (4.28).

Corollary 4.2.13. *The expectation of the unweighted group level variance is*

$$E({}_1\bar{S}_{yy}) = \bar{N}_{-1}(\bar{\Sigma} - \bar{\Delta}_W) + \frac{N-1}{N-\bar{N}}(\bar{\Delta}_W - \bar{\Delta}) + {}_1\bar{S}_{\mu\mu} \quad (4.53)$$

where

$$\bar{N}_{-1} = \frac{1}{M} \sum_g \frac{1}{N_g}$$

Proof. The proof is done by evaluating equation (4.35) with the value of $\bar{\Sigma}_g = \bar{\Sigma}'$ and $\bar{\Delta}_g = \bar{\Delta}'$. \square

Corollary 4.2.14. *expectation of the weighted variance is*

$$E({}_N\bar{S}_{yy}) = \bar{\Sigma} - \left(\frac{N-1}{M-1}\right) \bar{\Delta} + \left(\frac{N-M}{M-1}\right) \bar{\Delta}_W + {}_N\bar{S}_{\mu\mu} \quad (4.54)$$

Proof. The proof is done by evaluating equation (4.36) with the value of $\bar{\Sigma}_g = \bar{\Sigma}'$ and $\bar{\Delta}_g = \bar{\Delta}'$. \square

Expectation of the aggregation effect of the weighted variance can be derived easily from equation (4.54), that is

Corollary 4.2.15. *Expectation of the aggregation effect of the weighted variance is*

$$E({}_N\bar{S}_{yy} - \bar{S}_{yy}) = (\bar{N} - 1) \left(\frac{M}{M-1}\right) (\bar{\Delta}_W - \bar{\Delta}) + ({}_N\bar{S}_{\mu\mu} - S_{\mu\mu}) \quad (4.55)$$

Proof. The proof is done by modifying equation (4.54) and using (4.18), that is

$$E({}_N\bar{S}_{yy}) = \bar{\Sigma} - \bar{\Delta} + S_{\mu\mu} + \bar{\Delta} - \left(\frac{N-1}{M-1}\right) \bar{\Delta} + \left(\frac{N-M}{M-1}\right) \bar{\Delta}_W + {}_N\bar{S}_{\mu\mu} - S_{\mu\mu}$$

Simplifying this equation and taking the $E({}_N\bar{S}_{yy} - \bar{S}_{yy})$ completes the proof. \square

We may observe that equation (4.55) and (4.51) are same. This situation shows that constant N_g will give the same result as constant $\bar{\Sigma}_g$ and $\bar{\Delta}_g$.

4.3 Empirical perspective of squared differences of pairs of observations

In the previous section the effect of aggregation on the variance of a variable was theoretically investigated by considering the expectation of group and individual level variances under a simple general statistical model that allowed for covariances between different individuals. In this section the aggregation effect is related to empirical measures of the similarity of observations. The results obtained do not make any model assumptions at this stage and are purely algebraic. Later we will see how the relationships can be interpreted under some spatial models.

The empirical perspective is based on the value of squared differences of pairs of observations.

$$(Y_i - Y_j)^2 \quad (4.56)$$

Theorem 4.3.1. *The individual level variance S_{yy} can be represented in terms of the squared differences (4.56),*

$$S_{yy} = \frac{1}{N(N-1)} \sum_{i,j} \frac{1}{2} (Y_i - Y_j)^2 \quad (4.57)$$

Proof.

$$\begin{aligned} \sum_{i,j} \frac{1}{2} (Y_i - Y_j)^2 &= \sum_{i,j} \frac{1}{2} (Y_i^2 + Y_j^2 - 2Y_i Y_j) \\ &= N \sum_{i \in \mathcal{U}} Y_i^2 - \left(\sum_{i \in \mathcal{U}} Y_i \right)^2 \end{aligned}$$

Hence

$$\begin{aligned} \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{U}} \frac{1}{2} (Y_i - Y_j)^2 &= \frac{1}{N-1} \left[\sum_{i \in \mathcal{U}} Y_i^2 - \frac{1}{N} \left(\sum_{i \in \mathcal{U}} Y_i \right)^2 \right] \\ &= \frac{1}{N-1} \sum_{i \in \mathcal{U}} (Y_i - \bar{Y})^2 = S_{yy} \end{aligned}$$

□

Define the $\hat{\gamma}_{ij}$ by

$$\hat{\gamma}_{ij} = \frac{1}{2} (Y_i - Y_j)^2 \quad (4.58)$$

Corollary 4.3.2. *The individual level variance will be equal to the mean of the $\hat{\gamma}$, that is*

$$S_{yy} = \bar{\hat{\gamma}} \quad (4.59)$$

Proof.

$$\begin{aligned} S_{yy} &= \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{U}} \frac{1}{2} (Y_i - Y_j)^2 \\ &= \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{U}} \hat{\gamma}_{ij} = \bar{\hat{\gamma}} \end{aligned}$$

□

The same approach applies within groups

Corollary 4.3.3. *The individual level variance within the group can be formulated as*

$$S_{yy}^{<g>} = \bar{\hat{\gamma}}_g \quad (4.60)$$

where

$$\bar{\hat{\gamma}}_g = \frac{1}{N_g(N_g - 1)} \sum_{i,j \in \mathcal{U}_g} \hat{\gamma}_{ij}$$

The square of the differences between group means can also be constructed. Define $\hat{\Gamma}_{gh}$

$$\hat{\Gamma}_{gh} = \frac{1}{2} (\bar{Y}_g - \bar{Y}_h)^2 \quad (4.61)$$

The relationships between the unweighted and the weighted group level variance are given in the following theorems.

Theorem 4.3.4. *The average of $\hat{\Gamma}_{gh}$ is equal to the unweighted group level variance,*

$$\frac{1}{M(M-1)} \sum_{g \neq h} \hat{\Gamma}_{gh} = \bar{S}_{yy} \quad (4.62)$$

Proof. Same with theorem (4.3.1) with Y_i is replaced by \bar{Y}_g . □

Theorem 4.3.5. *The weighted group level variance can be expressed in term of $\hat{\Gamma}_{gh}$, that is*

$$N \bar{S}_{yy} = \frac{1}{N(M-1)} \sum_{g,h} N_g N_h \hat{\Gamma}_{gh} \quad (4.63)$$

Proof.

$$\begin{aligned} \sum_{g,h} N_g N_h \hat{\Gamma}_{gh} &= \sum_{g,h} N_g N_h \frac{1}{2} (\bar{Y}_g^2 + \bar{Y}_h^2 - 2\bar{Y}_g \bar{Y}_h) \\ &= N \sum_g N_g \bar{Y}_g^2 - \left(\sum_g N_g \bar{Y}_g \right)^2 = N \left(\sum_g N_g \bar{Y}_g^2 - N \bar{Y}^2 \right) \end{aligned}$$

The weighted group level variance is defined in equation (4.13), and can be expressed as

$$N\bar{S}_{yy} = \frac{1}{M-1} \left(\sum_g N_g \bar{Y}_g^2 - N\bar{Y}^2 \right)$$

Therefore

$$N\bar{S}_{yy} = \frac{1}{N(M-1)} \sum_{g,h} N_g N_h \hat{\Gamma}_{gh}$$

□

We will refer to $\hat{\gamma}_{ij}$ as the individual level empirical semivariogram value for the pair (i, j) and $\hat{\Gamma}_{gh}$ as the group level empirical semivariogram value for the pairs of groups (g, h) .

4.3.1 Relationship between $\hat{\Gamma}_{gh}$ and $\hat{\gamma}_{ij}$

Theorem 4.3.6. *Relationship between the group level semivariogram and the individual level semivariogram empirically can be formulated as*

$$\hat{\Gamma}_{gh} = \bar{\gamma}_{gh} - \frac{N_g - 1}{2N_g} \bar{\gamma}_g - \frac{N_h - 1}{2N_h} \bar{\gamma}_h \quad (4.64)$$

where

$$\bar{\gamma}_{gh} = \frac{1}{N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \hat{\gamma}_{ij}; \quad \bar{\gamma}_g = \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \hat{\gamma}_{ij}$$

Note that $\bar{\gamma}_{gg} = \frac{N_g - 1}{N_g} \bar{\gamma}_g$

Proof.

$$Y_i Y_j = \frac{Y_i^2}{2} + \frac{Y_j^2}{2} - \hat{\gamma}_{ij} \quad (4.65)$$

By definition we can derive

$$\begin{aligned} (\bar{Y}_g - \bar{Y}_h)^2 &= \bar{Y}_g^2 + \bar{Y}_h^2 - 2\bar{Y}_g \bar{Y}_h \\ &= \frac{1}{N_g^2} \left(\sum_{i \in \mathcal{U}_g} Y_i \right)^2 + \frac{1}{N_h^2} \left(\sum_{i \in \mathcal{U}_h} Y_i \right)^2 - 2 \left(\sum_{i \in \mathcal{U}_g} \frac{1}{N_g} Y_i \right) \cdot \left(\sum_{j \in \mathcal{U}_h} \frac{1}{N_h} Y_j \right) \\ &= \frac{1}{N_g^2} \left(\sum_{i \in \mathcal{U}_g} Y_i^2 + \sum_{i \neq j \in \mathcal{U}_g} Y_i Y_j \right) + \frac{1}{N_h^2} \left(\sum_{i \in \mathcal{U}_h} Y_i^2 + \sum_{i \neq j \in \mathcal{U}_h} Y_i Y_j \right) - \frac{2}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} Y_i Y_j \end{aligned}$$

Substituting (4.65) gives the result

$$\begin{aligned}
(\bar{Y}_g - \bar{Y}_h)^2 &= \frac{1}{N_g^2} \left(\sum_{i \in \mathcal{U}_g} Y_i^2 + \sum_{i \neq j \in \mathcal{U}_g} \frac{Y_i^2}{2} + \frac{Y_j^2}{2} - \hat{\gamma}_{ij} \right) + \frac{1}{N_h^2} \left(\sum_{i \in \mathcal{U}_h} Y_i^2 + \sum_{i \neq j \in \mathcal{U}_h} \frac{Y_i^2}{2} + \frac{Y_j^2}{2} - \hat{\gamma}_{ij} \right) \\
&\quad - \frac{2}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \frac{Y_i^2}{2} + \frac{Y_j^2}{2} - \hat{\gamma}_{ij} \\
&= \frac{1}{N_g^2} \left(\sum_{i \in \mathcal{U}_g} Y_i^2 + (N_g - 1) \sum_{i \in \mathcal{U}_g} Y_i^2 - \sum_{i \neq j \in \mathcal{U}_g} \hat{\gamma}_{ij} \right) \\
&\quad + \frac{1}{N_h^2} \left(\sum_{i \in \mathcal{U}_h} Y_i^2 + (N_h - 1) \sum_{i \in \mathcal{U}_h} Y_i^2 - \sum_{i \neq j \in \mathcal{U}_h} \hat{\gamma}_{ij} \right) \\
&\quad - \frac{2}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \frac{Y_i^2}{2} - \frac{2}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \frac{Y_j^2}{2} + \frac{2}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \hat{\gamma}_{ij} \\
&= \frac{1}{N_g^2} \sum_{i \in \mathcal{U}_g} Y_i^2 (1 + N_g - 1) - \frac{1}{N_g^2} \sum_{i \neq j \in \mathcal{U}_g} \hat{\gamma}_{ij} + \frac{1}{N_h^2} \sum_{i \in \mathcal{U}_h} Y_i^2 (1 + N_h - 1) - \frac{1}{N_h^2} \sum_{i \neq j \in \mathcal{U}_h} \hat{\gamma}_{ij} \\
&\quad - \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} Y_i^2 - \frac{1}{N_h} \sum_{j \in \mathcal{U}_h} Y_j^2 + \frac{2}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \hat{\gamma}_{ij} \\
&= 2\bar{\gamma}_{gh} - \frac{N_g - 1}{N_g} \bar{\gamma}_g - \frac{N_h - 1}{N_h} \bar{\gamma}_h
\end{aligned}$$

Hence

$$\hat{\Gamma}_{gh} = \bar{\gamma}_{gh} - \frac{N_g - 1}{2N_g} \bar{\gamma}_g - \frac{N_h - 1}{2N_h} \bar{\gamma}_h$$

□

Theorem (4.3.6) indicates that the empirical group level semivariogram for a pair of groups is composed of three components determined by the individual level semivariogram values for pairs of individuals in the groups. The first part is the average of the individual level semivariogram between individuals in the two groups. The second and third part are the averages of the individual level semivariogram for each pair of individuals within the same groups. We will use this relationship and the fact the variances can be expressed in term of semivariogram values to relate the aggregation effect to the individual level empirical semivariogram values.

Notice that (4.64) is purely an exact algebraic result and does not depend on any model assumptions. At this stage $\hat{\gamma}_{ij}$ is merely a measure of the similarity of the individuals and $\hat{\Gamma}_{gh}$ is a measure of the similarity of the groups. However, the statistics are the elements used in variogram analysis to investigate spatial relationship. We will expand on this in section 5.1.

4.3.2 Mean square error within the group – $S_{yy}^{<W>}$

The mean square error within the group is defined

$$S_{yy}^{<W>} = \frac{1}{N-M} \sum_g \sum_{i \in \mathcal{U}_g} (Y_i - \bar{Y}_g)^2 \quad (4.66)$$

Theorem 4.3.7. *The mean square error within the group is proportional to the mean of the within group of the individual level semivariogram,*

$$S_{yy}^{<W>} = \tilde{\gamma}_w \left(1 + \bar{C}_{N\bar{\gamma}} \frac{\bar{N}(M-1)}{M(\bar{N}-1)} \right) \quad (4.67)$$

where

$$\tilde{\gamma}_w = \frac{\sum_g \bar{\gamma}_g}{M}; \quad \bar{C}_{N\bar{\gamma}} = \frac{\bar{S}_{N\bar{\gamma}}}{\bar{N} \cdot \tilde{\gamma}_w}; \quad \bar{S}_{N\bar{\gamma}} = \frac{1}{M-1} \sum_g (N_g - \bar{N})(\bar{\gamma}_g - \tilde{\gamma}_w) \quad (4.68)$$

Proof. The mean square error of the within group can be rewritten

$$S_{yy}^{<W>} = \frac{1}{N-M} \sum_g (N_g - 1) S_{yy}^{<g>}$$

Applying theorem (4.3.1) for the within group elements, the mean square error of the within group becomes

$$\begin{aligned} S_{yy}^{<W>} &= \frac{1}{N-M} \sum_g \frac{1}{N_g} \sum_{i,j \in \mathcal{U}_g} \frac{1}{2} (Y_i - Y_j)^2 = \frac{1}{N-M} \sum_g (N_g - 1) \bar{\gamma}_g \\ &= \frac{1}{N-M} \left(\sum_g N_g \bar{\gamma}_g - \sum_g \bar{\gamma}_g \right) \end{aligned}$$

Hence

$$S_{yy}^{<W>} = \frac{1}{N-M} \left(\sum_g N_g \bar{\gamma}_g - M \tilde{\gamma}_w \right) \quad (4.69)$$

Now

$$\begin{aligned} \sum_g N_g \bar{\gamma}_g &= (M-1) \bar{S}_{N\bar{\gamma}} + M \bar{N} \tilde{\gamma}_w \\ &= (M-1) \bar{N} \tilde{\gamma}_w \bar{C}_{N\bar{\gamma}} + M \bar{N} \tilde{\gamma}_w \end{aligned} \quad (4.70)$$

Substituting (4.70) into (4.69) gives

$$\begin{aligned} S_{yy}^{<W>} &= \frac{1}{N-M} \left((M-1) \bar{N} \tilde{\gamma}_w \bar{C}_{N\bar{\gamma}} + M \bar{N} \tilde{\gamma}_w - M \tilde{\gamma}_w \right) \\ &= \frac{1}{N-M} \cdot \tilde{\gamma}_w \left(\bar{C}_{N\bar{\gamma}} \bar{N}(M-1) + M(\bar{N}-1) \right) \\ &= \tilde{\gamma}_w \left(\bar{C}_{N\bar{\gamma}} \frac{\bar{N}(M-1)}{M(\bar{N}-1)} + 1 \right) \end{aligned}$$

□

4.3.3 Empirical aggregation effect

Substituting (4.64) into (4.63) gives the empirical aggregation effect in terms of $\hat{\gamma}$.

Theorem 4.3.8. *The weighted group level variance is*

$${}_N\bar{S}_{yy} = \frac{N-1}{M-1} \bar{\gamma} - \bar{N} \tilde{\gamma}_w \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) \quad (4.71)$$

Proof. Substituting (4.64) into equation (4.63) will give

$${}_N\bar{S}_{yy} = \frac{1}{N(M-1)} \sum_{g,h} N_g N_h \left(\bar{\gamma}_{gh} - \frac{N_g-1}{2N_g} \bar{\gamma}_g - \frac{N_h-1}{2N_h} \bar{\gamma}_h \right)$$

Modifying this equation may derive into

$$\begin{aligned} {}_N\bar{S}_{yy} &= \frac{1}{N(M-1)} \left(\sum_{g,h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \gamma_{ij} - \sum_{g,h} \frac{N_h}{2} (N_g-1) \bar{\gamma}_g - \sum_{g,h} \frac{N_g}{2} (N_h-1) \bar{\gamma}_h \right) \\ &= \frac{1}{N(M-1)} \left(N(N-1) \bar{\gamma} - \sum_h \frac{N_h}{2} \sum_g (N_g-1) \bar{\gamma}_g - \sum_g \frac{N_g}{2} \sum_h (N_h-1) \bar{\gamma}_h \right) \\ &= \frac{1}{N(M-1)} \left(N(N-1) \bar{\gamma} - \sum_h \frac{N_h}{2} \left[(M-1) \bar{S}_{N\bar{\gamma}} + M \bar{N} \tilde{\gamma}_w - M \tilde{\gamma}_w \right] \right. \\ &\quad \left. - \sum_g \frac{N_g}{2} \left[(M-1) \bar{S}_{N\bar{\gamma}} + M \bar{N} \tilde{\gamma}_w - M \tilde{\gamma}_w \right] \right) \\ &= \frac{N-1}{M-1} \bar{\gamma} - \bar{S}_{N\bar{\gamma}} - \tilde{\gamma}_w \frac{M(\bar{N}-1)}{M-1} \end{aligned}$$

Substituting (4.68) for the $\bar{S}_{N\bar{\gamma}}$ completes the proof. \square

Corollary 4.3.9. *The difference between weighted group level variance and individual level variance is*

$${}_N\bar{S}_{yy} - S_{yy} = \frac{N-M}{M-1} \bar{\gamma} - \bar{N} \tilde{\gamma}_w \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) \quad (4.72)$$

Proof. Recall the $S_{yy} = \bar{\gamma}$, see (4.60). The difference of ${}_N\bar{S}_{yy}$ and S_{yy} is

$$\begin{aligned} {}_N\bar{S}_{yy} - S_{yy} &= \frac{N-1}{M-1} \bar{\gamma} - \bar{\gamma} - \bar{S}_{N\bar{\gamma}} - \tilde{\gamma}_w \frac{M(\bar{N}-1)}{M-1} \\ &= \frac{N-M}{M-1} \bar{\gamma} - \tilde{\gamma}_w \frac{N-M}{M-1} - \bar{C}_{N\bar{\gamma}} \bar{N} \tilde{\gamma}_w \end{aligned}$$

\square

Corollary 4.3.10. *The ratio between weighted group level variance and individual level variance is*

$$\frac{{}_N\bar{S}_{yy}}{S_{yy}} = \frac{N-1}{M-1} - \bar{N} \frac{\tilde{\gamma}_w}{\bar{\gamma}} \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) \quad (4.73)$$

Corollaries (4.3.9) and (4.3.10) formulate the aggregation effect in terms of the difference and ratio of the weighted group level variance and individual level variance. The key terms are $(\bar{\gamma} - \tilde{\gamma}_W)$ and $\frac{\tilde{\gamma}_W}{\bar{\gamma}}$ respectively. The value of $\bar{\gamma}$ is free from any zoning or scaling effect, but $\tilde{\gamma}_W$ depends on the zoning or scale. For $\bar{C}_{N\bar{\gamma}}$ small and M large, $N\bar{S}_{yy} - S_{yy} \approx (\bar{N} - 1)(\bar{\gamma} - \tilde{\gamma}_W)$ and so the scale effect is determined by \bar{N} and how scale $\tilde{\gamma}_W$. The input of different zoning at the same scale depends on how the different zoning affect $\tilde{\gamma}_W$.

4.4 Aggregation effects in term of variance of differences

In section (4.3) the aggregation effect was related to the differences for pairs of observations and $\hat{\gamma}_{ij}$ was treated as an empirical observation. In this section we look at the expectation of the aggregation effect and relate it to parameters of the distribution of the variables. Essentially the same results are obtained with the empirical semivariogram values replaced by the corresponding model based semivariogram values.

Consider the difference between Y_i and Y_j . For the assumptions given by (4.2), the expectation and variance of this difference are

$$E(Y_i - Y_j) = \mu_i - \mu_j; \quad i, j \in \mathcal{U} \quad (4.74)$$

and

$$V(Y_i - Y_j) = (\Sigma_i + \Sigma_j) - 2\Delta_{ij} \quad (4.75)$$

Define γ_{ij}

$$\begin{aligned} \gamma_{ij} &= \frac{1}{2} V(Y_i - Y_j) \\ &= \frac{1}{2} (\Sigma_i + \Sigma_j) - \Delta_{ij} \end{aligned} \quad (4.76)$$

The equation (4.76) implies a relationship between γ_{ij} and Δ_{ij} ,

$$\Delta_{ij} = \frac{1}{2} (\Sigma_i + \Sigma_j) - \gamma_{ij} \quad (4.77)$$

In spatial modeling, the γ_{ij} and Δ_{ij} are referred to as the semivariogram and covariogram for the pairs (i, j) , respectively (see Cressie, 1991; Haining, 1990). Equation (4.77) shows the relationship between the covariogram and the semivariogram. The semivariogram has advantages over the covariogram, since the

semivariogram estimator is more reliable than the covariogram estimator. The semivariogram is defined in cases when the covariogram is not (see Cressie, 1991, page 70). A simple case can be observed for a situation at a very close distance (or at distance zero), in which the covariogram is not define but the semivariogram is defined as a nugget. The relation between Δ_{ij} and γ_{ij} allows us to examine the aggregation effect in several ways. The semivariogram is often used as part of the kriging analysis, which is an analysis to estimate points values and construct the estimated surface map over the study region based on a few observations points (Cressie, 1991). Carrat and Valleron (1992) discussed such application in epidemiologic mapping. But Amrhein and Reynolds (1996) investigate other spatial statistics in the assessment of the aggregation effect, such as the Moran coefficient, Geary ratio, and $G_i(d)$ (see section 3.4.2).

Lemma 4.4.1. *The average of individual level covariance is*

$$\bar{\Delta} = \bar{\Sigma} - \bar{\gamma} \quad (4.78)$$

where

$$\bar{\gamma} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} \quad (4.79)$$

Proof. The $\bar{\Delta}$ is defined by

$$\bar{\Delta} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \Delta_{ij}$$

Then it may be modified into

$$\begin{aligned} \bar{\Delta} &= \frac{1}{N(N-1)} \left[\frac{1}{2} \sum_{i \neq j \in \mathcal{U}} (\Sigma_i + \Sigma_j) - \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} \right] \\ &= \frac{1}{N(N-1)} \left[\frac{1}{2} \sum_{i \in \mathcal{U}} \left\{ (N-1)\Sigma_i + \sum_{i \neq j \in \mathcal{U}} \Sigma_j \right\} - \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} \right] \\ &= \frac{1}{N(N-1)} \left[\frac{1}{2} \sum_{i \in \mathcal{U}} \left\{ (N-1)\Sigma_i + \sum_{j \in \mathcal{U}} \Sigma_j - \Sigma_i \right\} - \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} \right] \\ &= \frac{1}{N(N-1)} \left[\frac{1}{2} \cdot (N-2) \sum_{i \in \mathcal{U}} \Sigma_i + \frac{1}{2} N \sum_{j \in \mathcal{U}} \Sigma_j - \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} \right] \\ &= \frac{1}{N(N-1)} \cdot (N-1) \sum_{i \in \mathcal{U}} \Sigma_i - \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} \\ &= \bar{\Sigma} - \bar{\gamma} \end{aligned}$$

□

Substituting (4.78) in (4.4) gives

$$V(\bar{Y}) = \bar{\Sigma} - \left(\frac{N-1}{N} \right) \bar{\gamma} \quad (4.80)$$

and

Theorem 4.4.2. *Expectation of the population variance is*

$$E(S_{yy}) = \bar{\gamma} + S_{\mu\mu} \quad (4.81)$$

Proof. Use equations (4.18) and (4.78). □

Define the within group average of semivariogram, $\bar{\gamma}_g$ as

$$\bar{\gamma}_g = \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \gamma_{ij} \quad (4.82)$$

Then

$$\bar{\gamma}_g = \bar{\Sigma}_g - \bar{\Delta}_g \quad (4.83)$$

By applying theorem (4.4.2) and theorem (4.1.6) for the g th group, we have

Corollary 4.4.3.

$$E(S_{yy}^{<g>}) = \bar{\gamma}_g + S_{\mu\mu}^{<g>} = \bar{\Sigma}_g - \bar{\Delta}_g + S_{\mu\mu}^{<g>} \quad (4.84)$$

Proof. Applying theorem (4.4.2) for a particular group. □

Theorem 4.4.4. *The variance and covariance of the group level mean in term of γ 's are*

$$\begin{aligned} V(\bar{Y}_g) &= \bar{\Sigma}_g - \left(\frac{N_g - 1}{N_g} \right) \bar{\gamma}_g \\ \text{Cov}(\bar{Y}_g; \bar{Y}_h) &= \bar{\Delta}_{gh} = \frac{1}{2}(\bar{\Sigma}_g + \bar{\Sigma}_h) - \bar{\gamma}_{gh}; \quad \text{for } g \neq h \end{aligned} \quad (4.85)$$

where $\bar{\gamma}_{gh}$ is

$$\bar{\gamma}_{gh} = \frac{1}{N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \gamma_{ij}; \quad \text{for } g \neq h \quad (4.86)$$

Proof. Substituting (4.84) into (4.8-ii) to get $V(\bar{Y}_g)$, and applying (4.77) in theorem (4.1.2) to derive $Cov(\bar{Y}_g; \bar{Y}_h)$. \square

In the way similar to equation (4.28) and (4.29), we can define $\bar{\gamma}_W$ and $\bar{\gamma}_B$ as follow,

$$\bar{\gamma}_B = \frac{\sum_{g \neq h} N_g N_h \bar{\gamma}_{gh}}{\sum_{g \neq h} N_g N_h} \quad ; \text{ and } \quad \bar{\gamma}_W = \frac{\sum_g N_g (N_g - 1) \bar{\gamma}_g}{\sum_g N_g (N_g - 1)} \quad (4.87)$$

and the unweighted versions

$$\tilde{\gamma}_B = \frac{1}{M(M-1)} \sum_{g \neq h} \bar{\gamma}_{gh} \quad ; \text{ and } \quad \tilde{\gamma}_W = \frac{1}{M} \sum_g \bar{\gamma}_g \quad (4.88)$$

Theorem 4.4.5. *The $\bar{\gamma}$ can be expressed as function of $\bar{\gamma}_B$ and $\bar{\gamma}_W$.*

$$\bar{\gamma} = \frac{1}{N-1} \left\{ [\bar{N}(1+C^2) - 1] \bar{\gamma}_W + [N - \bar{N}(1+C^2)] \bar{\gamma}_B \right\} \quad (4.89)$$

where C^2 is defined in (4.31).

Proof. The proof can be obtained from the basic relation that

$$\sum_{i \neq j \in \mathcal{U}} V(Y_i - Y_j) = \sum_g \sum_{i \neq j \in \mathcal{U}_g} V(Y_i - Y_j) + \sum_{g \neq h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} V(Y_i - Y_j)$$

Multiplying both sides by $\frac{1}{2}$, then

$$\begin{aligned} \sum_{i \neq j \in \mathcal{U}} \gamma_{ij} &= \sum_g \sum_{i \neq j \in \mathcal{U}_g} \gamma_{ij} + \sum_{g \neq h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \gamma_{ij} \\ N(N-1)\bar{\gamma} &= \sum_g N_g(N_g-1)\bar{\gamma}_g + \sum_{g \neq h} N_g N_h \bar{\gamma}_{gh} \end{aligned}$$

Applying (4.87), gives

$$N(N-1)\bar{\gamma} = \bar{\gamma}_W \sum_g N_g(N_g-1) + \bar{\gamma}_B \sum_{g \neq h} N_g N_h$$

Substituting (4.32) and (4.33) into this equation completes the proof. \square

Corollary 4.4.6. *If N_g is constant then this implies that $C^2 = 0$, $\bar{\gamma}_W = \tilde{\gamma}_W$, and $\bar{\gamma}_B = \tilde{\gamma}_B$, and (4.89) becomes*

$$\bar{\gamma} = \frac{\bar{N}-1}{N-1} \tilde{\gamma}_W + \frac{\bar{N}(M-1)}{N-1} \tilde{\gamma}_B \quad (4.90)$$

The relative covariance between N_g and $\bar{\gamma}_g$ is useful in considering the expectation of the weighted variance. Define

$$\bar{C}_{N\bar{\gamma}} = \frac{\bar{S}_{N\bar{\gamma}}}{\bar{N} \cdot \bar{\gamma}_W} \quad (4.91)$$

where $\bar{S}_{N\bar{\gamma}}$ is the covariance between N_g and $\bar{\gamma}_g$,

$$\bar{S}_{N\bar{\gamma}} = \frac{1}{M-1} \sum_g (N_g - \bar{N}) \cdot (\bar{\gamma}_g - \bar{\gamma}_W) \quad (4.92)$$

Equation (4.92) can be rearranged to give

$$\sum_g N_g \cdot \bar{\gamma}_g = (M-1)\bar{S}_{N\bar{\gamma}} + M\bar{N}\bar{\gamma}_W \quad (4.93)$$

Theorem 4.4.7. *Expectation of the unweighted variance is*

$$E({}_1\bar{S}_{yy}) = \bar{\gamma}_B - \bar{\gamma}_W + \frac{1}{M} \sum_g \frac{\bar{\gamma}_g}{N_g} + {}_1\bar{S}_{\mu\mu} \quad (4.94)$$

Proof. By theorem (4.1.4) and (4.4.4)

$$\begin{aligned} E({}_1\bar{S}_{yy}) &= \frac{1}{M} \sum_g \left(\bar{\Sigma}_g - \left(\frac{N_g - 1}{N_g} \right) \bar{\gamma}_g \right) - \frac{1}{M(M-1)} \sum_{g \neq h} \left(\frac{1}{2} (\bar{\Sigma}_g + \bar{\Sigma}_h) - \bar{\gamma}_{gh} \right) \\ &\quad + {}_1\bar{S}_{\mu\mu} \\ &= \frac{1}{M} \sum_g \bar{\Sigma}_g - \frac{1}{M} \sum_g \left[1 - \frac{1}{N_g} \right] \bar{\gamma}_g - \frac{1}{M} \sum_g \bar{\Sigma}_g + \bar{\gamma}_B + {}_1\bar{S}_{\mu\mu} \end{aligned}$$

Then this equation can be simplified to complete the proof. □

Theorem 4.4.8. *Expectation of the weighted variance is*

$$E({}_N\bar{S}_{yy}) = \frac{N-1}{M-1} \bar{\gamma} - \bar{N}\bar{\gamma}_W \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) + {}_N\bar{S}_{\mu\mu} - \frac{N-1}{M-1} S_{\mu\mu} \quad (4.95)$$

Proof. Equation (4.23) of theorem (4.2.2) is modified into

$$\begin{aligned} E({}_N\bar{S}_{yy}) &= \frac{1}{M-1} \left((N-1)\bar{\gamma} - \sum_g (N_g - 1)\bar{\gamma}_g \right) + {}_N\bar{S}_{\mu\mu} - \frac{N-1}{M-1} S_{\mu\mu} \\ &= \left(\frac{N-1}{M-1} \right) \bar{\gamma} - \sum_g \left(\frac{N_g - 1}{M-1} \right) \bar{\gamma}_g + {}_N\bar{S}_{\mu\mu} - \frac{N-1}{M-1} S_{\mu\mu} \end{aligned}$$

Substituting (4.93), (4.92), and (4.91) into the above equation completes the proof. □

Theorem (4.4.8) shows that the expectation of the weighted group level variance may change for different values of \bar{N} , i.e. varying scales. The $\bar{\gamma}$ is not affected by the scale or zoning but $\tilde{\gamma}_W$ and $\bar{C}_{N\bar{\gamma}}$ are both affected by both the scale and the zoning. Theorem (4.4.8) can be used to obtain the aggregation effect, that is a difference or a ratio between group level variance and the individual level variance.

Theorem 4.4.9. *Expectation of the aggregation effect in term of the difference between the weighted group level variance and its individual level variance is*

$$E(N\bar{S}_{yy} - S_{yy}) = \left(\frac{N-M}{M-1} \right) \bar{\gamma} - \bar{N}\tilde{\gamma}_W \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) + N\bar{S}_{\mu\mu} - S_{\mu\mu} \quad (4.96)$$

Proof. The proof is immediately from theorem (4.4.8) and (4.4.2). \square

We may consider some special cases, such as the case when N_g constant at \bar{N} and the case when $\bar{\gamma}_g$ constant.

Corollary 4.4.10. *Expectation of the aggregation effect when the \bar{N} or $\bar{\gamma}_g$ is constant is*

$$E(N\bar{S}_{yy} - S_{yy}) = \left(\frac{N-M}{M-1} \right) (\bar{\gamma} - \tilde{\gamma}_W) + N\bar{S}_{\mu\mu} - S_{\mu\mu} \quad (4.97)$$

Proof. The constant N_g or $\bar{\gamma}_g$ will imply the $\bar{C}_{N\bar{\gamma}} = 0$ in equation (4.96). \square

Theorem 4.4.11. *Expectation of the aggregation effect in term of the ratio between the weighted group level variance and its individual level variance is, when $\mu_i = \mu$*

$$\frac{E(N\bar{S}_{yy})}{E(S_{yy})} = \frac{N-1}{M-1} - \bar{N} \frac{\tilde{\gamma}_W}{\bar{\gamma}} \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) \quad (4.98)$$

Proof. The proof can be done by dividing equation (4.95) by $E(S_{yy})$, then applying theorem (4.4.2) to substitute the $E(S_{yy})$ with $\bar{\gamma}$. \square

4.5 Summary

The aggregation effect may be considered in terms of the difference between group and individual level variance. In general it was presented in theorem (4.2.7) and corollary (4.2.8). Both formulations have a common factor, that the aggregation effect depend on $\bar{\Delta}$ and $\tilde{\Delta}_W$. The $\bar{\Delta}$ component is not affected by any scale or zoning, but $\tilde{\Delta}_W$ is. In general different scaling or zoning will produce different values of $\tilde{\Delta}_W$.

The square of the difference and the variance of the difference of a pair of observations are called the empirical and theoretical semivariogram (equation 4.56 and 4.76), respectively. By algebraic manipulation, the group level square of the difference of means ($\hat{\Gamma}_{gh}$) was decomposed into individual elements of square of difference (see theorem 4.3.6). The components are the mean of between group ($\bar{\gamma}_{gh}$) and the mean of within group ($\bar{\gamma}_g$ and $\bar{\gamma}_h$) empirical semivariogram values.

The aggregation effect is considered in two different ways, i.e. the difference and the ratio of between group and individual level variance. The empirical or theoretic perspective gave similar result, as we can look at equation (4.72) and (4.73) in the empirical perspective and equation (4.96) and (4.98) in theoretical perspective. Those results show that the key factors determining the aggregation effect are the average number of units in the group (\bar{N}) and the average of the within group semivariogram values, $\bar{\gamma}_W$ or $\bar{\gamma}_W$. The semivariogram is a well established approach to study spatial variation. Hence these results give a clear spatial explanation to the aggregation effect. To determine the effect of aggregation, we just determine the effect on the within group semivariogram values.

Chapter 5

The Role of Semivariogram In Aggregation Effect

The variogram and semivariogram are important tools in studying spatial variability (Cressie, 1991; Clifford et al., 1989; and Diggle et al., 1998). The variogram is defined by the variances of the difference between two observations at two locations and the semivariogram is obtained from the variogram by dividing by 2.

The objective of this chapter is to consider the use of the semivariogram in the analysis of aggregate social data. Aggregate group level data can be used to construct a group level semivariogram. In chapter four it was shown how the aggregation effect for variances can be related to the individual level semivariogram. In this chapter we will consider the relationship between the group level and individual level semivariogram in more detail. Estimation of the individual level semivariogram is useful as it gives information about spatial structure at the individual level. This information may be useful in its own right or it can be used to understand the aggregation effect on variances. We will consider some methods of estimating individual level semivariogram using group level data.

5.1 Variogram and semivariogram

In section 4.3.3 we related the empirical aggregation effect to the quantities $\hat{\gamma}_{ij} = \frac{1}{2}(Y_i - Y_j)^2$. The relationships obtained in 4.3.3 were purely arithmetic. In section 4.4 we considered the expectation of the aggregation effect and related it to $\gamma_{ij} = \frac{1}{2}V(Y_i - Y_j)$.

5.1.1 Definition and assumptions

In general γ_{ij} can depend on the individual involved and the locations of all the individuals in the population, which are contained in the matrix \mathbf{L} . In spatial analysis and modeling it is usually assumed that γ_{ij} depends only on the location of the two individuals concerned, that is

$$\gamma_{ij} = \gamma(\ell_i, \ell_j) \quad (5.1)$$

A further assumption is that only the relative position of the two individuals is relevant, that is

$$\gamma_{ij} = \gamma(\ell_i - \ell_j) \quad (5.2)$$

where $\ell_i - \ell_j$ is called the increment. Note $\gamma_{ii} = 0$, but for two different individuals at the same location, i.e. $\ell_i = \ell_j$ for $i \neq j$, $\gamma_{ij} = \gamma(0)$, which does not have to be zero.

The expression $2\gamma(\cdot)$ is called a variogram and $\gamma(\cdot)$ is called a semivariogram. The spatial process (Y_i, ℓ_i) is called intrinsically stationary if (5.2) applies and also

$$E(Y_i - Y_j) = 0 \quad (5.3)$$

Intrinsic stationarity is a weaker condition than second order stationarity. The second order stationarity assumptions involve three conditions;

$$\begin{aligned} (i) \quad & E(Y_i) = \mu; \quad \forall \quad \ell \in \mathcal{D} \subset \mathcal{R}^d \\ (ii) \quad & V(Y_i) = \sigma^2 \\ (iii) \quad & Cov(Y_i; Y_j) = C(\ell_i - \ell_j); \quad \forall \quad \ell_i, \ell_j \in \mathcal{D}, \quad i \neq j \end{aligned} \quad (5.4)$$

where it is assumed that the observations have a constant mean and variance over the population and the covariogram, $C(\cdot)$, is a function of the relative location of two observations. Note we will use $C(0)$ to denote the covariance of two different individuals at the same location, which in practice means within the same household.

A prominent aspect in dealing with spatial data is the distance between locations. A distance is usually defined as Euclidean distance, for example in \mathcal{R}^2 the distance between two individuals at location $\ell_i = (e_i, n_i)'$ and $\ell_j = (e_j, n_j)'$ is

$$d_{ij} = \sqrt{(e_i - e_j)^2 + (n_i - n_j)^2} = \|\ell_i - \ell_j\| \quad (5.5)$$

where the e_i and n_i indicate the location of the point e_i units to the east and n_i units to the north of an arbitrary origin point. The arbitrary origin point can be chosen to be the most westerly and most southerly point in the region under study.

If $\gamma(\ell_i - \ell_j)$ is a function of d_{ij} only, that is $\gamma(\ell_i - \ell_j) = \gamma(d_{ij})$, then it is called isotropic. The isotropic condition need not assume a constant mean over the population, but if the isotropic condition is assumed in conjunction with second order stationarity, then $C(\cdot)$ can be represented as a function of $\|\ell_i - \ell_j\|$ only (Christensen, 1991; and Grondona & Cressie, 1991). This condition is often reasonable in describing social phenomena, since in social data it is assumed that the interaction among individuals occurs in all directions. Therefore the direction is assumed not to have a significant influence in the analysis of spatial data.

The distance between two groups with centroids at $\ell_g = (e_g, n_g)'$ and $\ell_h = (e_h, n_h)'$ respectively can be defined by

$$d_{gh} = \sqrt{(e_g - e_h)^2 + (n_g - n_h)^2} = \|\ell_g - \ell_h\| \quad (5.6)$$

The distance between the centroids is not necessarily the same as the average distance between individuals in the g th and h th groups. Define,

$$\bar{d}_g = \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} d_{ij} \quad \text{and} \quad \bar{d}_{gh} = \frac{1}{N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} d_{ij} \quad (5.7)$$

The quantity \bar{d}_{gh} represents the average distance between all pairs of points within the g th and h th groups, whereas d_{gh} represents distance between the centroids of the g th and h th groups. In general \bar{d}_{gh} will not be equal to d_{gh} . The relationship between d_{gh} and \bar{d}_{gh} will be investigated later and is an additional issue that arises when the spatial aspects of aggregation are considered. The quantity \bar{d}_g is the average distance between all pairs of points within group g .

Corollary 5.1.1. *Relationship between semivariogram and covariogram is*

$$\gamma(d_{ij}) = \sigma^2 - C(d_{ij}) \quad (5.8)$$

The quantity $C(0)$ can be viewed as the covariogram value at distance zero, $d_{ij} = 0$, but is not necessarily equal to σ^2 . The reason is that observations corresponding to $d_{ij} = 0$ in social data may come

from one individual or two different individuals at the same location. Therefore the estimator of $C(0)$ in general is not the same as the estimator of σ^2 . This fact will affect the estimation of the semivariogram, if we cannot distinguish between individuals within the same household. This issue will be discussed in more detail in section (5.1.6).

5.1.2 Semivariogram model and its parameter

The semivariogram is often modeled through a distance function. Most models for the isotropic semivariogram contain three parameters, called the nugget (n), sill (s), and range (r), which are in the interval $[0, \infty)$. The theoretical semivariogram models are typically drawn as shown in figure (5.1).

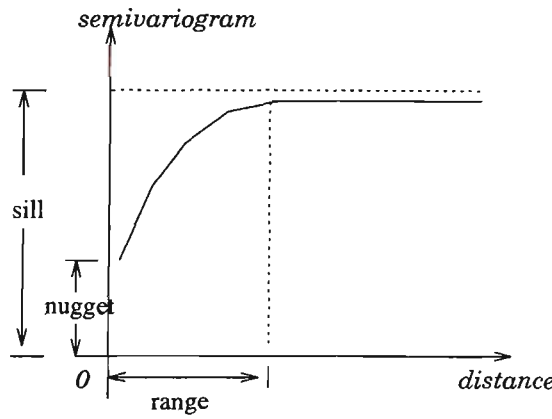


Figure 5.1. Graph of the theoretical semivariogram model

Cressie (1991) presented some common isotropic semivariogram models :

- Exponential model :

$$\gamma(d_{ij}) = n + (s - n)(1 - \exp\left[\frac{-d_{ij}}{r}\right]), \quad d_{ij} \geq 0 \quad (5.9)$$

- Spherical model :

$$\gamma(d_{ij}) = \begin{cases} n + (s - n) \left(\frac{3}{2} \frac{d_{ij}}{r} - \frac{1}{2} \left[\frac{d_{ij}}{r} \right]^3 \right) & 0 \leq d_{ij} < r, \\ s & d_{ij} \geq r. \end{cases} \quad (5.10)$$

- Gaussian model :

$$\gamma(d_{ij}) = n + (s - n) \left(1 - \exp\left[\frac{-d_{ij}^2}{r^2}\right] \right), \quad d_{ij} \geq 0 \quad (5.11)$$

Cressie (1991) cited other models such as the power model, quadratic model and wave model. These models are not discussed further in this section.

Figure (5.1) exhibits three important distances related to the parameters of a semivariogram model. They are the distances equal to zero, infinity, and r .

Theoretically, for geological and other physical data $d_{ij} = 0$ implies $Y_i = Y_j$ at $\ell_i = \ell_j$ and so the semivariogram will be zero at $d_{ij} = 0$, that is $\gamma(0) = 0$. This may be considered as the ideal situation where there is no measurement error when measuring objects at the same location (Cressie, 1991). But in practice, it is common to have $\gamma(d_{ij}) = n > 0$ as d_{ij} approaches 0. For physical data n is sometimes interpreted as the effect of measurement error. For social data additional issues arise and these will be discussed in section (5.1.6).

If $\gamma(d_{ij}) = s$ as d_{ij} approaches ∞ then s is called the sill. The sill is the asymptotic limit of the semivariogram model and will indicate the variance of the process being studied.

The last distance of interest is $d_{ij} = r$. We can evaluate the value of $\gamma(\cdot)$ for each model at this point, that is

- Exponential and Gaussian model :

$$\gamma(r) = s \left(1 - \frac{1}{e} \right) + n \cdot \frac{1}{e} = 0.632s + 0.368n \quad (5.12)$$

where e is equal to $2.718 \dots$.

- Spherical model :

$$\gamma(r) = s \quad (5.13)$$

For the spherical model d_o is a distance between observations such that the variance of the $(Y_i - Y_j)$ becomes constant. This distance can be interpreted as a situation where $\rho(d_o) = 0$, that is, a correlation between Y_i and Y_j is equal to zero for all $d_{ij} \geq d_o$. In other words, the observations are independent with each other at the distance greater or equal than d_o . For the exponential and Gaussian models observations will get close to independent for d_{ij} greater than some multiple of r . For example, bias in the exponential model $\rho(d) \leq 0.05$ for $d > 3r$ (Carr, 1995).

5.1.3 Estimation of the semivariogram

Theorem 5.1.2. Consider the estimator $\hat{\gamma}_{ij}$ of γ_{ij} then

$$E(\hat{\gamma}_{ij}) = \gamma_{ij} + \frac{1}{2} (E(Y_i - Y_j))^2 \quad (5.14)$$

where $\gamma_{ij} = \frac{1}{2} V(Y_i - Y_j)$

Proof.

$$\begin{aligned} E(\hat{\gamma}_{ij}) &= \frac{1}{2} E((Y_i - Y_j)^2) \\ &= \frac{1}{2} V(Y_i - Y_j) + \frac{1}{2} \cdot (E(Y_i - Y_j))^2 \\ &= \gamma_{ij} + \frac{1}{2} \cdot (E(Y_i - Y_j))^2 \end{aligned}$$

□

Corollary 5.1.3. If $E(Y_i - Y_j) = 0$ the $\hat{\gamma}_{ij}$ is unbiased for γ_{ij}

For an isotropic process, $\hat{\gamma}_{ij}$ is an unbiased estimator of $\gamma(d_{ij})$. Hence the data $\hat{\gamma}_{ij}$ can be used to model $\gamma(d_{ij})$ as a function of distance between observations.

The statistics $\hat{\gamma}_{ij}$ have been defined as unit level empirical semivariogram values. If there are N observations in the population then these will be $\frac{N(N-1)}{2}$ empirical semivariogram values, and corresponding distances, d_{ij} . The literature has described some ways of modeling the semivariogram using the distances as independent variables (Cressie, 1991).

Let $\hat{\gamma}(d_{ij})$ denote an estimator of the semivariogram at distance d_{ij} . Unless N is small the number of empirical semivariogram values will be very large and modeling using $\hat{\gamma}_{ij}$ may not be computationally feasible, or at least very computer intensive. The common initial way of modeling $\hat{\gamma}(d_{ij})$ is by categorizing the unit level empirical semivariogram values. The categories are created based on distance classes and the average of the unit semivariogram values within each distance class is calculated. The categorization is done by dividing the span distance of pairs of points within the region from minimum to maximum distance into K intervals. Define D_k , for $k = \{1, \dots, K\}$ as the distance class $D_k = [d_{k-1}, d_k)$ of width w_k , and the average distance \bar{d}_k , which is defined by

$$\bar{d}_k = \frac{1}{|N_{D_k}|} \sum_{d_{ij} \in D_k}^{N_{D_k}} d_{ij}; \quad k = 1, \dots, K \quad (5.15)$$

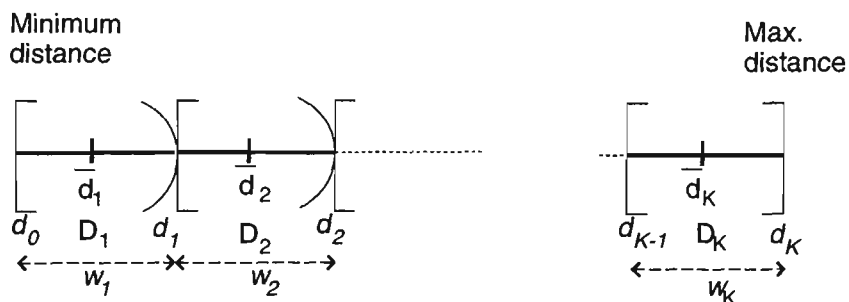


Figure 5.2. Distance classes

This can be drawn as in figure (5.2). The $|N_{D_k}|$ represents numbers of the semivariogram values within distance class D_k . The initial distance case is usually constructed by setting $d_0 = 0$. The classical categorized estimator of the variogram (Cressie, 1991) is,

$$\hat{\gamma}(\bar{d}_k) = \frac{1}{|N_{D_k}|} \sum_{d_{ij} \in D_k}^{N_{D_k}} \hat{\gamma}_{ij}; \quad k = 1, \dots, K \quad (5.16)$$

SAS (1996) noted a rule of thumb used in constructing categorical semivariograms, which is to use at least 30 pairs of points in computing a single value of the categorical semivariogram. If the interval width is set to small, there may be too few points within the categories. On the other hand, if the interval width is set too large, the number pairs of points within the categories may be much greater than that needed, thereby wasting data. Usually equal width intervals are used, $w_k = w$. Clark (1982) suggested that a good choice for the interval width (w) is 10 percent of the average distance of pairs of points within the region.

A semivariogram model is developed by fitting the $\hat{\gamma}_{ij}$ or $\hat{\gamma}(\bar{d}_k)$ to a theoretical semivariogram model, and estimating the parameters (n , s , and r). A relationship between $\hat{\gamma}$ and \bar{d}_k is the objective of the model fitting, and will be based on a semivariogram model. Several methods can be used. Cressie (1991) discussed several parametric methods to fit a variogram model, such as maximum likelihood (ML), restricted maximum likelihood method (REML), minimum norm quadratic (MINQ) estimation, generalized least squares and weighted least squares.

Zimmerman and Zimmerman (1991) compared several methods of variogram model fitting, ordinary least squares, weighted least squares, maximum likelihood, restricted maximum likelihood, and generalized minimum variance quadratic. Four different semivariogram models were considered, these being linear, exponential, spherical, and Gaussian. They concluded that different methods of model fitting can

be used for different semivariogram models, for example a linear model could be fitted by ordinary least square but for the exponential model using ML gives a better result.

Before the estimation procedure is applied, it is essential to make a plot of semivariogram values versus distance, either for the $\hat{\gamma}_{ij}$ or $\hat{\gamma}(\bar{d}_k)$. The plots may be used to determine initial values of the parameters and check that the theoretical semivariogram model is reasonable. Some estimation procedures can be used to fit $\hat{\gamma}_{ij}$ or $\hat{\gamma}(\bar{d}_k)$ to the model. If $\hat{\gamma}_{ij}$ is being used then non-linear regression or maximum likelihood methods can be applied. But in practice, the $\hat{\gamma}(\bar{d}_k)$ are often used to estimate the model. Then weighted least squares method could be used to estimate the semivariogram model parameters. Cressie (1985), suggested minimizing the weighted sum of squares;

$$\sum_{k=1}^K |N_{D_k}| \left(\frac{\hat{\gamma}(\bar{d}_k)}{\gamma(\bar{d}_k; n, s, r)} - 1 \right)^2 \quad (5.17)$$

where $\gamma(\bar{d}_k; n, s, r)$ is the specified variogram model with unknown parameters (n, s, r) . In this study we used the categorized approach and this weighted least squares method in estimating the parameters of the semivariogram models, for both individual and group level data. This leads to a non-linear regression method with weight

$$w = |N_{D_k}| \cdot \left(\frac{\hat{\gamma}(\bar{d}_k)}{\gamma(\bar{d}_k; \hat{n}, \hat{s}, \hat{r})} - 1 \right)^2 \quad (5.18)$$

5.1.4 Illustration of semivariogram from Illawarra dataset

As an illustration of semivariogram analysis we considered some data from the Illawarra region. The Illawarra region is located immediately south of Sydney in the state of New South Wales in Australia. The data used was at the collection district (CD) level. A CD is an area containing approximately 200-300 households (Castles, 1991). There are 377 collection districts in the 1087.3 km square total region area. The data were sourced from the Australian Census of Population and Housing 1991. Although these data are actually group level, they can be used to illustrate semivariogram analysis. Also, this type of analysis is used to analyse social data.

The labor participation rate will be used to illustrate the semivariogram model in the census data. The labor participation rate is calculated based on the total number of person older than 15 years old.

According to the 1991 census, there were 182,163 persons older than 15 years old in the Illawarra region. The distribution of this characteristic across CDs is described the box-plot in figure 5.3.

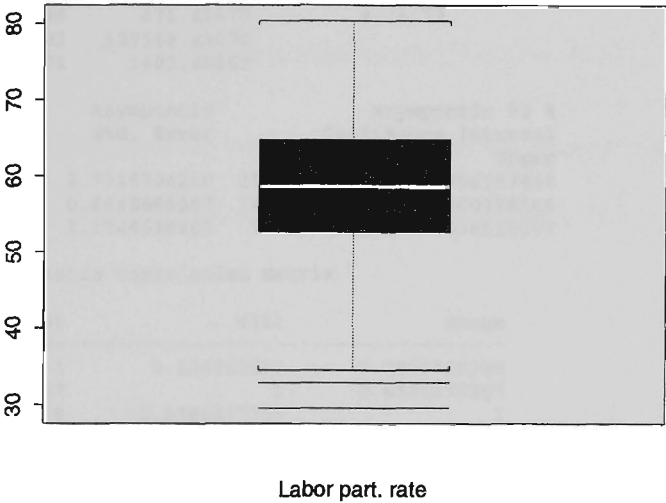


Figure 5.3. The CD distribution of the labor participation rate (%) in Illawarra NSW, mean=58.55, median=58.68, min.=29.64, max.=80.34, variance=73.38

The SAS software was used to develop a program to analyze the semivariogram associated with this variable. The exponential model is applied to explain the empirical model of the labor participation rate. The parameters of the model (nugget, sill, and range) are estimated by the weighted non-linear least squares method (Cressie, 1985). The weight is defined in (5.18). The results are shown in the following output. The complete computer code is listed in the appendix (D). Here we have reparameterized the exponential model by using $3d/r$ instead of d/r .

The output (5.1) shows that the estimated nugget, sill, and range are 44.49, 75.98, and 9.81, respectively. The estimated sill is close to the variance (73.38). The estimated nugget indicates that variation at distance zero of the labor participation rate of the CD's level is 44.49. In this context, then this variation shows a measure of dispersion of observations points within the CD. Meanwhile, the estimated range indicates that within the distance 9.81 km the CD level of the labor participation rate is dependent between each other but it is close to independent afterward. Hence the CD level values are affected by surrounding CD within the distance 9.81 kms. This results are presented in figure 5.4. The squares (\square) indicates the empirical semivariogram and the star (*) indicates the empirical exponential model of the semivariogram.

Table 5.1. The SAS output of the estimation procedure

Non-Linear Least Squares Summary Statistics
Dependent Variable: labor sv.

| Source | DF | Weighted SS | Weighted MS |
|-------------------|-----|--------------|-------------|
| Regression | 3 | 128688.00000 | 42896.00000 |
| Residual | 99 | 471.41070 | 4.76172 |
| Uncorrected Total | 102 | 129159.41070 | |
| (Corrected Total) | 101 | 1693.48569 | |

| Parameter | Estimate | Asymptotic Std. Error | Asymptotic 95 % Confidence Interval | |
|-----------|-------------|-----------------------|-------------------------------------|--------------|
| | | | Lower | Upper |
| nugget | 44.48633706 | 2.7315704260 | 39.066276653 | 49.906397468 |
| sill | 75.98161994 | 0.6948686057 | 74.602841717 | 77.360398164 |
| range | 9.81337146 | 1.3724638501 | 7.090089313 | 12.536653607 |

Asymptotic Correlation Matrix

| Corr | nugget | sill | range |
|--------|--------------|--------------|--------------|
| nugget | 1 | 0.324952037 | 0.7800186738 |
| sill | 0.324952037 | 1 | 0.6386277355 |
| range | 0.7800186738 | 0.6386277355 | 1 |

The distribution of inter-centroid CD distances would be of interest. Figure (5.4) suggests a minimum of about 0.2 km.

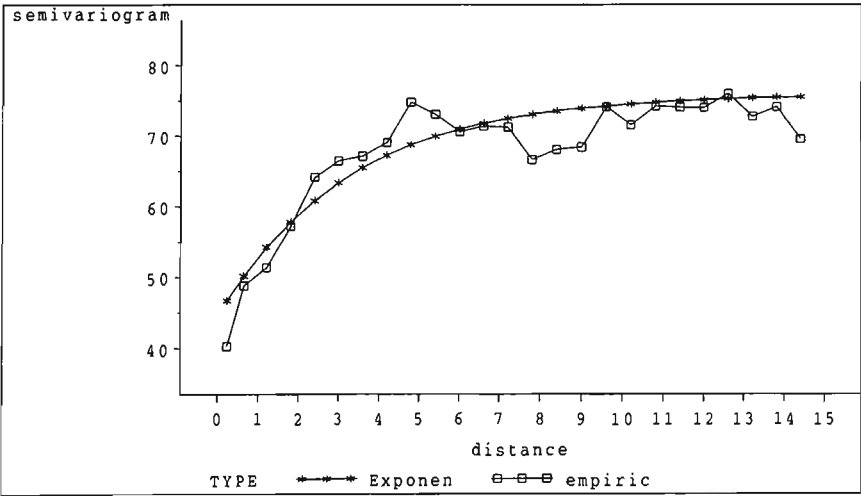


Figure 5.4. Empirical semivariogram and empirical exponential model with nugget 44.49, sill 75.98, and range 9.81

5.1.5 Relationship between semivariogram and spatial autocorrelation

For a second order stationary process, spatial autocorrelation was introduced in section (3.4.3) and explicitly formulated into

$$\rho(d_{ij}) = \frac{C(d_{ij})}{\sigma^2}$$

(5.19)

Theorem 5.1.4. *Relationship between semivariogram and spatial autocorrelation is*

$$\gamma(d_{ij}) = \sigma^2(1 - \rho(d_{ij})) \quad (5.20)$$

Proof. Assuming the isotropic condition and then substituting (5.19) into (5.8). \square

Given a semivariogram model, then the implied model of spatial autocorrelation can be derived using this relationship. The spatial autocorrelation model corresponding to (5.9), (5.10), and (5.11) are

- Exponential model :

$$\rho(d_{ij}) = \left(1 - \frac{n}{s}\right) \exp\left[\frac{-d_{ij}}{r}\right], \quad d_{ij} \geq 0 \quad (5.21)$$

- Spherical model :

$$\rho(d_{ij}) = \begin{cases} \left(1 - \frac{n}{s}\right) \left\{1 - \frac{3}{2} \frac{d_{ij}}{r} + \frac{1}{2} \left[\frac{d_{ij}}{r}\right]^3\right\} & 0 \leq d_{ij} \leq r, \\ 0 & d_{ij} > r. \end{cases} \quad (5.22)$$

- Gaussian model :

$$\rho(d_{ij}) = \left(1 - \frac{n}{s}\right) \exp\left[\frac{-d_{ij}^2}{r^2}\right], \quad d_{ij} \geq 0 \quad (5.23)$$

For the exponential and Gaussian models, we may again reparameterize the factor d/r by using $3d/r$.

There is a common factor in equation (5.21), (5.22), (5.23), that is

$$\left(1 - \frac{n}{s}\right) = \rho(0) \quad (5.24)$$

This quantity arises from the nugget effect, and will be discussed in (5.1.6). The typical spatial autocorrelation model can be drawn as in figure (5.5). The figure shows two quantities, those are the value of spatial autocorrelation at distance zero and the distance when the spatial autocorrelation becomes very small or zero. The first indicates nugget effect (5.24), and the second is the range.

For a given sill and nugget the greater the value of r will result higher spatial autocorrelation at closer distances. The spatial autocorrelation is getting smaller as the distance increases, because the factor involving distance in (5.21), (5.22) and (5.23) are all decreasing functions of d_{ij} .

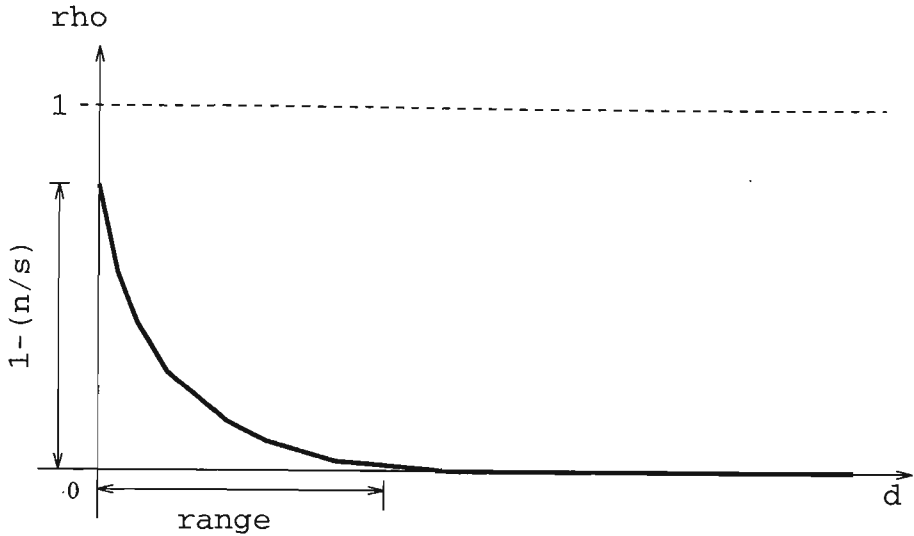


Figure 5.5. The positive side of spatial autocorrelation model based on parameters of semivariogram model

5.1.6 Nugget effect and spatial autocorrelation at zero distance

In social data, it is common to have several observations at the same location. For example, data from a household may contain observations on several different individuals, but the ℓ_i is recorded at the household location. Another situation is also can be considered, that the same locations may represents a group of households, e.g. apartment. This situation give a strong confidence that nugget effect may occur in social data. In both cases, the nugget effect will show co-variation of the observations within the same locations, either within one household or a group of households. This quantity will lead to considering the within household or a group of households correlation.

The relationship between nugget (n) and spatial autocorrelation may be used to consider the nugget effect.

Theorem 5.1.5. *Spatial autocorrelation at zero distance for exponential, spherical, and gaussian models is equal to*

$$\rho(0) = \left(1 - \frac{n}{s}\right) \tag{5.25}$$

Proof. Substituting $d_{ij} = 0$ into the models. □

Equation (5.25) shows the relation between $\rho(0)$ and the nugget value. The spatial correlation at zero distance will get smaller as the nugget value approaches the sill. The observations at zero distance will be perfectly independent if the nugget is equal to the sill. Hence, applying these semivariogram

models to individual level data implies a model for the within household correlation given by (5.25). More complex models for within household correlation are likely to be more realistic, however, without data on individuals within households to develop and estimate such models (5.25), is a useful working model.

5.1.7 Illustration of spatial autocorrelation from Illawarra dataset

The relationship between the semivariogram and spatial autocorrelation can be illustrated from the Illawarra dataset. Again the labor participation rate is taken as an example. The results from section (5.1.4) when the exponential model semivariogram was applied gave the estimated parameter nugget=44.49, sill=75.98, and range=9.81. The Moran coefficient (I) can be used to measure spatial autocorrelation (Cliff & Ord, 1981), and is defined

$$I = \frac{\sum_i \sum_j w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\hat{S}_{yy} \sum_i \sum_j w_{ij}} \quad (5.26)$$

This definition can be applied easily for the group level data. The w_{ij} indicates a connectivity matrix of the individual (or group) arrangement in the region, and in this application were defined as

$$w_{ij} = \begin{cases} 0 & \text{if } d_{ij} > \text{neighborhood distance} \\ 1 & \text{if } d_{ij} \leq \text{neighborhood distance} \end{cases} \quad (5.27)$$

where a neighborhood distance is a particular distance which the points within the distance are considered closed each other.

The Moran coefficient is used to measure spatial autocorrelation at a particular neighborhood distance. The *S+ Spatial Stats* package from S-Plus is used to calculate this statistic. The results are shown in table (5.2) and figure (5.6). Table (5.2) and figure (5.6) show the Moran coefficient and exponential model of the spatial autocorrelation such as defined in (5.21). For a particular neighborhood distance, e.g. h_1, h_2, \dots, h_k , the definition in (5.26) shows that the Moran coefficients for each neighborhood distances are defined within the interval 0 to the neighborhood distances. However the autocorrelation obtained from the semivariogram model refer to the distance classes $[0, 1), [1, 2)$, etc. Therefore we might expect a difference between the spatial autocorrelation defined by Moran coefficient and the semivariogram model.

Table 5.2. Moran coefficient of the labor participation rate of the Illawarra data at difference neighborhood distance

| Neighborhood distance (km) | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 7.0 | 9.0 | 10.0 |
|----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| Moran coefficient | 0.459 | 0.268 | 0.197 | 0.164 | 0.140 | 0.106 | 0.087 | 0.077 |
| The exponential model | 0.305 | 0.225 | 0.166 | 0.122 | 0.090 | 0.049 | 0.026 | 0.019 |

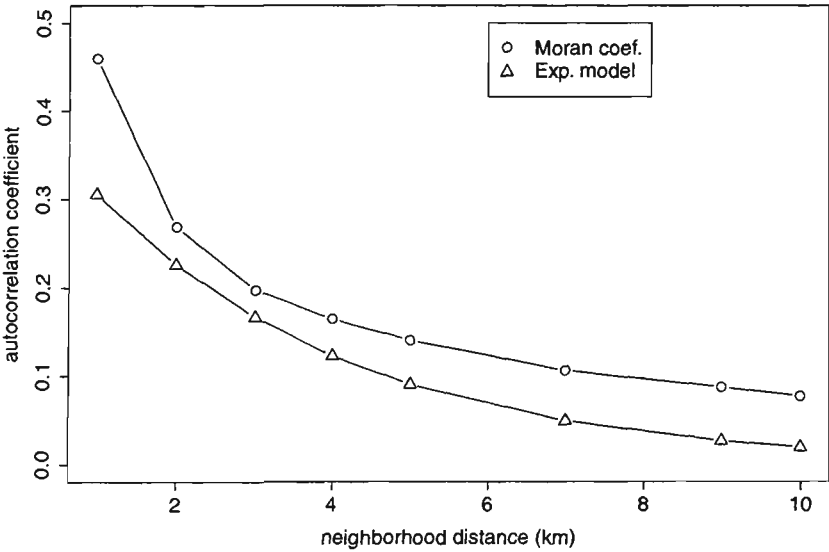


Figure 5.6. Moran coefficient and exponential model of spatial autocorrelation for the labor participation rate of the Illawarra data

Figure (5.6) shows that the Moran coefficient at neighborhood distance 1 km is a measure of spatial autocorrelation within the interval 0-1 km. The Moran coefficient at neighborhood distance 2 km indicates spatial autocorrelation within the interval 0-2 km, and the same way for another neighborhood distance. Meanwhile, the spatial correlation derived from the exponential model at distance 1 km indicates spatial correlation of points within the distance $[0, 1)$ km, and then $[1, 2)$ for the next interval, and so on. This description shows that the Moran coefficient cannot be compared directly with the exponential model, since they have a different definition of the distance except for the first interval where the results are similar. But indirectly both coefficients shows a typical trend of the spatial correlation, that is decaying as the distance increases.

5.1.8 Generating random observations based on semivariogram model

Simulation methods can be used to generate random observations according to a specific probability distribution function or other conditions. Arbia (1989a) discussed two methods of simulation of spatial data, that is distribution-based methods and model-based approaches. The first approach, that is distribution-based methods, will be applied to generate simulated data for this thesis.

The distribution-based approach generates random observations according to the distribution of the process under study, such as the random process defined in (4.1). We will generate observations from an isotropic, second order stationary spatial process. To generate observations from such a random process we need to define a variance-covariance matrix, which characterizes this random process. Since the random process is defined in two dimensional space, then we can use a spatial model. If the spatial model is specified in the form of a semivariogram model then the variance-covariance matrix can be derived from a relation between semivariogram and covariogram (5.8).

Let \mathbf{V} be the variance-covariance matrix of the random process generating Y_i (4.1). The elements of \mathbf{V} are equal to the $C(d_{ij})$ and σ^2 is defined to be the variance of the random process (\mathbf{Y}, \mathbf{L}) . For a given semivariogram model the covariances can be determined from the relation

$$C(d_{ij}) = \sigma^2 - \gamma(d_{ij}) \quad (5.28)$$

In a semivariogram model the σ^2 is equal to the sill of the model. The matrix \mathbf{V} is

$$\begin{pmatrix}
 \sigma^2 & C(d_{12}) & \cdots & C(d_{1j}) & \cdots & C(d_{1N}) \\
 C(d_{21}) & \sigma^2 & \cdots & \cdots & \cdots & C(d_{2N}) \\
 \vdots & \cdots & \sigma^2 & C(d_{ij}) & \cdots & \vdots \\
 C(d_{j1}) & \cdots & C(d_{ji}) & \sigma^2 & \cdots & C(d_{jN}) \\
 \vdots & \cdots & \cdots & \cdots & \sigma^2 & \vdots \\
 C(d_{N1}) & \cdots & \cdots & C(d_{Nj}) & \cdots & \sigma^2
 \end{pmatrix} \quad (5.29)$$

Arbia (1989a) noted that the distribution-based approach is based on decomposing the \mathbf{V} into lower and upper triangular matrix. The Choleski decomposition method can be applied to do this task, that is we find a matrix \mathbf{A} such that

$$\mathbf{A}\mathbf{A}^\top = \mathbf{V} \quad (5.30)$$

where, \mathbf{A} is a particular form of a lower triangular matrix.

Assume that \mathbf{e} is a vector of independent identically distributed standard normal random variables ($N(0, 1)$). Then the random process with variance-covariance matrix \mathbf{V} may be generated from \mathbf{e} by using the following relation,

$$\mathbf{Z} = \mathbf{A} \cdot \mathbf{e} \quad (5.31)$$

(see Arbia, 1989a).

5.2 Group level variogram ($\Gamma(d_{gh})$)

In some applications the group means constitute the main data available. Variogram analysis and modeling may be attempted using the groups means and the distances between the group, which will often be taken to be the distance between centroids, d_{gh} . The analysis of the Illawarra region in section (5.1.4) took this approach. Define a group level semivariogram value

$$\Gamma_{gh} = \frac{1}{2} V(\bar{Y}_g - \bar{Y}_h) \quad (5.32)$$

Theorem 5.2.1. *Consider an intrinsically stationarity process, then an unbiased estimator of Γ_{gh} , is*

$$\hat{\Gamma}_{gh} = \frac{1}{2}(\bar{Y}_g - \bar{Y}_h)^2 \quad (5.33)$$

Proof.

$$\bar{Y}_g - \bar{Y}_h = \frac{1}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} (Y_i - Y_j)$$

Hence for an intrinsically stationary process

$$E(\bar{Y}_g - \bar{Y}_h) = 0$$

Thus

$$\begin{aligned} E(\hat{\Gamma}_{gh}) &= \frac{1}{2} E((\bar{Y}_g - \bar{Y}_h)^2) \\ &= \frac{1}{2} V(\bar{Y}_g - \bar{Y}_h) = \Gamma_{gh} \end{aligned}$$

□

Theorem (5.2.1) implies that the results of section (4.3) concerning the empirical group level semivariogram can be applied to the group level semivariogram Γ_{gh} . For example theorems (4.3.4) and (4.3.5), are about the relationship with the unweighted and weighted group level variance.

We can derive a relationship between group level semivariogram and individual level semivariogram. For an intrinsically stationarity spatial process we have seen that $\hat{\gamma}_{ij}$ is unbiased for γ_{ij} (5.14) and $\hat{\Gamma}_{gh}$ is unbiased for Γ_{gh} , hence taking expectation of (4.64) gives :

Theorem 5.2.2. *The group level semivariogram can be expressed in term of individual level semivariogram*

$$\Gamma_{gh} = \bar{\gamma}_{gh} - \left(\frac{N_g - 1}{2N_g} \right) \bar{\gamma}_g - \left(\frac{N_h - 1}{2N_h} \right) \bar{\gamma}_h \quad (5.34)$$

where

$$\bar{\gamma}_{gh} = \frac{1}{N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} \gamma_{ij}; \quad \bar{\gamma}_g = \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \gamma_{ij}$$

Theorem (5.2.2) shows that the group level semivariogram can be expressed as a function of the individual level semivariogram. This result shows that Γ_{gh} will depend on the distances d_{ij} for pairs of

individuals within the same group and between pairs in which one individual is in group g and the other is in group h . Later we will consider how Γ_{gh} can be related to some measure of the distance between groups, such as d_{gh} or \bar{d}_{gh} .

For a second order stationary process, the group level semivariogram can be related to the individual level spatial autocorrelation. Assume that the population variance is constant at, σ^2 , and define the average spatial autocorrelation within the group as

$$\bar{\rho}_g = \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \rho(d_{ij}) \quad (5.35)$$

then

$$V(\bar{Y}_g) = \frac{\sigma^2}{N_g} (1 + (N_g - 1) \cdot \bar{\rho}_g) \quad (5.36)$$

And $Cov(\bar{Y}_g; \bar{Y}_h)$ may be rewritten as,

$$\begin{aligned} Cov(\bar{Y}_g; \bar{Y}_h) &= \sigma^2 - \bar{\gamma}_{gh} \\ &= \sigma^2 \bar{\rho}_{gh} \end{aligned} \quad (5.37)$$

where $\bar{\rho}_{gh} = \frac{1}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \rho(d_{ij})$ is the average spatial correlation between the individuals in the two groups.

Theorem 5.2.3. Relationship between Γ_{gh} and within group spatial autocorrelation ($\bar{\rho}_g$ and $\bar{\rho}_h$) is

$$\Gamma_{gh} = \bar{\gamma}_{gh} - \frac{\sigma^2}{2} \left(\frac{N_g - 1}{N_g} (1 - \bar{\rho}_g) + \frac{N_h - 1}{N_h} (1 - \bar{\rho}_h) \right) \quad (5.38)$$

or

$$\Gamma_{gh} = \sigma^2 (1 - \bar{\rho}_{gh}) - \frac{\sigma^2}{2} \left(\frac{N_g - 1}{N_g} (1 - \bar{\rho}_g) + \frac{N_h - 1}{N_h} (1 - \bar{\rho}_h) \right) \quad (5.39)$$

Proof. Substituting equation (5.20) into (5.34). □

Corollary 5.2.4. Assume that γ_{ij} is constant at σ^2 and there is no spatial autocorrelation, then

$$\Gamma_{gh} = \frac{\sigma^2}{2} \cdot \left(\frac{1}{N_g} + \frac{1}{N_h} \right) \quad (5.40)$$

Corollary 5.2.5. *Consider a constant within group spatial autocorrelation at $\bar{\rho}$, and a constant group size at \bar{N} then*

$$\Gamma_{gh} = \bar{\gamma}_{gh} - \sigma^2 \left(1 - \frac{1}{\bar{N}}\right) (1 - \bar{\rho}) \quad (5.41)$$

5.2.1 Fitting a group level semivariogram model

Assuming an isotropic process, the group level data can be used in the same way as the individual data to model $\hat{\Gamma}_{gh}$ in terms of the distance between the group as reflected by d_{gh} . The first step would be to calculate

$$\hat{\Gamma}(\bar{d}_f) = \frac{1}{|M_{D_f}|} \sum_{d_{gh} \in D_f}^{|M_{D_f}|} \hat{\Gamma}_{gh} \quad (5.42)$$

where D_f is a distance class used for the group level data and \bar{d}_f is the average of the between group distance for this class. The $|M_{D_f}|$ represents the numbers of pairs of groups which fall within the distance class D_f .

The next step is to fit a model to $\hat{\Gamma}(\bar{d}_f)$. There is no standard procedure to choose an appropriate model, but creating a scatter plot of $\hat{\Gamma}(\bar{d}_f)$ versus \bar{d}_f will be helpful in determining the appropriate model. Then a model fitting procedure can be used to estimate the parameters of the model, such as the nugget, sill, and range. The weighted least squares method can be considered as discussed in section (5.1.3), with the weight defined by (5.18), with M_{D_f} replacing N_{D_k} .

Analysis of the group level semivariogram models will produce estimates of the nugget, sill, and range. However, it is not clear how the estimates obtained from a group level analysis relate to the parameters of the underlying individual level semivariogram. From (5.34) and (5.38) we expect the average within group semivariogram or average within group spatial correlation to play a key role.

5.3 Deriving group level semivariogram in terms of individual level variogram by Taylor series expansion

Equation (5.34) gives the relationship between the group level and individual level semivariograms. However, this relationship does not immediately indicate how the analysis of Γ_{gh} is determined by γ_{ij} . In this section the relationship between Γ and γ will be investigated further by Taylor series expansion. Assume

that the individual level semivariogram is isotropic and represented by a distance function, $\gamma(d_{ij})$. Using a Taylor series expansion about the distance d_o , the distance function may be written as,

$$\gamma(d_{ij}) = \gamma(d_o) + \gamma'(d_o)(d_{ij} - d_o) + \frac{\gamma''(d_o)}{2}(d_{ij} - d_o)^2 + \dots + \frac{\gamma^{(p)}(d_o)}{p}(d_{ij} - d_o)^p + \mathbb{R}_p(d_{ij}) \quad (5.43)$$

where $\mathbb{R}_p(d_{ij})$ is the reminder of order p , which is expressed in terms of

$$\mathbb{R}_p(d_{ij}) = \frac{\gamma^{(p+1)}(c)}{(p+1)!} (d_{ij} - d_o)^{p+1}$$

the $\gamma^{(p+1)}(c)$ is the $(p+1)$ th derivative of the $\gamma(d_{ij})$ evaluated at c for some c between d_o and d_{ij} .

Suppose the $\gamma(d_{ij})$ is an exponential model with parameter n , s , and r . The likely value of $\mathbb{R}_p(d_{ij})$ can be illustrated as follow

$$\gamma(d_{ij}) = n + (s - n) \left(1 - \exp \left[\frac{-d_{ij}}{r} \right] \right)$$

$$\gamma'(d_{ij}) = \frac{s - n}{r} \exp \left[\frac{-d_{ij}}{r} \right]$$

$$\gamma''(d_{ij}) = -\frac{s - n}{r^2} \exp \left[\frac{-d_{ij}}{r} \right]$$

$$\gamma'''(d_{ij}) = \frac{s - n}{r^3} \exp \left[\frac{-d_{ij}}{r} \right]$$

\vdots

$$\gamma^{(2k)}(d_{ij}) = (-1)^{2k} \frac{s - n}{r^{2k}} \exp \left[\frac{-d_{ij}}{r} \right]$$

$$\gamma^{(2k+1)}(d_{ij}) = (-1)^{2k+1} \frac{s - n}{r^{2k+1}} \exp \left[\frac{-d_{ij}}{r} \right]$$

All the derivatives are bounded in magnitude $(s - n)$ and a constant $\frac{1}{r}$. Applying the Taylor's theorem and the Remainder Estimation theorem, we can determine

$$|\mathbb{R}_{2k+1}(d_{ij})| \leq (s - n) \left(\frac{d_{ij} - d_o}{r} \right)^{2k+2} \frac{1}{(2k+2)!} \cdot \exp \left[\frac{-c}{r} \right] \quad (5.44)$$

Provided the distance $d_{ij} - d_o$ is less than r , then the remainder term should be small, which will usually be the case in our applications. This $\mathbb{R}_{2k+1}(d_{ij}) \rightarrow 0$ as $k \rightarrow \infty$.

Consider (5.43) up to the second order and ignoring \mathbb{R} gives the approximation for $\gamma(d_{ij})$

$$\gamma(d_{ij}) \approx \gamma(d_o) + \gamma'(d_o)(d_{ij} - d_o) + \frac{\gamma''(d_o)}{2}(d_{ij} - d_o)^2 \quad (5.45)$$

Equation (5.45) can be used to derive an approximation of the group level variogram by substituting into equation (5.34). Equation (5.34) contains two main components, these are $\bar{\gamma}_{gh}$ and $\bar{\gamma}_g$ (or $\bar{\gamma}_h$). These components represent the average of the between and within group semivariogram values respectively.

Evaluating $\bar{\gamma}_{gh}$

Substituting approximation (5.45) into $\bar{\gamma}_{gh}$, we get

$$\begin{aligned} \bar{\gamma}_{gh} &= \gamma(d_o) + \frac{\gamma'(d_o)}{N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} (d_{ij} - d_o) + \frac{\gamma''(d_o)}{2N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} (d_{ij} - d_o)^2 \\ &= \gamma(d_o) + \gamma'(d_o)(\bar{d}_{gh} - d_o) + \frac{\gamma''(d_o)}{2N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} (d_{ij} - \bar{d}_{gh})^2 + \frac{\gamma''(d_o)}{2}(\bar{d}_{gh} - d_o)^2 \\ &= \gamma(d_o) + \gamma'(d_o)(\bar{d}_{gh} - d_o) + \frac{\gamma''(d_o)}{2}S_{d_{gh}}^2 + \frac{\gamma''(d_o)}{2}(\bar{d}_{gh} - d_o)^2 \end{aligned} \quad (5.46)$$

where \bar{d}_{gh} and $S_{d_{gh}}^2$ indicate the mean and variance of the distance between all pairs of individuals between the two groups. The term \bar{d}_{gh} is defined in (5.7) and $S_{d_{gh}}^2$ is defined as

$$S_{d_{gh}}^2 = \frac{1}{N_g N_h} \sum_{\substack{i \in \mathcal{U}_g \\ j \in \mathcal{U}_h}} (d_{ij} - \bar{d}_{gh})^2$$

In Equation (5.46), the value d_o represents some chosen distance to represent the distance between two non overlapping groups. Two different choices for d_o seem sensible. First, d_o is the average distance of all pair of individuals between the two groups, d_{gh} , giving

$$\bar{\gamma}_{gh} = \gamma(\bar{d}_{gh}) + \frac{\gamma''(\bar{d}_{gh})}{2}S_{d_{gh}}^2 \quad (5.47)$$

Often we will not have \bar{d}_{gh} , but instead have the distance between the centroids of the two groups, say $d_{gh} = \|\ell_g - \ell_h\|$. Substituting d_{gh} for d_o gives

$$\bar{\gamma}_{gh} = \gamma(d_{gh}) + \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) + \frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2 + \frac{\gamma''(d_{gh})}{2} (\bar{d}_{gh} - d_{gh})^2 \quad (5.48)$$

If $d_{gh} = \bar{d}_{gh}$ then the two choices for d_o will give the same result for $\bar{\gamma}_{gh}$. In practice, the centroid location is often recorded but not the individuals' locations. Therefore, we will usually have a representation of d_{gh} in term of distance between two centroids rather than the average distance of all pairs of locations. The centroid of the group depends on the boundary of the group, since the centroid is derived from manipulation of the boundary coordinates (Griffith & Amrhein, 1991).

Evaluating $\bar{\gamma}_g$

In the same way, we can formulate an approximation for $\bar{\gamma}_g$, that is

$$\begin{aligned} \bar{\gamma}_g &= \gamma(d'_o) + \frac{\gamma'(d'_o)}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} (d_{ij} - d'_o) + \frac{\gamma''(d'_o)}{2N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} (d_{ij} - d'_o)^2 \\ &= \gamma(d'_o) + \gamma'(d'_o)(\bar{d}_g - d'_o) + \frac{\gamma''(d'_o)}{2N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} (d_{ij} - \bar{d}_g)^2 + \frac{\gamma''(d'_o)}{2} (\bar{d}_g - d'_o)^2 \\ &= \gamma(d'_o) + \gamma'(d'_o)(\bar{d}_g - d'_o) + \frac{\gamma''(d'_o)}{2} S_{d_g}^2 + \frac{\gamma''(d'_o)}{2} (\bar{d}_g - d'_o)^2 \end{aligned} \quad (5.49)$$

Taking $d'_o = \bar{d}_g$, the average distance of all pairs of individuals within the group, that is defined in (5.7) gives

$$\bar{\gamma}_g = \gamma(\bar{d}_g) + \frac{\gamma''(\bar{d}_g)}{2} S_{d_g}^2 \quad (5.50)$$

The terms \bar{d}_g and $S_{d_g}^2$ indicate the mean and variance of the distance of all pairs of individuals within the group.

5.3.1 Approximation of $\Gamma(d_{gh})$

Given these approximations for $\bar{\gamma}_{gh}$ and $\bar{\gamma}_g$ ($\bar{\gamma}_h$), then we may substitute into (5.34) to give an approximation for the Γ_{gh} , say $\tilde{\Gamma}_{gh}$, that is

$$\begin{aligned} \tilde{\Gamma}_{gh} = & \gamma(d_{gh}) \\ & + \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) + \frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2 + \frac{\gamma''(d_{gh})}{2} (\bar{d}_{gh} - d_{gh})^2 \\ & - \frac{N_g - 1}{2N_g} \left[\gamma(\bar{d}_g) + \frac{\gamma''(\bar{d}_g)}{2} S_{d_g}^2 \right] - \frac{N_h - 1}{2N_h} \left[\gamma(\bar{d}_h) + \frac{\gamma''(\bar{d}_h)}{2} S_{d_h}^2 \right] \end{aligned} \quad (5.51)$$

Consider the semivariogram models, the exponential and Gaussian model. The value of γ , γ' , and γ'' are calculated based on the results of simulation of the properties of the distance in section (5.3.2) of page 99. The simulated average distance within and between groups are run over 1200 times of simulation, based on the population of individual level semivariogram with $n = 15$, $s = 25$, and $r = 10$. Figure (5.7) shows $\gamma'(\bar{d}_g)$, $\gamma''(\bar{d}_g)$, $\gamma'(d_{gh})$, and $\gamma''(d_{gh})$ for different semivariogram models. The figure shows that the second derivative (γ'') of the three models are relatively small compared with the first derivative, either for the within group or between group component. The exponential and Gaussian models show a small value of their first and second derivative.

Some expected points from this figures are

- γ'' is small compared with γ' , which is small compared with γ , which has a minimum of 15 and maximum of 25. Hence ignoring the first and second derivative will not introduce much error.
- also, the factor $(\bar{d}_{gh} - d_{gh})$ and $S_{d_{gh}}^2$, $S_{d_g}^2$ are likely to further reduce the impact of the first and second order terms.

The bias components

Equation (5.51) shows clearly how the group level semivariogram is related to the individual level semivariogram, and the average and variance of distances between points between groups, and average and variance of distance between points within groups. Equation (5.51) shows that the bias of $\hat{\Gamma}_{gh}$ as an estimate of $\gamma(d_{gh})$ has two component of bias, the first is;

$$+ \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) + \frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2 + \frac{\gamma''(d_{gh})}{2} (\bar{d}_{gh} - d_{gh})^2 \quad (5.52)$$

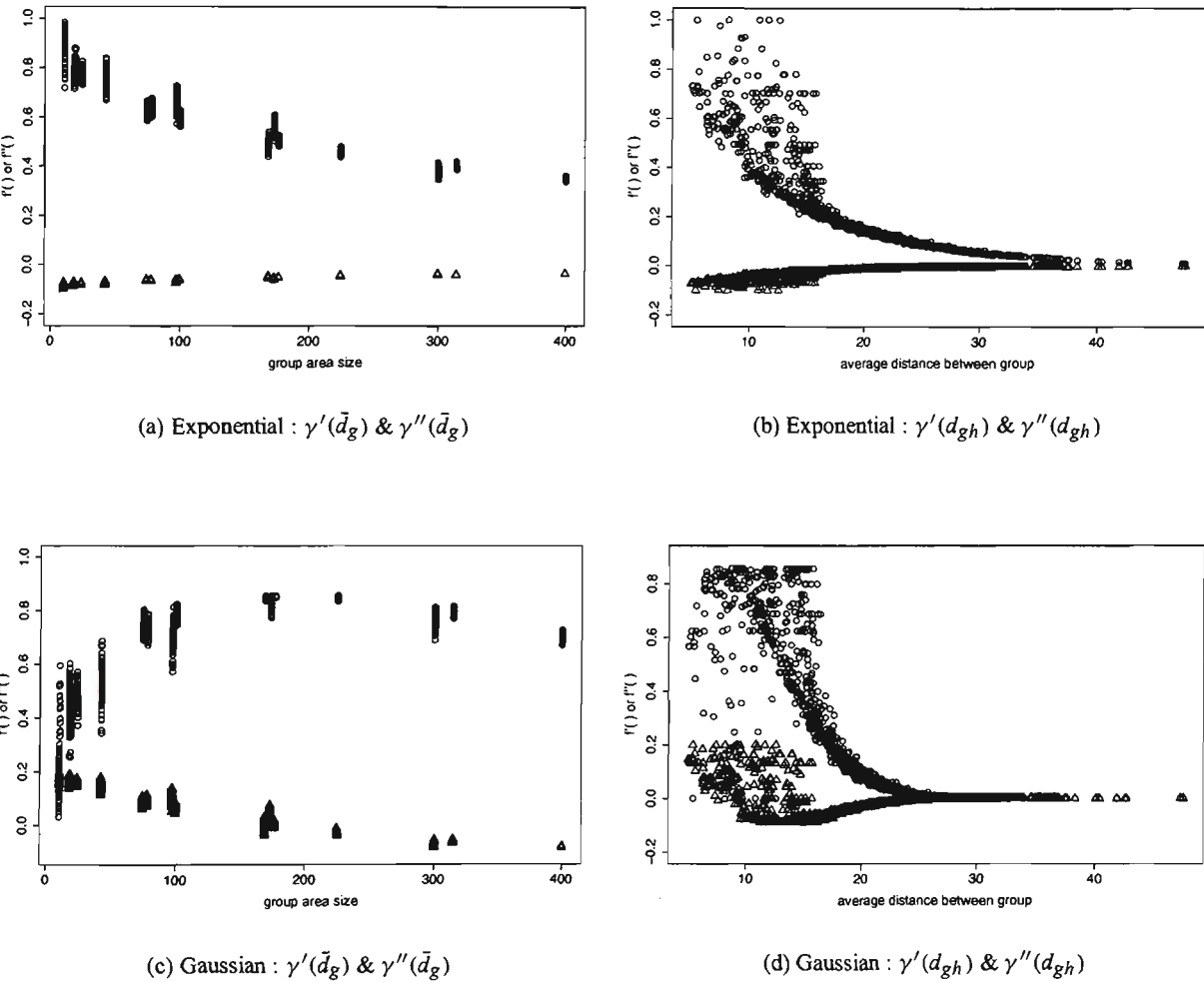


Figure 5.7. The likely relative value of γ' and γ'' compared with the average distance between group and group area size, the symbol $\circ = \gamma'$, and $\Delta = \gamma''$.

This component of bias involves the difference of the true average distance (\bar{d}_{gh}) and the distance between centroids (d_{gh}), and the variation of distance between two groups ($S_{d_{gh}}^2$). The difference between \bar{d}_{gh} and d_{gh} will depend on how points are distributed within the groups.

The second component of bias is,

$$-\frac{N_g - 1}{2N_g} \left[\gamma(\bar{d}_g) + \frac{\gamma''(\bar{d}_g)}{2} S_{d_g}^2 \right] - \frac{N_h - 1}{2N_h} \left[\gamma(\bar{d}_h) + \frac{\gamma''(\bar{d}_h)}{2} S_{d_h}^2 \right] \quad (5.53)$$

The second component of bias produces a negative effect. By considering a particular shape of the group (for instance circle, square, rectangle, etc), approximation can be developed for this component. Given the result on the values of first and second derivatives, we expect that the main bias component will be

$$-\frac{1}{2} \left(\frac{N_g - 1}{N_g} \gamma(\bar{d}_g) + \frac{N_h - 1}{N_h} \gamma(\bar{d}_h) \right) \quad (5.54)$$

This will give a negative bias and so the group level semivariogram will be below the corresponding individual level semivariogram. An illustration of the bias is given in section (5.5.1) using simulated data.

A knowledge of the probability density function of the distance of all pairs of individuals within the group will be useful in evaluating $\bar{\gamma}_g$ and $S_{d_g}^2$ closely by giving an idea of the \bar{d}_g , \bar{d}_h , $S_{d_g}^2$, and $S_{d_h}^2$ values. We can also develop an approximation for the \bar{d}_{gh} and $S_{d_{gh}}^2$ for a particular shape of the group (section 5.3.2).

Some cases

Corollary 5.3.1. *Consider groups that are very small so that $\gamma(d)$ within the groups are approximately constant at $\gamma(0)$, then we have*

$$\begin{aligned} \Gamma_{gh} \approx \tilde{\Gamma}_{gh} = & \gamma(d_{gh}) \\ & + \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) + \frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2 + \frac{\gamma''(d_{gh})}{2} (\bar{d}_{gh} - d_{gh})^2 \\ & - \gamma(0) \left(1 - \frac{1}{2} \left[\frac{1}{N_g} + \frac{1}{N_h} \right] \right) \end{aligned} \quad (5.55)$$

This case indicates the role that the individual level nugget plays in this relationship. The nugget, $\gamma(0)$, will take a positive value, and the whole last term of (5.55) will be negative.

Consider the special case that there is no spatial relationship at the individual level, for example that $\gamma(d)$ is constant at σ^2 . This situation implies the first and second derivative of $\gamma()$ is zero, then we may

simplify equation (5.55),

$$\tilde{\Gamma}_{gh} = \gamma(d_{gh}) - \gamma(0) \left(1 - \frac{1}{2} \left[\frac{1}{N_g} + \frac{1}{N_h} \right] \right) \quad (5.56)$$

and

$$\tilde{\Gamma}_{gh} = \frac{1}{2} \sigma^2 \left(\frac{1}{N_g} + \frac{1}{N_h} \right) \quad (5.57)$$

While this case is not usually realistic, it may suggest a weighting approach (see section 5.5).

Theorem (5.2.2) suggested that the group level semivariogram is less than the corresponding individual level semivariogram. Equation (5.56) shows explicitly in this case the position of group level semivariogram compared with the individual level semivariogram. It shows that the group level semivariogram is always less than or equal to the individual level semivariogram. This situation is shown in the following graph (Figure 5.8).

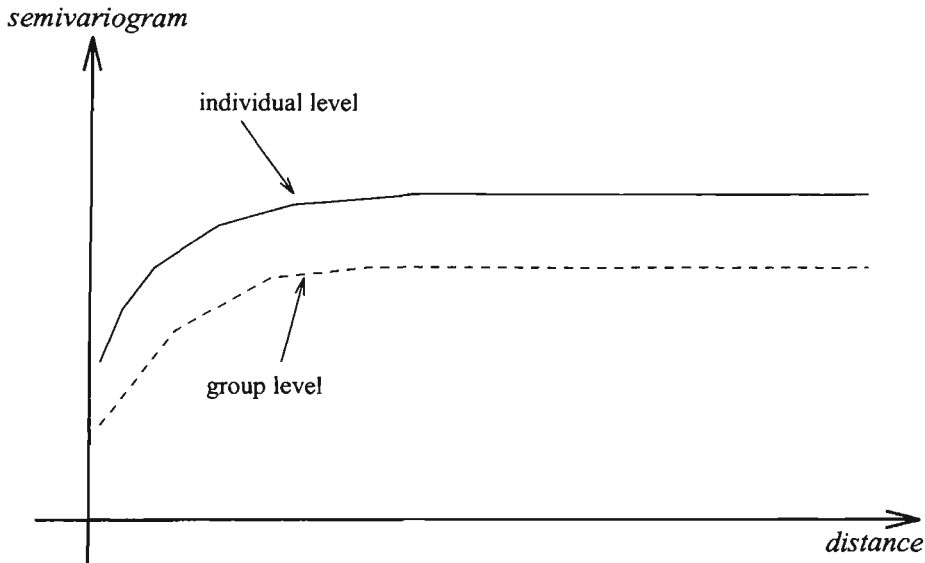


Figure 5.8. Group level semivariogram and its individual level semivariogram

5.3.2 Approximation of moment structure of random distance within and between groups.

Probability density function for distance of pairs of points in a convex region

Equation (5.51) shows that the bias in using the group level semivariogram to estimate the individual level semivariogram depends on the mean and variance of the distribution of distances between points within and between groups. We can investigate the effect of aggregation on variogram analysis by considering the

distribution of distances between points. Assume that the points are randomly distributed in a particular region. The random distance within a particular region has been derived for a range of applications, such as in forestry, biology, and ecology (Ghosh,1951; Bartlett,1964; Matérn,1986). Different shapes of the region have been considered, such as circle, rectangle, square, triangle, and hexagon.

Rectangle and square

Ghosh (1951) derived the probability density function (*pdf*) of the random distance (d) within a rectangle with the sides of lengths a and b , for $a \geq b$. Denote w as the ratio of the two sides of a rectangle, $w = \frac{a}{b}$ for $a \geq b$, and A is area of the rectangle, $A = a \cdot b$. The *pdf* is given by

$$f_r(d) = \frac{4d}{A^2} \cdot g_r(d) \quad (5.58)$$

where $g_r(d)$ is defined for three intervals of d ,

$$\begin{aligned} g_r(d) &= 0.5\pi A - \frac{d\sqrt{A}}{\sqrt{w}} \cdot (w+1) + 0.5 \cdot d^2, \quad \text{for } 0 \leq d \leq \sqrt{\frac{A}{w}} \\ g_r(d) &= A \cdot \arcsin\left(\frac{\sqrt{Aw}}{d}\right) + [Aw d^2 - A^2]^{\frac{1}{2}} - d\sqrt{Aw} - 0.5 \cdot \frac{A}{w}, \quad \text{for } \sqrt{\frac{A}{w}} \leq d \leq \sqrt{Aw} \\ g_r(d) &= A \cdot \left\{ \arcsin\left(\frac{1}{d} \left[\frac{A}{w}\right]^{\frac{1}{2}}\right) - \arccos\left(\frac{\sqrt{Aw}}{d}\right) \right\} + [Aw d^2 - A^2]^{\frac{1}{2}} + \left[\frac{Ad^2}{w} - A^2\right]^{\frac{1}{2}} \\ &\quad - 0.5 \cdot \left(d^2 + Aw + \frac{A}{w}\right), \quad \text{for } \sqrt{Aw} \leq d \leq \left[Aw + \frac{A}{w}\right]^{\frac{1}{2}} \end{aligned}$$

The probability density function of the random distance within a square region can be obtained as a special case of equation (5.58) when $w = 1$, that is $a = b$. This *pdf* was discussed by Bartlett (1964) and it can be formulated as

$$f_s(d) = \frac{4d}{A^2} \cdot g_s(d) \quad (5.59)$$

where $g_s(d)$ is defined within two intervals of d ,

$$\begin{aligned} g_s(d) &= \frac{\pi A}{2} - \frac{d}{\sqrt{A}} + \frac{d^2}{2}, \quad \text{for } 0 \leq d \leq \sqrt{A}, \\ g_s(d) &= A \cdot \left\{ \arcsin\left(\frac{\sqrt{A}}{d}\right) - \arccos\left(\frac{\sqrt{A}}{d}\right) \right\} + 2\sqrt{A} \cdot [d^2 - A]^{\frac{1}{2}} \\ &\quad - \frac{1}{2} \cdot (d^2 + 2A), \quad \text{for } \sqrt{A} \leq d \leq \sqrt{2A} \end{aligned}$$

Circle

Bartlett (1964) also defined the probability density function of the random distance within the circle of radius R . The *pdf* can be expressed in terms of the area (A) of the circle and is

$$f_c(d) = \frac{4d}{A} \cdot \left\{ \arccos\left(\frac{d\sqrt{\pi}}{2\sqrt{A}}\right) - \frac{d\sqrt{\pi}}{4\sqrt{A}} \cdot \left[4 - \frac{d^2\pi}{A}\right]^{\frac{1}{2}} \right\}, \quad \text{for } 0 \leq d \leq 2\sqrt{A/\pi} \quad (5.60)$$

Hexagon and equilateral triangle

Matérn (1986) discussed the probability density functions of the random distance within a regular hexagon and equilateral triangle. The regular hexagon with sides s has area

$$A_h = \frac{3}{2}s^2\sqrt{3} \quad (5.61)$$

The equilateral triangle with sides s has area

$$A_t = \frac{1}{4}s^2\sqrt{3} \quad (5.62)$$

Define ν as the ratio of the random distance d over the side s , then the probability density function for the regular hexagon (see Matérn, 1986) is

$$f_h(\nu) = \frac{4}{9\sqrt{3}} \cdot \nu \cdot g_h(\nu) \quad (5.63)$$

where $g_h(\nu)$ is defined in three intervals and is zero otherwise,

$$g_h(\nu) = 3\pi - 4\nu\sqrt{3} + \nu^2 \cdot \left(\sqrt{3} - \frac{\pi}{3}\right), \quad \text{for } 0 \leq \nu \leq 1$$

$$g_h(\nu) = \pi(5 + \nu^2) - 3 \left[12\nu^2 - 9\right]^{\frac{1}{2}} - (4\nu^2 + 6) \cdot \arcsin\left(\frac{\sqrt{3}}{2\nu}\right), \quad \text{for } 1 \leq \nu \leq \sqrt{3}$$

$$g_h(d) = (2\nu^2 + 24) \cdot \left\{ \arcsin\left(\frac{\sqrt{3}}{\nu}\right) - \frac{\pi}{2} \right\} - \sqrt{3} \cdot (\nu^2 + 6) \\ + 10 \left[3\nu^2 - 9\right]^{\frac{1}{2}}, \quad \text{for } \sqrt{3} \leq \nu \leq 2$$

The probability density function of the equilateral triangle is,

$$f_t(\nu) = \frac{8}{\sqrt{3}} \cdot (g_t(\nu) + h_t(\nu)) \quad (5.64)$$

where $g_t(\nu)$ and $h_t(\nu)$ are

$$g_t(\nu) = \begin{cases} \pi - 4\nu\sqrt{3} + \nu^2 \left(\sqrt{3} + \frac{2\pi}{3}\right) & 0 \leq \nu \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$h_t(v) = \begin{cases} 3 \left[12v^2 - 9 \right]^{\frac{1}{2}} - (4v^2 + 6) \arccos \left(\frac{\sqrt{3}}{2v} \right) & \frac{\sqrt{2}}{2} \leq v \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Appendix (C) give *Maple* codes for evaluating these probability density functions. The mean and variance are calculated for a rectangle, square, circle, regular hexagon, and equilateral triangle, each of unit area. The calculation of the mean and variance of the distances were done analytically by evaluating the integral of the *pdf* of the particular group shape. The results are shown in table (5.3.2).

Table 5.3. The mean and variances of random distances within a unit area

| Shape of the region | Perimeter | Mean | Variance |
|----------------------|-----------|---------|----------|
| Circle | 3.54491 | 0.51083 | 0.05737 |
| Regular hexagon | 3.72242 | 0.51261 | 0.05798 |
| Square | 4.00000 | 0.52141 | 0.06147 |
| Equilateral triangle | 4.55901 | 0.55436 | 0.07758 |
| Rectangle w=2 | 4.24264 | 0.56906 | 0.09284 |
| Rectangle w=4 | 5.00000 | 0.71374 | 0.19890 |
| Rectangle w=6 | 5.71546 | 0.84636 | 0.31145 |

These result shows a positive relationship between perimeters of the shape at the unit area with the mean and variance of random distance. The result are clearly shown in the figure (5.9).

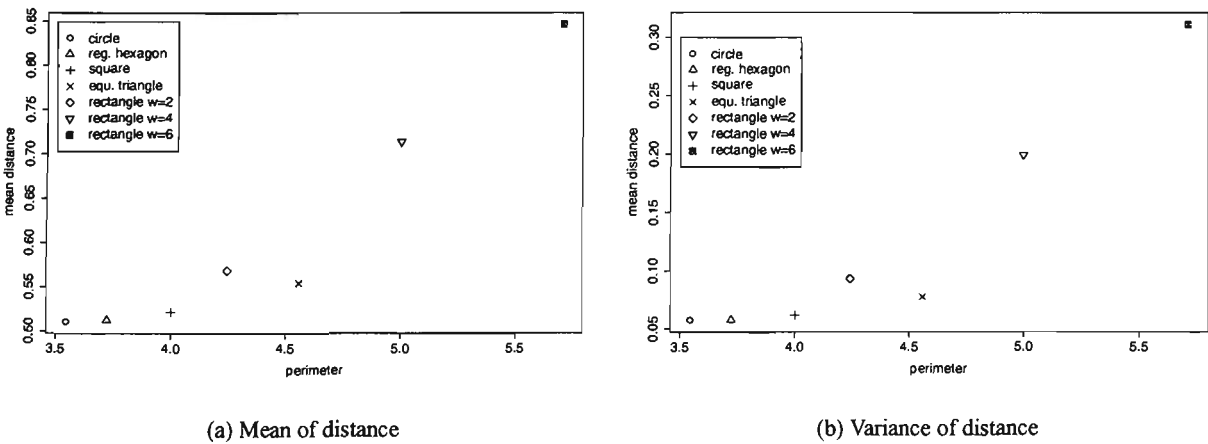


Figure 5.9. Relationship of the mean and variance of the random distance with perimeter of the unit area

These are similar to the results presented by Matérn (1986), who noted that the moment structures of random distance within a region is affected by the length of region’s perimeter. The mean and variance

of the random distance increase in the order as circle, regular hexagon, square, equilateral triangle and rectangle with $w = 2$ (Figure 5.9). The difference of the mean or variance among these shapes is small, which Matérn (1986) regarded as insignificance differences. However, for a rectangular shape the mean and variance are appreciatively larger when w becomes large, corresponding to very elongated shapes.

A further investigation is done to relate the mean and variance of the random distance to the area of the region. Based on the previous calculation, the mean and variance of the distance within the group can be computed at a particular area. The area of the groups were set from 1 to 20 square unit. The result are shown in Figure (5.10).

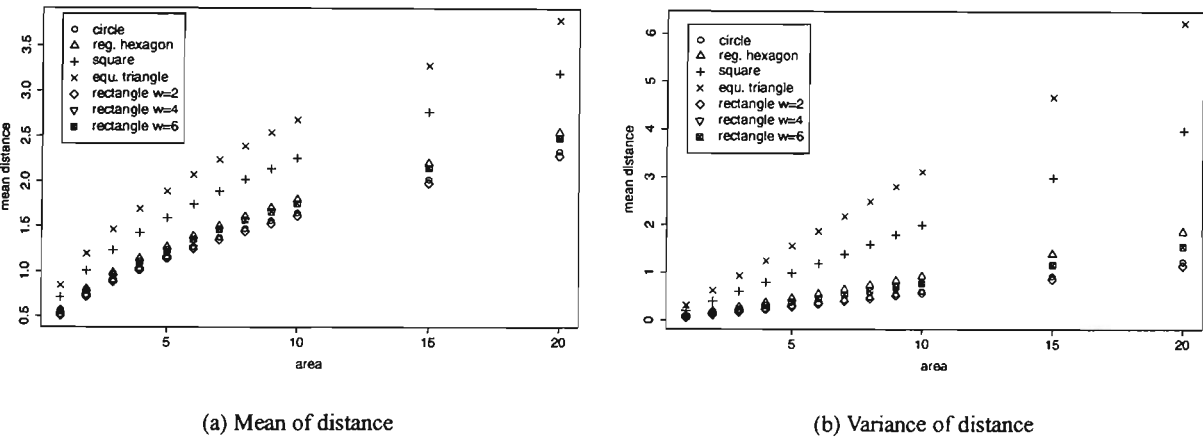


Figure 5.10. Relationship of the mean and variance of the random distance within the region with area of different shapes of the region

For a given shape, the larger area will give a larger mean and variance of the random distance within the region. Figure (5.10) shows that the mean and variance of random distance within the region increase with the area of the region. The expectation and variance of the random distance within the region are related to the area A as follows;

$$E(d_{ij}) \propto \sqrt{A}; \quad \text{and} \quad V(d_{ij}) \propto A \tag{5.65}$$

The plot in figure (5.10) shows this, and hence we can write

$$E(d_{ij}) = k_1 \sqrt{A}; \quad \text{and} \quad V(d_{ij}) = k_2 A \tag{5.66}$$

where k_1 and k_2 depend on the shape of the region. The values of k_1 and k_2 are given in table (5.3.2). The next section will look at an approximation of the values k_1 and k_2 , which are applied for all the shapes

considered. An interesting point can be observed when comparing between shape of the region. The circle, square, regular hexagon, rectangle with $w = 2$, and equilateral triangle have similar mean and variance values at different area.

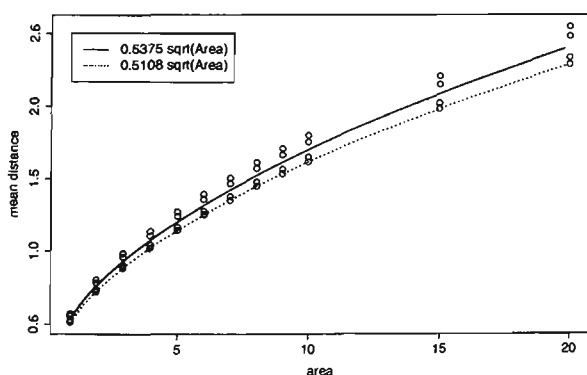
Approximation for \bar{d}_g and $S_{d_g}^2$

An approximation for the mean and variance of distances within a group can be found using a regression method, with the mean of distance or variance of distance as dependent variable. Excluding the rectangle with $w = 4$ and $w = 6$, the regression line can be computed

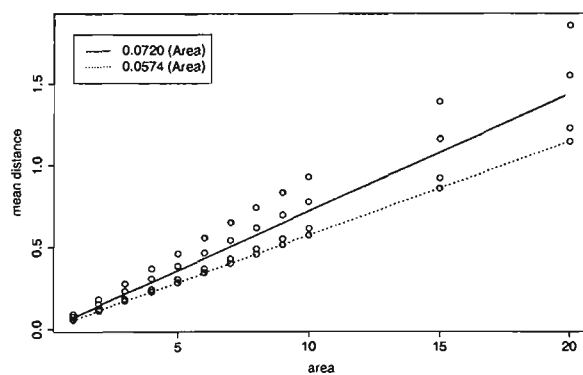
$$E(d) \approx 0.537465 \cdot \sqrt{A}, \quad (\text{residual square error} = 0.06377899) \quad (5.67)$$

$$V(d) \approx 0.071998 \cdot A, \quad (\text{residual square error} = 0.1240943)$$

These approximations are drawn in Figure (5.11) with a solid lines, and effectively average values of k_1 and k_2 for these shapes, with have reasonably similar values of k_1 and k_2 .



(a) Relationship of mean distance and area



(b) Relationship of variance of distance area

Figure 5.11. Approximation of the mean and variance of random distance within the region by regression method

Another approximation was suggested by Matérn (1986) that was based on the probability density function for the distance for pairs of individuals within the circle. Matérn (1986) indicated that in a convex and small area the moment structure of the random distance can be approximated by the moment structure of the random distance of the circle. This approximation is shown as a dash line in Figure (5.11) and can be defined analytically as follow

$$E(d) \approx k_1 \sqrt{A} \quad \text{and} \quad V(d) \approx k_2 \cdot A \quad (5.68)$$

where

$$k_1 = \frac{128}{45\pi\sqrt{\pi}} = 0.51082 \dots \quad \text{and} \quad k_2 = \left(\frac{1}{\pi} - \frac{128^2}{45^2\pi^3} \right) = 0.05736 \dots$$

These values agree with those given in table (5.3.2) for the case of a circle.

Figure (5.11) shows that the approximation (5.68) is an underestimate of the mean or the variance of random distance between points within regions that are not circles. But it could be accepted for regions with small area. The figure also indicates that for a small area these two approximations will be very close. In the following discussion we will use approximation (5.68), such as suggested by Matérn (1986).

Approximations of \bar{d}_{gh} and $S_{d_{gh}}^2$

This approximation can be found described briefly in Vaughan (1984). However, not all the details of the derivation of the approximation are presented. Therefore in this section, details are presented and some improvements of the approximation are made.

Consider the g th and h th groups are arranged as in Figure (5.12), which have area A and B , respectively. Assume the centroid of the first group is located at coordinate $(0, 0)$, and the second group's centroid is located at $(d_{gh}, 0)$. The distance between the two centroids is d_{gh} . One point is located in the g th group, that is p_a and one point is located in the h th group, p_b . The relative location of p_a and p_b can be defined as the coordinate (x, y) and $(d_{gh} + u, v)$, respectively. Therefore the distance between the two points p_a and p_b is

$$d = \left[(d_{gh} + u - x)^2 + (v - y)^2 \right]^{\frac{1}{2}} \quad (5.69)$$

The mean distance between the two groups is defined as the expected valued of the direct distance between a random point selected in the g th region and a random point selected in the h th region, that is

$$E(d) = \int_A \int_B \frac{d}{A \cdot B} dA dB \quad (5.70)$$

The approximation is proceeded by modifying equation (5.69) to

$$d = d_{gh} \cdot \left[1 + \left[\frac{2p}{d_{gh}} + \frac{q^2}{d_{gh}^2} \right] \right]^{\frac{1}{2}} \quad (5.71)$$

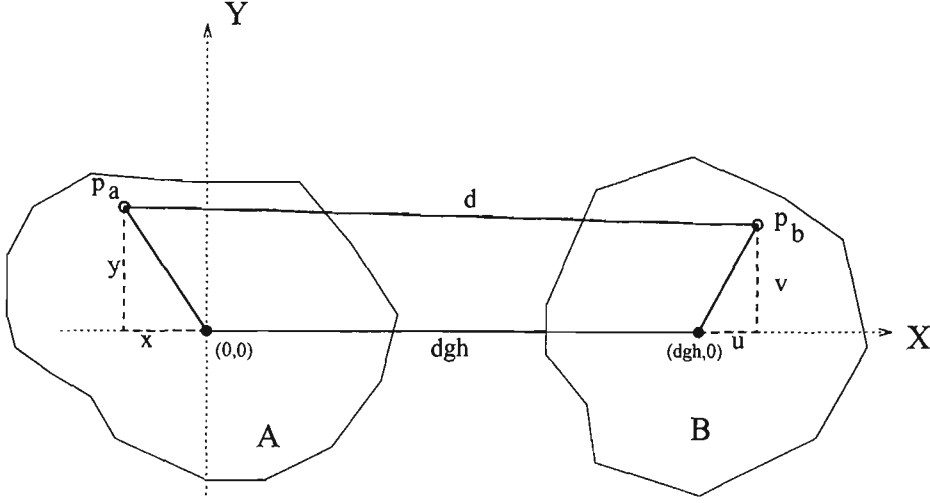


Figure 5.12. The random distance between points in two groups

where

$$p = (u - x) \quad \text{and} \quad q^2 = (u - x)^2 + (v - y)^2$$

Applying Taylor series expansion to (5.71) then

$$d = d_{gh} \cdot \left(1 + \frac{1}{2}t - \frac{1}{8}t^2 + \frac{1}{16}t^3 - \frac{5}{128}t^4 + O(t^5) \right) \quad (5.72)$$

where

$$t = \left(\frac{2p}{d_{gh}} + \frac{q^2}{d_{gh}^2} \right)$$

Taking the $\frac{1}{d_{gh}}$ factor out of (5.72)

$$\begin{aligned} d &= d_{gh} \cdot \left[1 + \frac{p}{d_{gh}} + \frac{1}{2d_{gh}^2}(q^2 - p^2) + \frac{1}{2d_{gh}^3}(p^3 - pq^2) + \frac{1}{8d_{gh}^4}(6p^2q^2 - 5p^4 - q^4) \right. \\ &\quad + \frac{1}{8d_{gh}^5}(3pq^4 - 10p^3q^2) + \frac{1}{16d_{gh}^6}(q^6 - 15p^2q^4) \\ &\quad \left. - \frac{5}{16d_{gh}^7}pq^6 - \frac{5}{128d_{gh}^8}q^8 + O\left(\left[\frac{2p}{d_{gh}} + \frac{q^2}{d_{gh}^2}\right]^5\right) \right] \end{aligned} \quad (5.73)$$

Assume that the d_{gh} are relatively larger than p and q^2 , such that the fifth and greater order of d_{gh} may be ignored, then the approximation of d can be defined as

$$\begin{aligned} d \approx \tilde{d} &= d_{gh} \cdot \left[1 + \frac{p}{d_{gh}} + \frac{1}{2d_{gh}^2}(q^2 - p^2) + \frac{1}{2d_{gh}^3}(p^3 - pq^2) \right. \\ &\quad \left. + \frac{1}{8d_{gh}^4}(6p^2q^2 - 5p^4 - q^4) \right] \end{aligned} \quad (5.74)$$

or

$$\begin{aligned}
 \tilde{d} = & d_{gh} \cdot \left[1 + \frac{u-x}{d_{gh}} + \frac{1}{2d_{gh}^2} (v^2 - 2vy + y^2) \right. \\
 & + \frac{1}{2d_{gh}^3} (xy^2 + xv - 2xyv - uv^2 + 2uvy - uy^2) \\
 & + \frac{1}{8d_{gh}^4} (4u^2y^2 + 4u^2v^2 + 4x^2y^2 + 4x^2v^2 - 6v^2y^2 - 8u^2vy - 8uxy^2 \\
 & \left. - 8x^2yv - 8uv^2x + 4vy^3 + 4v^3y - v^4 - y^4 + 16uvxy) \right]
 \end{aligned} \tag{5.75}$$

Vaughan (1984) defined the moments of continuous variables $\{x_1, \dots, x_i\}$ and $\{y_1, \dots, y_j\}$ within the region with area A as

$$A_{x_1x_2\dots x_i y_1y_2\dots y_j} = \int_A \frac{x^i y^j}{A} dA \tag{5.76}$$

This definition is analog with the moments of the discrete variables.

Substituting (5.75) into (5.70) and applying (5.76), then we can determine the approximation of \bar{d}_{gh} as

$$\begin{aligned}
 \bar{d}_{gh} \approx E(\tilde{d}) = & d_{gh} \cdot \left[1 + \frac{B_u - A_x}{d_{gh}} + \frac{1}{2d_{gh}^2} (B_{vv} - 2B_v A_y + A_{yy}) \right. \\
 & + \frac{1}{2d_{gh}^3} (A_{xy} + A_x B_v - 2A_{xy} B_v - B_{uvv} + 2B_{uv} A_y - B_u A_{yy}) \\
 & + \frac{1}{8d_{gh}^4} (4B_{uu} A_{yy} + 4B_{uuvv} + 4A_{xxyy} + 4A_{xx} B_{vv} - 6B_{vv} A_{yy} \\
 & - 8B_{uuv} A_y - 8B_u A_{xyy} - 8A_{xxy} B_v - 8B_{uvv} A_x \\
 & \left. + 4B_v A_{yyy} + 4B_{vvv} A_y - B_{vvvv} - A_{yyyy} + 16B_{uv} A_{xy}) \right]
 \end{aligned} \tag{5.77}$$

Vaughan (1984) noticed that in general $A_x = A_y = B_u = B_v = 0$, and also the odd moments, such as A_{xxx} , are zero. Hence (5.77) can be simplified into

$$\begin{aligned}
 E(\tilde{d}) = & d_{gh} \cdot \left[1 + \frac{1}{2d_{gh}^2} (B_{vv} + A_{yy}) + \frac{1}{8d_{gh}^4} (4B_{uu} A_{yy} + 4B_{uuvv} + 4A_{xxyy} \right. \\
 & \left. + 4A_{xx} B_{vv} - 6B_{vv} A_{yy} - B_{vvvv} - A_{yyyy}) \right]
 \end{aligned} \tag{5.78}$$

In general the moments of relative locations of the points within the region depend on the shape of the regions involved. A further approximation can be made by assuming a particular shape of the regions, such as a circle. Assume that the g th circle has a radius r_g and the h th circle has a radius r_h . Consider the

point p_a and p_b are located in the g th and h th region, respectively. The points will have a distance to its centroid as $r_g(a)$ and $r_h(b)$, respectively. Therefore the moment of relative location p_a and p_b within the g th and h th circles are

- First order moment :

$$\begin{aligned} A_y &= \int_A \frac{y}{A} dA \\ &= \frac{1}{\pi r_g^2} \int_0^{2\pi} \int_0^{r_g} r_g(a) \sin \theta r_g(a) dr_g(a) d\theta = 0 \end{aligned} \quad (5.79)$$

$$\begin{aligned} B_v &= \int_B \frac{v}{B} dB \\ &= \frac{1}{\pi r_h^2} \int_0^{2\pi} \int_0^{r_h} r_h(b) \sin \theta r_h(b) dr_h(b) d\theta = 0 \end{aligned} \quad (5.80)$$

- Second order moment :

$$\begin{aligned} A_{yy} &= \int_A \frac{y^2}{A} dA \\ &= \frac{1}{\pi r_g^2} \int_0^{2\pi} \int_0^{r_g} r_g(a)^2 \sin^2 \theta r_g(a) dr_g(a) d\theta = \frac{r_g^2}{4} = \frac{A}{4\pi} \end{aligned} \quad (5.81)$$

$$B_{vv} = \int_B \frac{v^2}{B} dB = \frac{B}{4\pi} \quad (5.82)$$

- Fourth order moment :

$$\begin{aligned} A_{yyyy} &= \int_A \frac{y^4}{A} dA \\ &= \frac{1}{\pi r_g^2} \int_0^{2\pi} \int_0^{r_g} r_g(a)^4 \sin^4(\theta) r_g(a) dr_g(a) d\theta \\ &= \frac{A^2}{8\pi^2} \end{aligned} \quad (5.83)$$

$$B_{vvvv} = \frac{B^2}{8\pi^2} \quad (5.84)$$

• Other :

$$\begin{aligned}
 A_{xxyy} &= \int_A \frac{x^2 \cdot y^2}{A} dA \\
 &= \frac{1}{\pi r_g^2} \int_0^{2\pi} \int_0^{r_g} r_g(a)^2 \cos^2 \theta r_g(a)^2 \sin^2(\theta) r_g(a) dr_g(a) d\theta \\
 &= \frac{A^2}{24\pi^2}
 \end{aligned} \tag{5.85}$$

$$\begin{aligned}
 B_{uuvv} &= \int_B \frac{u^2 \cdot v^2}{B} dB \\
 &= \frac{B^2}{24\pi^2}
 \end{aligned} \tag{5.86}$$

Substituting all these moments into (5.78) gives

$$E(\tilde{d}) = d_{gh} \cdot \left(1 + \frac{1}{8\pi d_{gh}^2} (A + B) + \frac{1}{192\pi^2 d_{gh}^4} (A + B)^2 + \frac{1}{192\pi^2 d_{gh}^4} AB \right) \tag{5.87}$$

Vaughan (1984) derived the approximation that involved only the first two terms within the bracket of equation (5.87). The last two extra terms of equation (5.87) are introduced that may enhance the approximation.

The second moment of the d , that is $E(d^2)$, can be determined by evaluating equation (5.71)

$$\begin{aligned}
 E(\tilde{d}^2) &= E \left(d_{gh}^2 \left(1 + \frac{2p}{d_{gh}} + \frac{q^2}{d_{gh}^2} \right) \right) \\
 &= E \left(d_{gh}^2 \left(1 + \frac{2}{d_{gh}} (u - x) + \frac{1}{d_{gh}^2} [(u - x)^2 + (v - y)^2] \right) \right) \\
 &= d_{gh}^2 \left(1 + \frac{2}{d_{gh}^2} (B_u - A_x) \frac{1}{d_{gh}^2} [(B_{uu} - 2B_u A_x + A_{xx}) + (B_{vv} - 2B_v A_y + A_{yy})] \right) \\
 &= d_{gh}^2 + \frac{1}{2\pi} (A + B)
 \end{aligned} \tag{5.88}$$

This result is similar to the result presented by Wilson (1990). The value $\frac{1}{2\pi}$ represents a constant value for the circle shape. Value for other shapes are also given in Wilson (1990).

Using (5.87) and (5.88), we can develop an approximation for the variance of the distance d , that is

$$\begin{aligned}
 V(\tilde{d}) &= E(\tilde{d}^2) - (E(\tilde{d}))^2 \\
 &= d_{gh}^2 + \frac{1}{2\pi} (A + B) - \left[d_{gh} \cdot \left(1 + \frac{1}{8\pi d_{gh}^2} (A + B) + \frac{1}{192\pi^2 d_{gh}^4} (A + B)^2 \right. \right. \\
 &\quad \left. \left. + \frac{1}{192\pi^2 d_{gh}^4} AB \right) \right]^2
 \end{aligned} \tag{5.89}$$

Illustration from simulation

This section gives some illustrations of how these formulas apply by performing simulations. The simulations were done by generating the individual level data uniformly distributed within a particular region. Two non-overlapping groups were created within the region boundary. Five groups shapes were considered, the square, rectangle, circle, equilateral triangle, and "L-shape", as illustrated in figure (5.13). The important contrast may be between the convex shape of points (a)-(d), and the non-convex shape (e).

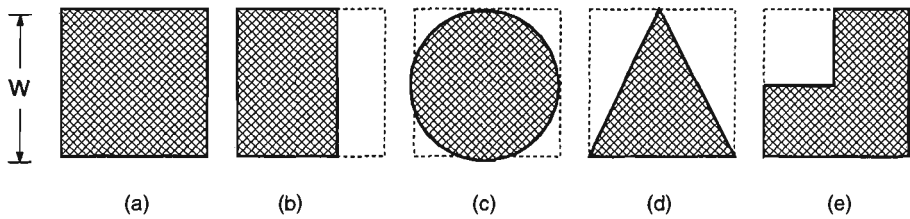
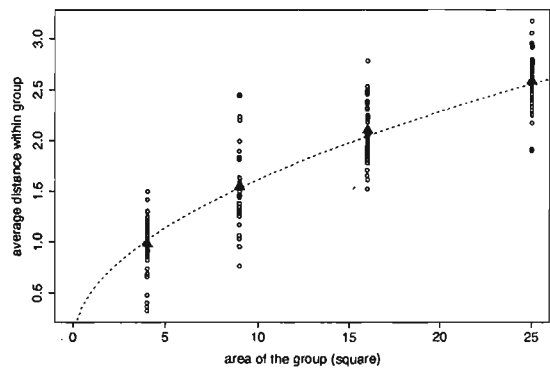


Figure 5.13. The shapes considered in the simulation, (a) square, (b) rectangle, (c) circle, (d) equilateral triangle, (e) "L-shape".

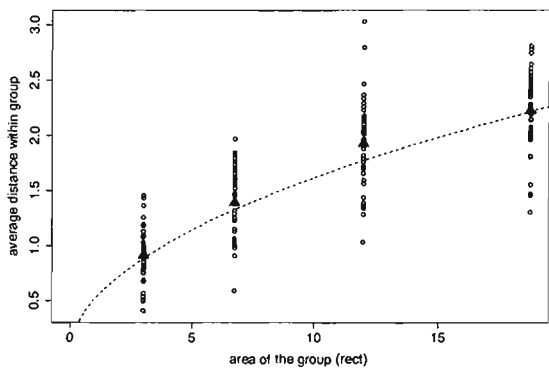
The simulations were repeated 1200 times and five different lengths (w) of the groups area were considered, these were 2, 3, 4, 5, or 6 unit. The population was generated in the region which was bounded at left lower corner (20,20) and right upper corner (90,80).

The mean and variance of distance between members within the same group, the mean and variance of distance between members of two different groups, and the centroid distance between groups were considered.

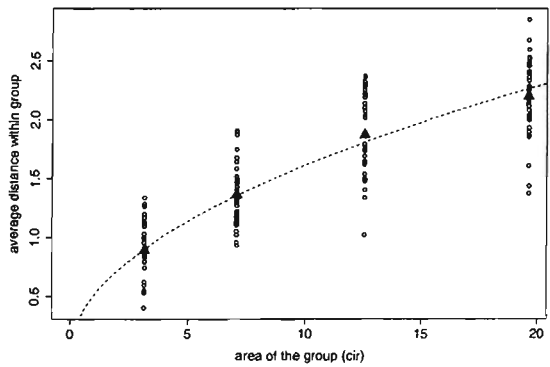
Figure (5.14) shows the relationship between the group's area with the average distance within the group for different shapes of the groups. It indicates a square root relationship between group's area and the average of distance within the group. Figure (5.15) shows the relationship between the group's area with the variance of the distances within the group for different shapes of the groups. It indicates a linear relationship between them. These figures agree with the relationships as formulated in (5.65). Table (5.4) gives some descriptions of comparison between the value of the average and variance of distance within the group obtained from the simulation with the approximations given in (5.68). The dash lines in figures (5.14) and (5.15) indicate the approximation (5.68).



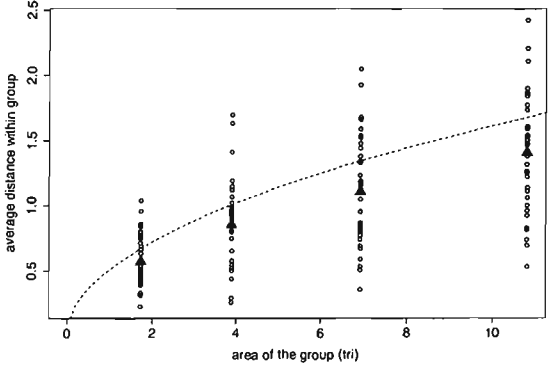
(a) square



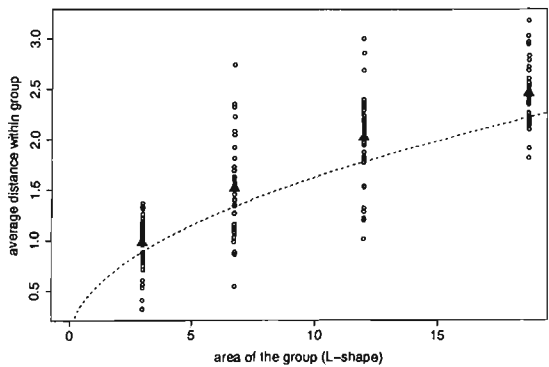
(b) rectangle



(c) circle

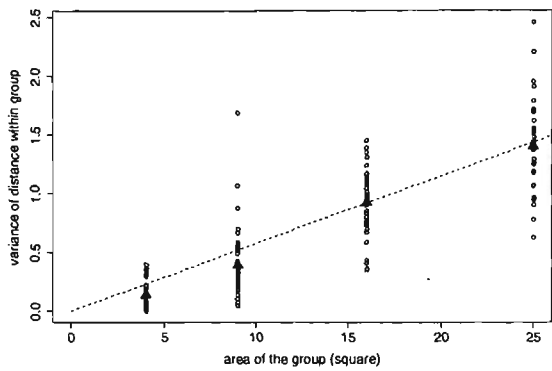


(d) triangle

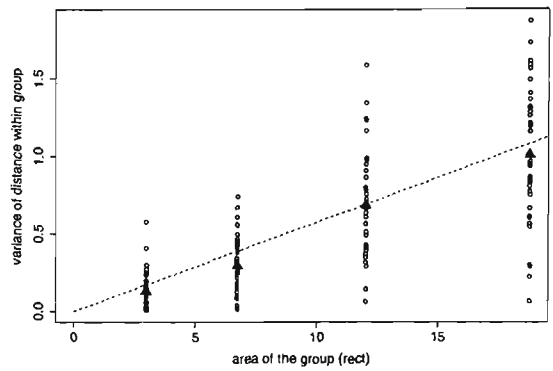


(e) L-shape

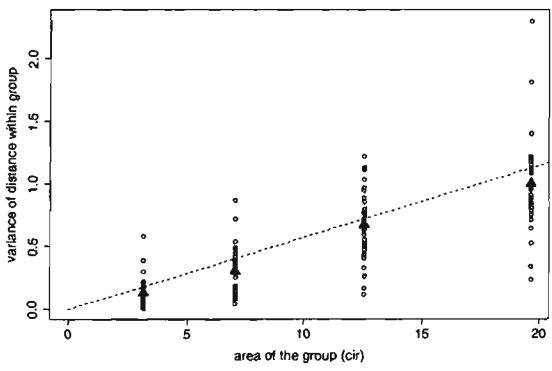
Figure 5.14. Relationship between average distance within group with the area of the group at different shapes of the groups, note : the \blacktriangle indicates the mean point.



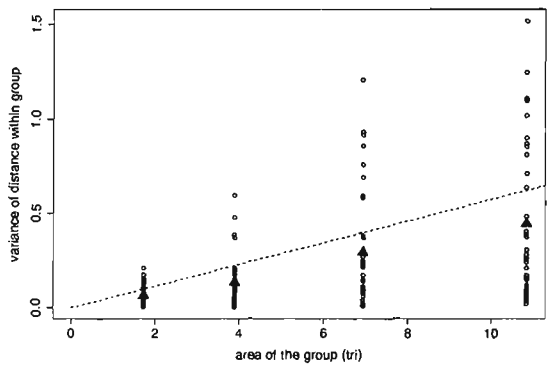
(a) square



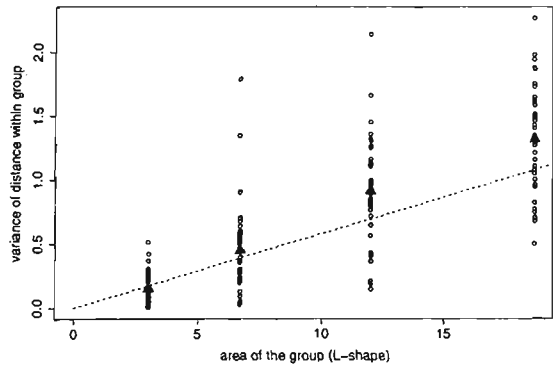
(b) rectangle



(c) circle



(d) triangle



(e) L-shape

Figure 5.15. Relationship between variance of distance within group with area of the group at different shapes of the groups, note : the \blacktriangle indicates the mean point.

Table 5.4. The mean and variance of the distance within groups. Comparison between the simulated value and its approximation

| Groups shape | Within group distances | | | |
|------------------------|-------------------------|--------------------|-----------------------------|------------------------|
| | Average of sim. mean | Approx. to mean | Average of sim. variance | Approx. to variance |
| square | 1.833 | 1.788 | 0.714 | 0.744 |
| rectangle (3/4) | 1.602 | 1.548 | 0.558 | 0.581 |
| circle | 1.583 | 1.584 | 0.509 | 0.608 |
| Triangle (equal sides) | 1.008 | 1.177 | 0.227 | 0.335 |
| L shape | 1.738 | 1.548 | 0.708 | 0.581 |

Table (5.5) and figure (5.16) show a comparison of the mean and variance of distance between points in two groups with the approximated values as defined in (5.87) and (5.89), respectively. Three situations concerning the distance between the centroids of the groups, d_{gh} were considered, corresponding to $d_{gh} \approx 5, 12$ and 19 . The approximation for the average distance between groups is very close to the true values as indicated by the correlation $r = 0.968$. Meanwhile the approximation for the variance of distance between groups is fairly close as $r = 0.740$. Table (5.5) shows illustration of the approximation of the mean and variance at difference values of d_{gh} . Part (a) of table (5.5) indicates the distance between centroids around 5, meanwhile part (b) and part (c) are parted by the distance around 12 unit and 19 unit, respectively. The table shows that approximation (5.87) and (5.89) work very well for the mean and reasonably well for the variance. Moreover, d_{gh} is also close to the simulated means although slightly less in all cases.

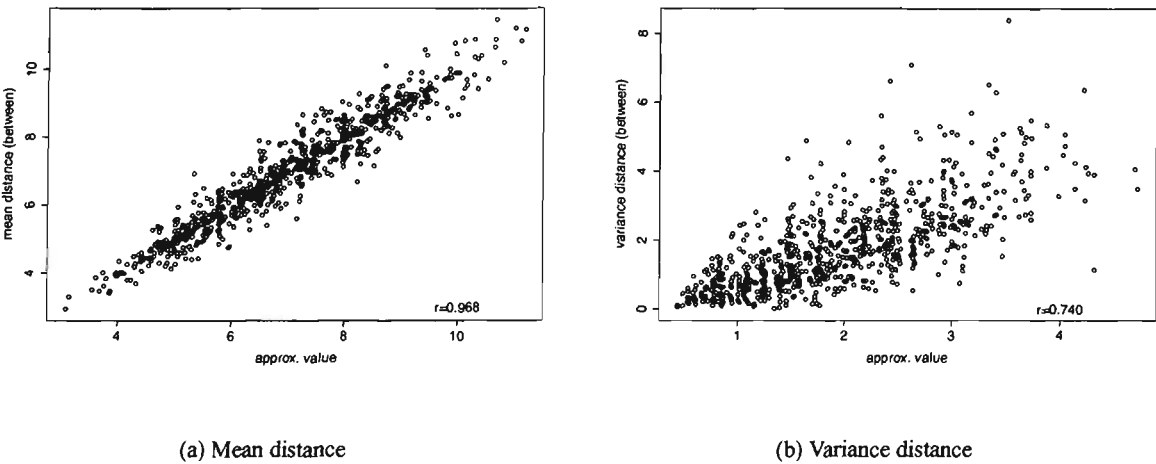


Figure 5.16. The plot between the simulated and approximated of the mean and variance distance between groups

Table 5.5. The description of the mean and variance of the distance between groups. Comparison between the simulated values and its approximation (5.87) and (5.89) respectively

| Groups shape | Between group | | | | |
|------------------------|---------------|--|-----------------------------------|--|---------------------------------------|
| | d_{gh} | Average of sim. mean (std. err.) | Approx. to mean (std. err.) | Average of sim. variance (std. err.) | Approx. to variance (std. err.) |
| (a) | | | | | |
| square | 5.226 | 5.477 (1.522) | 5.442 (1.432) | 2.031 (1.253) | 2.190 (0.928) |
| rectangle (3/4) | 5.465 | 5.701 (1.529) | 5.646 (1.452) | 1.800 (1.256) | 1.946 (0.829) |
| circle | 5.260 | 5.451 (1.503) | 5.452 (1.426) | 1.813 (1.204) | 1.978 (0.843) |
| Triangle (equal sides) | 6.228 | 6.325 (1.703) | 6.362 (1.712) | 1.282 (1.019) | 1.632 (0.724) |
| L shape | 5.214 | 5.392 (1.482) | 5.407 (1.450) | 2.259 (1.582) | 1.938 (0.825) |
| (b) | | | | | |
| square | 12.189 | 12.324 (1.613) | 12.280 (1.510) | 2.093 (1.367) | 2.261 (0.951) |
| rectangle (3/4) | 12.322 | 12.399 (1.656) | 12.402 (1.557) | 1.882 (1.164) | 1.997 (0.847) |
| circle | 12.086 | 12.162 (1.534) | 12.170 (1.486) | 1.744 (1.049) | 2.034 (0.861) |
| Triangle (equal sides) | 13.316 | 13.346 (1.842) | 12.232 (1.840) | 1.338 (1.085) | 1.661 (0.743) |
| L shape | 12.151 | 12.229 (1.716) | 12.232 (1.660) | 2.101 (1.380) | 1.996 (0.847) |
| (c) | | | | | |
| square | 19.184 | 19.272 (1.644) | 19.242 (1.552) | 2.049 (1.277) | 2.271 (0.957) |
| rectangle (3/4) | 19.312 | 19.405 (1.734) | 19.363 (1.639) | 1.814 (1.121) | 2.004 (0.852) |
| circle | 19.210 | 19.314 (1.623) | 19.262 (1.569) | 1.761 (1.277) | 2.042 (0.866) |
| Triangle (equal sides) | 20.001 | 20.059 (1.881) | 20.042 (1.809) | 1.376 (1.038) | 1.665 (0.745) |
| L shape | 19.321 | 19.386 (1.626) | 19.372 (1.600) | 2.005 (1.381) | 2.004 (0.852) |

5.3.3 Evaluating $\tilde{\Gamma}_{gh}$ for exponential model

Consider the situation where the data are available at the group level along with the centroids of each group. Using the approximation (5.68) for \bar{d}_g and $S_{d_g}^2$ we may rewrite equation (5.51)

$$\begin{aligned} \tilde{\Gamma}_{gh} = & \gamma(d_{gh}) \\ & + \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) + \frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2 + \gamma''(d_{gh})(\bar{d}_{gh} - d_{gh})^2 \\ & - \frac{N_g - 1}{2N_g} \left(\gamma(k_1 \sqrt{\mathcal{A}_g}) + \gamma''(k_1 \sqrt{\mathcal{A}_g}) \frac{k_2}{2} \mathcal{A}_g \right) \\ & - \frac{N_h - 1}{2N_h} \left(\gamma(k_1 \sqrt{\mathcal{A}_h}) + \gamma''(k_1 \sqrt{\mathcal{A}_h}) \frac{k_2}{2} \mathcal{A}_h \right) \end{aligned} \tag{5.90}$$

Table (5.5) gave illustrations of the approximation of the mean distance of points between groups (\bar{d}_{gh}) compared with the distance between centroid of the groups (d_{gh}). They are very similar, hence their difference is approximately zero. This implies that the term in (5.90) which involved the $(\bar{d}_{gh} - d_{gh})$ is very small, then it can be ignored, at least for uniformly distributed individuals.

Figure (5.7) shows the plot of $\gamma'()$ and $\gamma''()$ for two different semivariogram models, the exponential and Gaussian. The $\gamma''()$ is small relative to $\gamma(d_{gh})$ and gets smaller as the average distance between groups

gets larger. Meanwhile, table (5.5) and result (5.89) suggest that the variance of distance between groups are small and similar regardless of the distance between groups unless the areas of the groups are large.

Therefore the term $\frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2$ in (5.90) should have only a small effect on the value of $\tilde{\Gamma}_{gh}$.

This suggests that the bias mainly arises from the within groups factors, which involve \bar{d}_g and $S_{d_g}^2$. This leads to focussing on the zero order term, which will be discussed in section (5.4.1). For example, let us consider an exponential semivariogram model (5.9) to illustrate the likely bias. Here we parameterized again the exponential model by using $3d/r$ instead of d/r . The corresponding of equation (5.51) can be expressed in term of this model, that is;

$$\begin{aligned} \tilde{\Gamma}_{gh} = & s - (s - n) \cdot \exp\left[\frac{-3d_{gh}}{r}\right] \\ & + \frac{3}{r}(s - n) \cdot \exp\left[\frac{-3d_{gh}}{r}\right] \left\{ (\bar{d}_{gh} - d_{gh}) - \frac{3}{2r} S_{d_{gh}}^2 - \frac{3}{r} (\bar{d}_{gh} - d_{gh})^2 \right\} \\ & - s \left\{ 1 - \frac{1}{2N_g} - \frac{1}{2N_h} \right\} \\ & + \frac{N_g - 1}{2N_g} (s - n) \cdot \exp\left[\frac{-3\bar{d}_g}{r}\right] \left\{ 1 + \frac{9}{2r^2} \cdot S_{d_g}^2 \right\} \\ & + \frac{N_h - 1}{2N_h} (s - n) \cdot \exp\left[\frac{-3\bar{d}_h}{r}\right] \left\{ 1 + \frac{9}{2r^2} \cdot S_{d_h}^2 \right\} \end{aligned} \quad (5.91)$$

Assume that the d_{gh}/r is large enough that $\exp\left[\frac{-3d_{gh}}{r}\right]$ will be almost zero, and hence equation (5.91) will become,

$$\begin{aligned} \tilde{\Gamma}_{gh} \approx & \frac{s}{2} \left(\frac{1}{N_g} + \frac{1}{N_h} \right) + \frac{N_g - 1}{2N_g} (s - n) \cdot \exp\left[\frac{-3\bar{d}_g}{r}\right] \left\{ 1 + \frac{9}{2r^2} \cdot \bar{d}_g \right\} \\ & + \frac{N_h - 1}{2N_h} (s - n) \cdot \exp\left[\frac{-3\bar{d}_h}{r}\right] \left\{ 1 + \frac{9}{2r^2} \cdot \bar{d}_h \right\} \end{aligned} \quad (5.92)$$

Equation (5.92) shows that the individual level sill (s) will be affected by the within each group factors, such as the N_g , N_h , \bar{d}_g , \bar{d}_h , $S_{d_g}^2$, and $S_{d_h}^2$. Consider the case that \bar{d}_g and \bar{d}_h are large so that $\exp(-3\bar{d}_g/r)$ and $\exp(-3\bar{d}_h/r)$ are very small, then

$$\tilde{\Gamma}_{gh} \approx s \cdot \frac{1}{2} \left(\frac{1}{N_g} + \frac{1}{N_h} \right)$$

This equation shows that the sill will be reduced by the factor $\frac{1}{2} \left(\frac{1}{N_g} + \frac{1}{N_h} \right)$.

Another case is when the groups are much smaller than r so that $\frac{3\bar{d}_g}{r}$ is close to zero and so $\exp(-3\bar{d}_g/r) \approx 1$, then

$$\tilde{\Gamma}_{gh} \approx s - n \left(1 - \frac{1}{2} \left[\frac{1}{N_g} + \frac{1}{N_h} \right] \right)$$

Here the sill is again reduced, effectively by the value of the nugget parameter. Suppose that the two groups are much smaller than r and very close together so that $3d_{gh}/r$ is close to zero and hence $\exp(-3d_{gh}/r) \approx 1$.

Under these conditional the second term in (5.91) will also be close to zero and

$$\hat{\Gamma}_{gh} \approx \frac{n}{2} \left(\frac{1}{N_g} + \frac{1}{N_h} \right)$$

This suggests that analysis of the group level semivariogram values will give an estimated nugget of approximately $n \cdot (\frac{1}{N})$.

Equation (5.91) suggests that given the group level semivariogram value $\hat{\Gamma}_{gh}$, d_{gh} , N_g , N_h , and approximate values for \bar{d}_g , \bar{d}_h , $S_{d_g}^2$, $S_{d_h}^2$, \bar{d}_{gh} , $S_{d_{gh}}^2$, then it may be possible to estimate the parameters of the individual level semivariogram. A non-linear estimation method can be developed to estimate those parameters. This issue will be discussed in the following section.

5.4 Estimation of individual level semivariogram parameters from the group level semivariogram

This section will propose and evaluate a method to estimate parameters of the individual semivariogram from the observed group level semivariogram values. This is done by applying theorem (5.2.2) concerning the relationship between group level semivariogram and individual level semivariogram and the approximation developed in section (5.3).

5.4.1 Development of the method

Using (5.51) we can develop methods to estimate the individual level semivariogram parameters (n , s , and r) from group level data. Two methods will be developed depending on whether some non-spatial individual level data are available. The method will use equation (5.51) to develop a non-linear model relating the empirical group level semivariogram values to the unknown parameters.

We can consider three different situation of equation (5.51). The first situations is a zero order, that is, we ignore all but zero order terms of $\gamma()$, giving

$$\Gamma_{gh}^0 \approx \gamma(d_{gh}) - \frac{N_g - 1}{2N_g} \gamma(\bar{d}_g) - \frac{N_h - 1}{2N_h} \gamma(\bar{d}_h) \quad (5.93)$$

The second situation is where we include all first order terms of equation (5.51), that is,

$$\Gamma_{gh}^1 \approx \gamma(d_{gh}) + \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) - \frac{N_g - 1}{2N_g} \gamma(\bar{d}_g) - \frac{N_h - 1}{2N_h} \gamma(\bar{d}_h) \quad (5.94)$$

The third situation is where all second order terms are included , that is equation (5.51) itself.

The $\hat{\Gamma}_{gh} = \frac{1}{2}(\bar{Y}_g - \bar{Y}_h)^2$ shows the empirical value of the group level semivariogram of group g and h (5.33). The adequacy of the approximated (5.93), can be investigated by examining

$$\hat{\Gamma}_{gh} - \Gamma_{gh}^0 \quad (5.95)$$

This term represents the extra terms in (5.51). The simulated data as discussed in page (99) were used to show the property of (5.95). The simulated individual level data followed the exponential semivariogram model with $n = 5$, $s = 20$, and $r = 10$. A pair of groups was created and empirical group level semivariogram value and Γ_{gh}^0 were calculated. The simulations were repeated 1200 times. Figure (5.17) shows that the points of the difference $(\hat{\Gamma}_{gh} - \Gamma_{gh}^0)$ are distributed around the zero for all d_{gh} (distance between groups centroid). The boxplot shows that the difference has mean 0.6767 and variance 32.6841. This illustrates that the zero order (5.93) is expected to give a good estimation of the group level semivariogram values, with a reasonable small error.

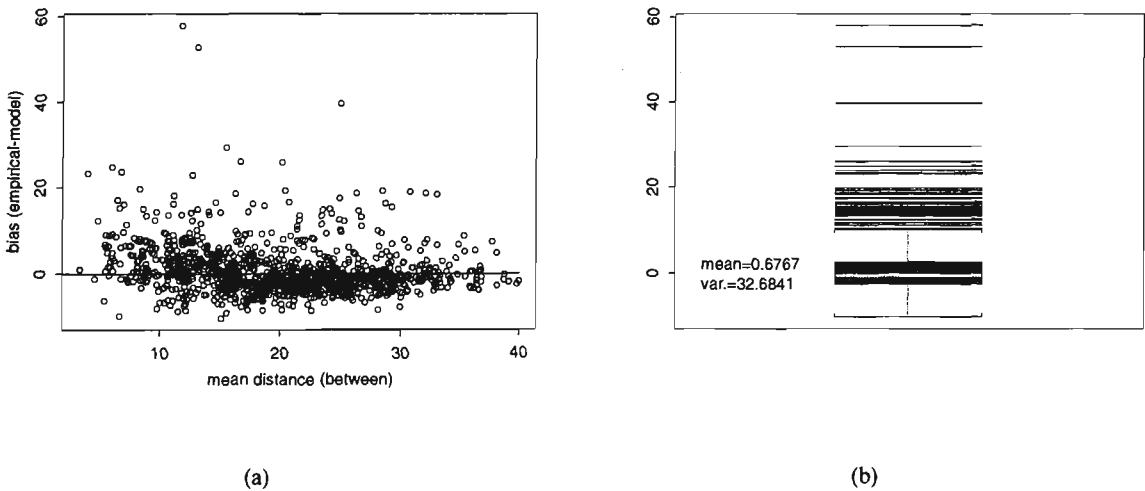


Figure 5.17. (a) Relationship between $(\hat{\Gamma}_{gh} - \Gamma_{gh}^0)$ and \bar{d}_{gh} , (b) boxplot of $(\hat{\Gamma}_{gh} - \Gamma_{gh}^0)$.

Consider again the exponential model, then corresponding to equation (5.93) is,

$$\begin{aligned}\hat{\Gamma}_{gh} = & s - (s - n) \cdot \exp\left[\frac{-3d_{gh}}{r}\right] - s \left\{1 - \frac{1}{2N_g} - \frac{1}{2N_h}\right\} \\ & + \frac{N_g - 1}{2N_g}(s - n) \cdot \exp\left[\frac{-3\bar{d}_g}{r}\right] + \frac{N_h - 1}{2N_h}(s - n) \cdot \exp\left[\frac{-3\bar{d}_h}{r}\right]\end{aligned}\quad (5.96)$$

The ignored term from (5.91) is

$$\begin{aligned}R_0 = & \frac{3}{r}(s - n) \cdot \exp\left[\frac{-3d_{gh}}{r}\right] \left\{(\bar{d}_{gh} - d_{gh}) - \frac{3}{2r}S_{d_{gh}}^2 - \frac{3}{r}(\bar{d}_{gh} - d_{gh})^2\right\} \\ & + \frac{N_g - 1}{2N_g}(s - n) \cdot \exp\left[\frac{-3\bar{d}_g}{r}\right] \cdot \left(\frac{9}{2r^2}\right) \cdot S_{d_g}^2 \\ & + \frac{N_h - 1}{2N_h}(s - n) \cdot \exp\left[\frac{-3\bar{d}_h}{r}\right] \cdot \left(\frac{9}{2r^2}\right) \cdot S_{d_h}^2\end{aligned}\quad (5.97)$$

In section (5.3) we showed $\bar{d}_g \approx k_1\sqrt{\mathcal{A}_g}$ and $S_{d_g}^2 \approx k_2\mathcal{A}_g$ where k_1, k_2 are both less than 1 for most shapes. Hence provided $\sqrt{\mathcal{A}_g}/r$ and $\sqrt{\mathcal{A}_h}/r$ are small the last two terms in R_0 should be small. Similarly (5.87) and (5.89) can be used to show the first term in R_0 will be small provided $\sqrt{\mathcal{A}_g}/r$ and $\sqrt{\mathcal{A}_h}/r$ are small. The results in figure (5.17) confirm this. Hence we will focus on the use of (5.93) in developing an estimation procedure.

We can develop an estimation method for the individual level semivariogram parameters based on equation (5.96) and $E(\hat{\Gamma}_{gh}) = \Gamma_{gh} \approx \hat{\Gamma}_{gh}$. All components in (5.96) are known except for the n , s , and r . We can estimate these parameters (n , s , and r) using a non-linear least squares method with $\hat{\Gamma}_{gh}$ as the dependent variable. The SAS code was developed to implement this estimation (appendix E). This procedure is based on an iterative process with a given initial value. Obtaining good results will depend on the appropriate initial value.

An important issue in the non-linear estimation procedure is determining initial values for the nugget, sill, and range. Poor initial values can lead to convergence problems. There are two situation that will be considered to determine the initial values. The first situation is when non-spatial individual level sample data are available. The second situation is when an individual level sample is not available.

The individual level sample data are available

Theorem 5.4.1. *Suppose that a simple random sample K of size n is taken from the population \mathcal{U} . The sample variance, s^2 , is an unbiased estimator of the population individual level semivariogram mean,*

Proof. Define

$$s^2 = \frac{1}{n(n-1)} \sum_{i,j \in K} \frac{1}{2} (Y_i - Y_j)^2$$

Hence

$$\begin{aligned} E_p(s^2) &= \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{U}} \text{Prob}((i,j) \in K) \frac{1}{2} (Y_i - Y_j)^2 \\ &= \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{U}} \frac{n(n-1)}{N(N-1)} \frac{1}{2} (Y_i - Y_j)^2 \\ &= \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{U}} \hat{\gamma}_{ij} = \bar{\gamma} \end{aligned}$$

where $E_p()$ represents expectation with respect to the simple random sampling process. Then we have

$$E(s^2) = E_\xi [E_p(s^2)] = E_\xi(\bar{\gamma}) = \bar{\gamma}$$

where we have used the " ξ " subscript to indicate taking expectation over the superpopulation process that generated the population. \square

We will take the initial value for the sill as the variance of the individual level sample data (s^2). The sample may be taken by simple random sample from the population. This will be an underestimate of the sill because $\gamma_{ij} \leq s$, but will provide a reasonable starting value. For illustration, consider the exponential semivariogram model

$$\bar{\gamma} = s - (s - n) \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \exp \left[\frac{-3d_{ij}}{r} \right]$$

Let

$$\bar{d}_R = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} d_{ij}$$

be the average distance between points within the region R , then using a second order Taylor series approximation for $\exp \left[\frac{-3d_{ij}}{r} \right]$ gives

$$\begin{aligned} \bar{\gamma} &= s - (s - n) \left(\exp \left[\frac{-3\bar{d}_R}{r} \right] + \frac{9}{2r^2} \cdot \exp \left[\frac{-3\bar{d}_R}{r} \right] \cdot S_{\bar{d}_R}^2 \right) \\ &= s - (s - n) \cdot \exp \left[\frac{-3\bar{d}_R}{r} \right] \left(1 + \frac{9}{2r^2} \cdot S_{\bar{d}_R}^2 \right) \end{aligned}$$

where

$$S_{d_R}^2 = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} (d_{ij} - \bar{d}_R)^2$$

Using the result for the mean and variance of distances between points within a region (5.68), we can write this as

$$\bar{\gamma} \approx s - (s - n) \cdot \exp \left[\frac{-3k_1 \sqrt{\mathcal{A}_R}}{r} \right] \left(1 + \frac{9}{2r^2} \cdot k_2 \mathcal{A}_R \right) \quad (5.98)$$

For most shapes $k_1 \leq 1.0$ and $k_2 \leq 0.5$ (see table 5.3.2). Provided $3 \frac{\sqrt{\mathcal{A}_R}}{r}$ is large, i.e. the region is much larger than the range, then $\bar{\gamma}$ will be close to s . Another case, is $3 \frac{\sqrt{\mathcal{A}_R}}{r} = c$ then the absolute value of the second term in (5.98) is less than

$$(s - n) \exp(-k_1 \cdot c) \cdot \left(1 + \frac{9}{2} k_2 c^2 \right)$$

For the case of a circle region with radius R , then $\mathcal{A}_R = \pi \cdot R^2$. Hence $3 \frac{\sqrt{\mathcal{A}_R}}{r} = 3\sqrt{\pi} \frac{R}{r}$. Moreover $k_1 = 0.511$ and $k_2 = 0.0574$ (see table 5.3.2). Thus if $\frac{R}{r} = 2$, we get the absolute value of the second term in (5.98) is $0.0142 (s - n)$.

The initial value for the nugget may be chosen between zero to the sill value, which gives two extreme cases. In the first case when $n = 0$, we start by assuming that there is no variation among observations at distance zero. And in the second case when $n = s$, we assume that initially there is no spatial autocorrelation in the data.

The initial value of the range can be developed on the basis section (4.4) and (5.1). The basic relationship between individual and group level variance was shown in equation (4.24). This equation can be modified in terms of semivariogram, and it has expectation

$$E(N\bar{S}_{yy}) = \frac{N-1}{M-1} \bar{\gamma} - \frac{N-M}{M-1} \cdot E(S_{yy}^{<W>}) \quad (5.99)$$

Meanwhile by taking expectation of (4.67), we can define

$$E(S_{yy}^{<W>}) = \frac{M-1}{M(\bar{N}-1)} \left(\bar{S}_{N\bar{\gamma}} + \frac{M(\bar{N}-1)}{M-1} \tilde{\gamma}_W \right) \quad (5.100)$$

Therefore

$$E(N\bar{S}_{yy}) = \frac{N-1}{M-1}\bar{\gamma} - \bar{S}_{N\bar{\gamma}} - \frac{M(\bar{N}-1)}{M-1}\tilde{\gamma}_W \quad (5.101)$$

Consider a case N_g is a constant at \bar{N} , then $\bar{S}_{N\bar{\gamma}} = 0$. The estimate $\tilde{\gamma}_W$ can be developed by evaluating equation (5.101) and applying (5.99), that is

$$\tilde{\gamma}_W = \frac{\bar{N}}{\bar{N}-1} \left(\frac{N-1}{N} \cdot S_{yy} - \frac{M-1}{N} \cdot N\bar{S}_{yy} \right) \quad (5.102)$$

Therefore the estimator of the average within group autocorrelation can be defined

$$\hat{\rho}_W = 1 - \frac{\tilde{\gamma}_W}{\sigma^2} \quad (5.103)$$

The initial value of the range is defined by the following derivation of $\tilde{\gamma}_W$. For the exponential semi-variogram model with parameter (n, s , and r), then we may rewrite $\tilde{\gamma}_W$ as;

$$\tilde{\gamma}_W = s - (s - n) \frac{1}{M} \sum_g \exp \left[\frac{-3 \cdot k_1 \sqrt{\mathcal{A}_g}}{r} \right] \quad (5.104)$$

Assume that \mathcal{A}_g is constant at $\bar{\mathcal{A}}$ and divide by σ^2 , then

$$\tilde{\rho}_W = \left(1 - \frac{n}{C(0)} \right) \exp \left[\frac{-3 \cdot k_1 \sqrt{\bar{\mathcal{A}}}}{r} \right] \quad (5.105)$$

where the σ^2 is equal to the S_{yy} and estimated by s^2 . Consider the nugget is zero, then we may rearrange equation (5.105) to get,

$$r = \frac{-3 \cdot k_1 \sqrt{\bar{\mathcal{A}}}}{\log \tilde{\rho}_W} \quad (5.106)$$

The estimate $\tilde{\gamma}_W$ can be obtained from equation (5.102), then substituting into equation (5.103) to get $\hat{\rho}_W$.

It can be noted that equation (5.106) can be obtained directly from (5.102) and (5.104).

The individual level sample data are not available

There are two approaches that can be applied in this situation. The first approach is developed by choosing the estimators of the group level semivariogram model parameters as the initial value. The second approach is developed as follows. Suppose that the nugget is zero and \mathcal{A} is the area of the region, an approximation for $\bar{\gamma}$ is given by

$$\bar{\gamma} \approx s \cdot \left(1 - \exp \left[\frac{-3 \cdot k_1 \sqrt{\mathcal{A}}}{r} \right] \right) \quad (5.107)$$

Suppose $\mathcal{A}_g = \mathcal{A}_h = \bar{\mathcal{A}}$, then the same sort of approximation can be used for $\tilde{\gamma}_W$, that is

$$\tilde{\gamma}_W = s \left(1 - \exp \left[\frac{-3k_1\sqrt{\bar{\mathcal{A}}}}{r} \right] \right) \quad (5.108)$$

Using Taylor series approximation for the $\exp(-x) = 1 - x$, we have

$$\exp \left[\frac{-3 \cdot k_1\sqrt{\bar{\mathcal{A}}}}{r} \right] = 1 - \frac{3 \cdot k_1\sqrt{\bar{\mathcal{A}}}}{r} \quad (5.109)$$

Consider (5.101) in the case of N_g constant at \bar{N} , and then using (5.107), (5.108), and (5.109) gives

$$\frac{s}{r} = \frac{\left(\frac{M-1}{\bar{N}-1} \right) N\bar{S}_{yy}}{3k_1 \left(\sqrt{\bar{\mathcal{A}}} - \frac{N-M}{\bar{N}-1} \sqrt{\bar{\mathcal{A}}} \right)} \quad (5.110)$$

Applying the empirical relationship between group level semivariogram and individual level semivariogram such as defined in (4.64). Consider the exponential model at large d_{gh} and assume zero nugget, then

$$\tilde{\gamma}_{gh} = s; \quad \text{and} \quad \tilde{\gamma}_g = s \left(1 - \exp \left[\frac{-3\bar{d}_g}{r} \right] \right)$$

By Taylor series approximation,

$$\tilde{\gamma}_g = \frac{s}{r} 3 \cdot k_1 \sqrt{\mathcal{A}_g}$$

Hence the empirical group level semivariogram in (4.64) becomes

$$\hat{\Gamma}_{gh} = s - \frac{s}{r} \cdot \frac{\bar{N} - 1}{2\bar{N}} \cdot 3k_1 \left(\sqrt{\mathcal{A}_g} + \sqrt{\mathcal{A}_h} \right) \quad (5.111)$$

Substituting (5.110) into (5.111) and solving for s

$$s = \hat{\Gamma}_{gh} + \frac{M-1}{N-1} N\bar{S}_{yy} \frac{(\sqrt{\mathcal{A}_g} + \sqrt{\mathcal{A}_h})}{\sqrt{\bar{\mathcal{A}}} - \frac{N-M}{\bar{N}-1} \sqrt{\bar{\mathcal{A}}}} \cdot \frac{\bar{N} - 1}{2\bar{N}} \quad (5.112)$$

In (5.112) the s actually varies depend on the $\hat{\Gamma}_{gh}$, \mathcal{A}_g , and \mathcal{A}_h , therefore the sill is written as s_{gh} . The average of s_{gh} is

$$\bar{s} = \frac{1}{M(M-1)} \sum_{g \neq h} s_{gh}$$

The \bar{s} can be used as the initial value of the sill. The initial value of the nugget (n) is defined to be zero or equal to the initial value of the sill, and the initial value of the range is obtained from equation (5.110) using \bar{s} as the s and solve for the r .

5.4.2 Illustration from simulated data

This section illustrates using group level data to estimate the individual level semivariogram using simulated data. The individual level population values are generated according to the exponential semivariogram model with the parameter values of the nugget, sill, and range being 0, 20, and 10, respectively. Locations were generated randomly within the region. The population size is 1500 individuals in the square region with boundary defined by the coordinates (20,10) and (90,80). One hundred and fifty groups were created by dividing the region into a 10 by 15 grid of equal size rectangles. Note although the groups are equal in area, they will vary in terms of the number of individuals that they contain because of the random generation of locations. One of the simulation results is shown in figure (5.18). The square dots (\square) indicate the individual categorized semivariogram, and the solid square dots indicate the categorized group level semivariogram. Figure (5.18-a) is consistent with the figure (5.8), and shows the negative effect of the within group component of the bias discussed in section (5.3).

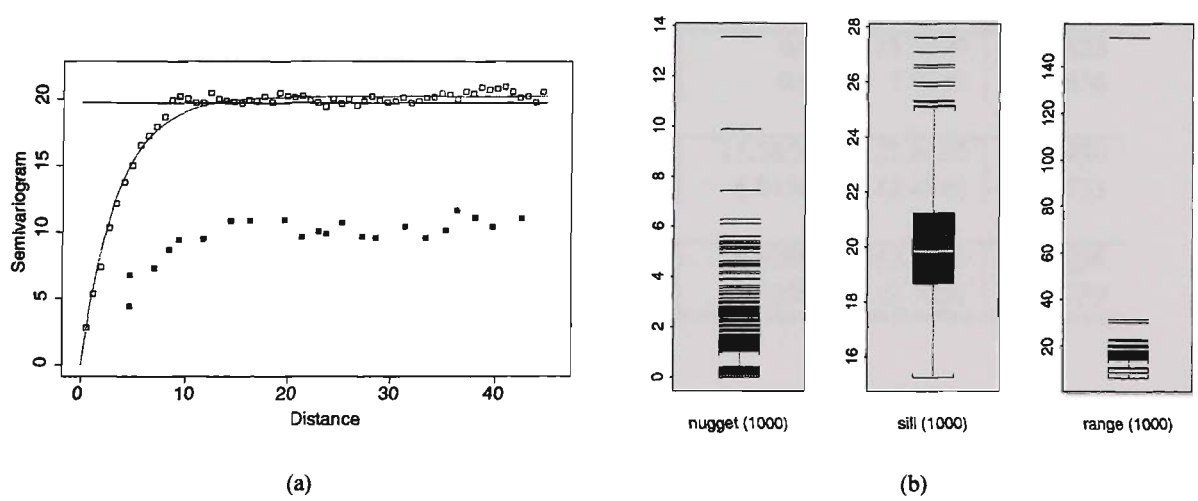


Figure 5.18. (a) The categorized version of the individual and group level semivariogram; (b) distribution of estimated parameters of individual level semivariogram model from 1000 simulations

The simulations were repeated a thousands times. Formal estimation of the semivariogram model parameters was done by the weighted least squares methods discussed in section (5.1.3) (see Cressie, 1991). Figure (5.18-b) and Table (5.6) show the distribution of the estimates of the semivariogram model parameters from 1000 simulations. Table (5.6) indicates that distribution of the estimators using the individual level data are consistent with the true values ($n = 0$, $s = 20$, and $r = 10$). For the sill and the range, the

mean and median are very close to the parameters value. For the nugget, the median is the same as the parameter value $n = 0$.

Meanwhile the group level semivariogram is described by figure (5.19). In 50 simulations the estimation process failed to converge, with unreliable estimates when the estimation process stopped. Hence, in a small proportion of cases, the use of group level data may not produce estimates.

After deleting these cases, there are 950 simulations remaining and described in figure (5.19-b), and associated statistics are found in table (5.6). The median of the estimate of the group level nugget, \hat{n} , is the same as the estimated of the individual level nugget, and the mean is slightly less. The mean and median of the group level sill, \hat{s} , is smaller than the estimated individual level. But the estimated group level range, \hat{r} , is larger, for both the mean and median than the estimated individual level range.

Table 5.6. Descriptive values of Figure (5.18-b)

| Parameter | Nbr. of sim. | Mean | Median | Minimum | Maximum | Standard error |
|---------------------|--------------|---------|---------|---------|----------|----------------|
| Nugget | | | | | | |
| individual (n) | 1000 | 0.4505 | 0.0 | 0.0 | 13.5520 | 1.0825 |
| group (\hat{n}) | 950 | 0.3991 | 0.0 | 0.0 | 7.0240 | 1.0656 |
| Sill | | | | | | |
| individual (s) | 1000 | 20.0032 | 19.8370 | 15.2650 | 27.6150 | 1.9490 |
| group (\hat{s}) | 950 | 10.0595 | 10.0595 | 6.0190 | 16.4860 | 1.6733 |
| Range | | | | | | |
| individual (r) | 1000 | 10.5491 | 9.9200 | 6.6390 | 152.5360 | 5.0996 |
| group (\hat{r}) | 950 | 16.7454 | 15.5970 | 3.3800 | 52.7030 | 5.8739 |

Table 5.7. Descriptive values of difference between the estimated parameters of individual level and group level semivariogram

| Parameter | Mean | Median | Minimum | Maximum | Standard error |
|-------------------------------|---------|---------|----------|---------|----------------|
| Nugget bias ($n - \hat{n}$) | -0.0319 | 0.0 | -7.0240 | 6.1180 | 1.2482 |
| Sill bias ($s - \hat{s}$) | 9.8579 | 9.8465 | 6.9520 | 12.7020 | 0.7630 |
| Range bias ($r - \hat{r}$) | -6.4596 | -5.5145 | -42.7330 | 3.7500 | 4.9889 |

Figure (5.18-a) shows that the group level semivariogram lies below the individual level as predicted by the theory in section (5.3.1). Table (5.7) shows the difference between the estimated parameters of the group level and the individual level semivariogram. Compared with the individual level estimates, the group level sill decreases, the group level range increases, but group level nugget does not change much.

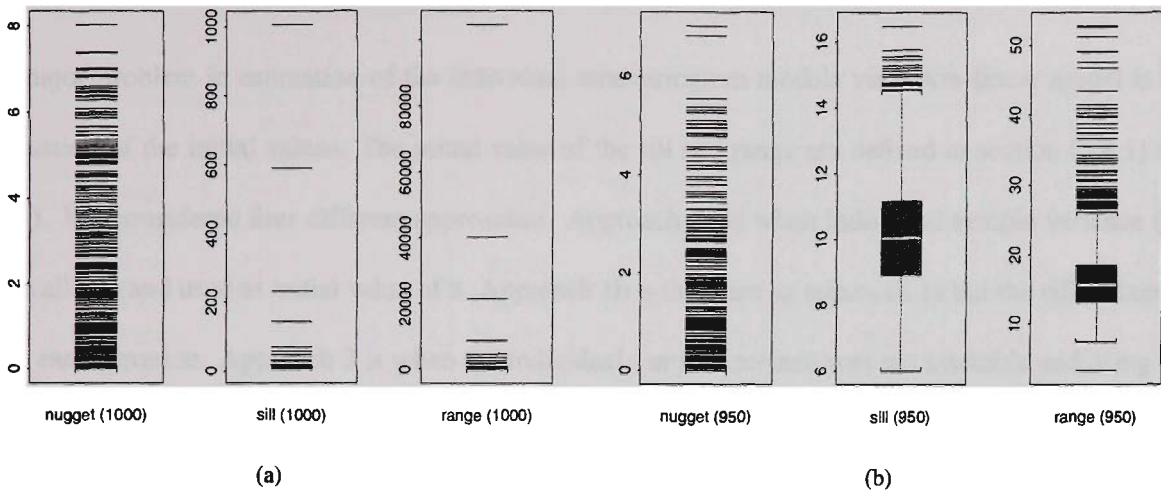


Figure 5.19. Distribution of the group level semivariogram estimated parameters nugget, sill, and range, (a) 1000 simulations (b) 950 simulation

These results clearly demonstrate the biases that rise from using the group level semivariogram to estimate the parameters of the individual level semivariogram. The look of bias for the nugget may be due to it being 0 at the individual level. A non-zero nugget is considered later.

5.4.3 Output of the estimation process using group level data

Defining the model

The zero order model is considered as defined in (5.93) The exponential semivariogram model is implemented as defined in (5.96). Let us consider the approximation for \bar{d}_g and \bar{d}_h given in (5.68), then we have

$$\Gamma_{gh} = s \left(\frac{1}{2N_g} - \frac{1}{2N_h} \right) - (s - n) \cdot \exp \left[\frac{-3d_{gh}}{r} \right] + \frac{N_g - 1}{2N_g} (s - n) \cdot \exp \left[\frac{-3k_1 \sqrt{A_g}}{r} \right] + \frac{N_h - 1}{2N_h} (s - n) \cdot \exp \left[\frac{-3k_1 \sqrt{A_h}}{r} \right] \quad (5.113)$$

We have observations of the $\hat{\Gamma}_{gh}$ as empirical group level semivariogram values, d_{gh} as a distance between centroid of the groups g and h , N_g and N_h as the size of group g and h , A_g and A_h as the areas of group g and h , respectively. The unknown parameters of the individual level semivariogram n , s , and r can be estimated by applying a non-linear regression method.

The initial values of n , s , and r

The major problem in estimation of the individual semivariogram models via a non-linear model is determination of the initial values. The initial value of the sill and range are defined in section (5.4.1) and (5.4.1). We considered four different approaches. Approach 1a is when individual sample variance (s^2) was available and used as initial value of s . Approach 1b is the same as approach 1a but the sill is fixed at s^2 for each iteration. Approach 2 is when the individual sample variance was not available and using the initial value of the sill and range such as described in section (5.4.1). The approach 3 used the estimated group level parameters sill and range as the initial value of the respective parameter.

There are two cases for the initial value of the nugget, one is equal to the initial value of the sill and the other is equal to zero. The mean of the differences of these two cases on the estimated parameters are shown in Table (5.8). These results demonstrate that initial value of the nugget can be determined as zero or equal to the initial value of the sill.

Table 5.8. The mean of difference between the estimated parameters based on two different initial values of the nugget, equal the sill and zero

| Parameter | Approach 1a (std. err.) | Approach 1b (std. err.) | Approach 2 (std. err.) | Approach 3 (std. err.) |
|-------------------|----------------------------|----------------------------|---------------------------|---------------------------|
| Nugget difference | -0.0305 (0.018) | 0.0 (0.011) | 1.9372 (1.208) | 0.0140 (0.003) |
| Sill difference | -0.0437 (0.017) | – | 1.8201 (1.157) | 0.0017 (0.037) |
| Range difference | -0.0665 (0.513) | 0.0002 (0.047) | -1.4074 (1.121) | -0.2051 (0.161) |

To check if this conclusion was due to the individual level nugget being zero, another population of exponential semivariogram model with nugget equal to 5, sill equal to 20, and range equal to 13 was generated. There were 400 simulations done. The result of the estimated parameters is found in table (5.9). These results show that whether the initial value of the nugget was equal to zero or equal to the initial value of the sill there was little difference. Therefore we have two alternatives for the initial value of the nugget, that is equal to zero or equal to initial value of the sill.

Individual sample data are available

In this case we can use approach 1a and approach 1b. The initial value of the nugget is defined to be zero. Figure (5.20) shows boxplots of the estimated parameter of nugget, sill, and range. The figure shows

Table 5.9. The estimated parameters based on two different initial values of the nugget, equal the sill and zero

| Est. Par. | | approach 1a | | approach 1b | | approach 2 | | approach 3 | |
|-----------|----------|-------------|-------------|-------------|-------------|------------|-------------|------------|-------------|
| | | $n_0 = 0$ | $n_0 = s_0$ | $n_0 = 0$ | $n_0 = s_0$ | $n_0 = 0$ | $n_0 = s_0$ | $n_0 = 0$ | $n_0 = s_0$ |
| \hat{n} | mean | 9.153 | 9.509 | 4.008 | 3.968 | 11.246 | 10.131 | 9.315 | 9.315 |
| | median | 0.0 | 0.0 | 3.827 | 3.794 | 4.934 | 1.328 | 0.0 | 0.0 |
| | std.err. | 13.627 | 13.938 | 3.287 | 3.282 | 14.904 | 14.027 | 13.549 | 13.548 |
| \hat{s} | mean | 23.622 | 23.803 | 19.598 | 19.598 | 24.269 | 23.079 | 23.065 | 23.069 |
| | median | 19.370 | 19.573 | 19.536 | 19.536 | 19.659 | 18.395 | 18.662 | 18.662 |
| | std.err. | 10.645 | 11.060 | 2.554 | 2.554 | 12.253 | 11.805 | 11.039 | 11.036 |
| \hat{r} | mean | 14.953 | 15.320 | 13.884 | 13.886 | 18.270 | 21.076 | 16.047 | 16.002 |
| | median | 13.105 | 13.673 | 12.517 | 12.575 | 15.545 | 16.233 | 14.722 | 14.722 |
| | std.err. | 7.467 | 7.502 | 5.528 | 5.520 | 9.482 | 12.693 | 7.212 | 7.131 |

n_0 is initial value of nugget and s_0 is initial value of sill.

True value $n = 5, s = 20, r = 13$.

that generally the location of the distribution of the estimator of nugget, sill, and range are close to the parameter value of $n = 0, s = 20$, and $r = 10$. Although we may observe some extreme values in the estimates for each parameter. Summary statistics are given in table (5.10). Compared with the group level estimated parameters given in table 5.6, the median for approach 1a is much closer to the individual level median for the sill and range. The means for the estimates of the sill and range are larger than the median, as they are affected by a long positive tail in the distribution. The mean of the estimated nugget is affected by some large positive values, but the median is still zero. However, approach 1a gives increased standard deviation of the estimated parameters compared with their corresponding estimated parameters in table (5.6). Meanwhile, setting the sill estimate to the sample variance improves the estimation of nugget, sill and range considerably, as indicated in the standard deviation of the estimated parameters. The standard deviation of the nugget, sill and range are reduced compared with the approach 1a and the means are considerably improved. Therefore the approach 1b gives some improvement on the estimated parameter of nugget, sill and range. As we can observe from Figure (5.20), there is considerably less variability in the estimates.

Table (5.11) shows the correlations among the estimated parameters and the initial values. The correlation of the parameters with their initial values are small for the approach 1a. But in approach 1b, the correlation between the initial value of the range and its estimate is moderate (0.4974). The estimated nugget is strongly correlated with the estimated sill (0.9713) in approach 1a. Meanwhile, in approach 1b

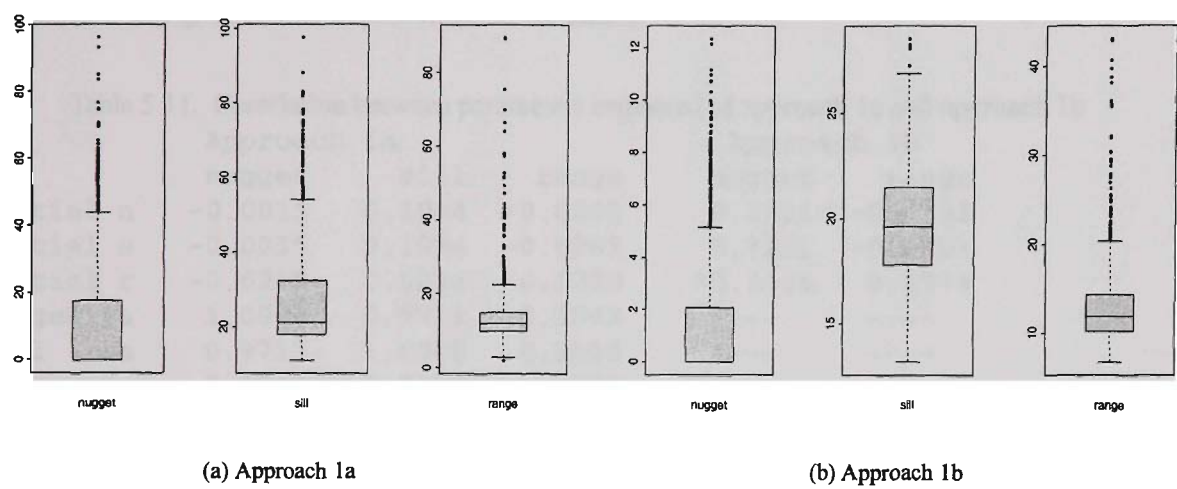


Figure 5.20. Distribution of parameters estimated by approach 1a and approach 1b

Table 5.10. Descriptive values of the Fig. 5.20

| Approach 1a | Mean | Median | Minimum | Maximum | Standard error |
|----------------------|---------|---------|---------|---------|----------------|
| Nugget (\hat{n}) | 10.4713 | 0.0 | 0.0 | 96.2560 | 17.7089 |
| Sill (\hat{s}) | 27.1393 | 21.3595 | 11.3900 | 87.6200 | 14.1134 |
| Range (\hat{r}) | 14.2607 | 11.9890 | 1.8080 | 89.5070 | 8.7364 |
| Approach 1b | | | | | |
| Nugget (\hat{n}) | 1.3923 | 0.0 | 0.0 | 12.3270 | 2.4366 |
| Sill (\hat{s}) | 19.733 | 19.618 | 13.176 | 28.529 | 2.714 |
| Range (\hat{r}) | 13.0438 | 11.8810 | 6.7240 | 43.0900 | 4.7102 |

True value $n = 0, s = 20, r = 10$

the estimated nugget is moderately correlated with the estimated range (0.4925). Therefore, the overall performance of the approach 1a is good, since it can be reduce the biases significantly for estimation of the sill and range, but it has high standard deviation. The approach 1b reduces the standard deviation of the estimates, and gives further improvement in the biases.

Table 5.11. Correlation between parameters estimated of approach 1a and approach 1b

| | Approach 1a | | | Approach 1b | |
|-----------|-------------|---------|---------|-------------|---------|
| | nugget | sill | range | nugget | range |
| initial n | -0.0015 | 0.1094 | -0.0205 | 0.3221 | -0.0793 |
| initial s | -0.0015 | 0.1094 | -0.0205 | 0.3221 | -0.0793 |
| initial r | -0.0289 | 0.0234 | 0.1230 | -0.3164 | 0.4974 |
| nugget 1a | 1.0000 | 0.9713 | 0.3842 | ---- | ---- |
| sill 1a | 0.9713 | 1.0000 | 0.3038 | ---- | ---- |
| range 1a | 0.3842 | 0.3038 | 1.0000 | ---- | ---- |
| nugget 1b | -0.0688 | -0.1379 | 0.2921 | 1.0000 | 0.4925 |
| range 1b | 0.0322 | 0.0062 | 0.4846 | 0.4925 | 1.0000 |

Individual sample data are not available

This situation will use approach 2a where initial values are obtained from (5.110) and (5.112), and the option of using the estimated group level parameters as the initial value (approach 3a). The initial value of the nugget is defined to be zero.

The distribution of the estimated parameters are exhibited in Figure (5.21) and described in table (5.12). The distributions of the estimated parameters indicate a marked skewness. Hence the median is an appropriate statistics to be used in evaluating the method. The approaches show very similar results for the nugget and sill with medians close to the individual level parameters . But approach 3 give a better result for the estimation of the range. The approach 3 also shows a smaller standard deviation than approach 2, particularly for the range.

This illustration demonstrates that the individual level semivariogram parameters can be estimated through the non-linear regression method by using the estimated group level semivariogram parameters as the initial value of the nugget, sill, and range. Some implications will be mentioned later in this section.

The performance of approach 3 is comparable with that of approach 1a, except for a higher standard deviation for the range estimate. Hence, these is no great advantage in using the individual level data,

unless we use it to estimate the sill at each iteration (approach 1b). In that case the standard deviations of the estimates are greatly reduced.

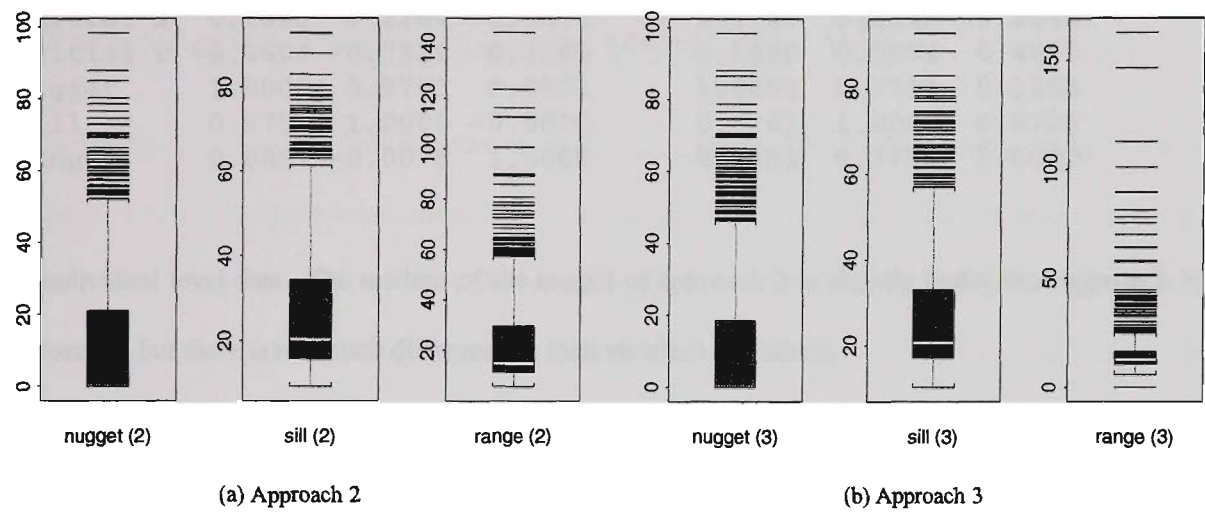


Figure 5.21. Distribution of parameters estimated by approach 2 and approach 3

Table 5.12. Description of distribution of the estimated parameters by approach 2 and approach 3

| Parameter | Mean | Median | Minimum | Maximum | Standard error |
|------------|---------|---------|---------|----------|----------------|
| Approach 2 | | | | | |
| Nugget (n) | 12.0337 | 0.0 | 0.0 | 98.4740 | 18.6173 |
| Sill (s) | 26.9495 | 20.1680 | 9.1830 | 92.9540 | 15.7444 |
| Range (r) | 22.8640 | 14.9005 | 5.6790 | 146.2850 | 18.0290 |
| Approach 3 | | | | | |
| Nugget (n) | 10.7540 | 0.0 | 0.0 | 98.4310 | 18.0681 |
| Sill (s) | 26.9703 | 20.6455 | 10.2710 | 92.9340 | 14.7738 |
| Range (r) | 15.4472 | 12.6730 | 0.0 | 162.5730 | 11.8291 |

True value n = 0, s = 20, r = 10

Table (5.13) shows correlation matrix of the estimated parameters. Correlation between initial values with their estimated are small for both approach. Except the correlation between initial value of range with its estimated range in approach 3 is moderately strong at 0.4693. The estimated sill is strongly correlated (0.9753) with the estimated nugget for both approach.

Table (5.9) shows similar median values for the sill and range between approach 2 and approach 3. They are also comparable with the median of the approach 1. Their standard deviations are also the same, hence this situation confirms with the previous simulation results, that there is no great advantage in using

Table 5.13. Correlation between parameters estimated of approach 2 and approach 3

| | Approach 2 | | | Approach 3 | | |
|-----------|------------|---------|---------|------------|--------|--------|
| | nugget | sill | range | nugget | sill | range |
| initial n | 0.1093 | 0.2264 | 0.0677 | 0.0665 | 0.0172 | 0.3844 |
| initial s | 0.1093 | 0.2264 | 0.0677 | 0.1727 | 0.2967 | 0.2044 |
| initial r | -0.3604 | -0.3929 | 0.3345 | 0.0698 | 0.0404 | 0.4693 |
| nugget | 1.0000 | 0.9753 | 0.0851 | 1.0000 | 0.9743 | 0.3353 |
| sill | 0.9753 | 1.0000 | -0.0076 | 0.9743 | 1.0000 | 0.2789 |
| range | 0.0851 | -0.0076 | 1.0000 | 0.3353 | 0.2789 | 1.0000 |

the individual level data. The median of the nugget of approach 2 is slightly better than approach 3 or approach 1, but there is not much difference in their standard deviations.

Implications

Estimation of the individual level semivariogram parameters using the group level data shows some success, although some outliers exist, either for nugget, sill, or range. Having a good estimates of the individual level semivariogram parameters, may lead to some implications,

- The individual level nugget may indicate the variation between individuals at very close distance, for instance within household. This value may lead to exploring within household correlation.
- The individual level sill provides an estimate of the individual level variance. It may give an adjustment of the individual level variance from a sample or from group level variance, and can be used to explore the aggregation effect.
- The individual level range may give a new perspective on spatial autocorrelation. Then we could examine the intercorrelation between individuals within the region in terms of distance.
- The information on the individual level semivariogram can help in explaining the Modifiable Areal Unit Problem (MAUP) using the result in chapter (4).

Given group level data and group’s centroid location, we may develop a framework to estimate parameters of the individual level semivariogram model. The non-linear estimation procedure can be applied on two different situation, when the individual sample is available and not available. In the first situation, keeping the sill fixed factor at the sample s^2 is a better alternative, leading to a smaller standard deviation

on the estimates. In some cases, the second situation is more realistic, and can provide reasonable adjusted estimates but it gives more variable estimates than the first one.

5.5 The weighted version of group level semivariogram

In this section we investigate whether use of weighting factors in the group level semivariogram can be applied to produce less biased estimates. The main feature of the group level semivariogram is that it is shifted down compared with the individual level semivariogram. The purpose of weighting is to adjust the group level semivariogram so that it is closer to individual level semivariogram. We saw in theorem (4.3.4) that the unweighted average of $\hat{\Gamma}_{gh}$ is ${}_1\bar{S}_{yy}$ and that the weighted average of $\hat{\Gamma}_{gh}$, using weight $N_g N_h$ is ${}_N\bar{S}_{yy}$. We know that generally ${}_1\bar{S}_{yy}$ will be less than S_{yy} which will probably be close to the sill of the individual level semivariogram. Also ${}_N\bar{S}_{yy}$ will usually be larger than S_{yy} . Hence, there may be some weighting that gets close to the individual level.

The weighting factors can be applied in situations where there is only limited software or hardware resources that make performing non-linear regression analysis difficult. This gives an advantage to the weighting factors analysis over the non-linear regression analysis. However, the use of weighting factors may not adequately adjust for the biases involved in using group level data.

For groups g and h define the empirical weighted group level semivariogram as

$${}_N\hat{\Gamma}_{gh} = \frac{1}{2} \left(\sqrt{N_g}(\bar{Y}_g - \bar{\bar{Y}}) - \sqrt{N_h}(\bar{Y}_h - \bar{\bar{Y}}) \right)^2; \quad \text{where } \bar{\bar{Y}} = \frac{1}{M} \sum_g \bar{Y}_g \quad (5.114)$$

and its average as

$${}_N\bar{\hat{\Gamma}} = \frac{1}{M(M-1)} \sum_{g \neq h} {}_N\hat{\Gamma}_{gh} \quad (5.115)$$

Theorem 5.5.1. *The mean of the weighted group level semivariogram is equal to the weighted group level variance, that is*

$${}_N\bar{\hat{\Gamma}} = {}_N\bar{S}_{yy} \quad (5.116)$$

Proof. Equation (5.114) can be modified into

$$\begin{aligned}
 {}_N\hat{\tilde{\Gamma}} &= \frac{1}{M(M-1)} \sum_{g \neq h} \frac{1}{2} \left(\sqrt{N_g}(\bar{Y}_g - \bar{\bar{Y}}) - \sqrt{N_h}(\bar{Y}_h - \bar{\bar{Y}}) \right)^2 \\
 &= \frac{1}{2M(M-1)} \sum_{g \neq h} \left(N_g(\bar{Y}_g - \bar{\bar{Y}})^2 + N_h(\bar{Y}_h - \bar{\bar{Y}})^2 - 2\sqrt{N_g N_h}(\bar{Y}_g - \bar{\bar{Y}})(\bar{Y}_h - \bar{\bar{Y}}) \right) \\
 &= \frac{1}{2M(M-1)} \left((M-1) \sum_g N_g(\bar{Y}_g - \bar{\bar{Y}})^2 + (M-1) \sum_h N_h(\bar{Y}_h - \bar{\bar{Y}})^2 \right. \\
 &\quad \left. - 2 \sum_{g \neq h} \sqrt{N_g N_h}(\bar{Y}_g - \bar{\bar{Y}})(\bar{Y}_h - \bar{\bar{Y}}) \right) \\
 &= \frac{1}{2M(M-1)} \left(2(M-1)^2 {}_N\bar{S}_{yy} - 2 \sum_{g,h} \sqrt{N_g N_h}(\bar{Y}_g - \bar{\bar{Y}})(\bar{Y}_h - \bar{\bar{Y}}) \right. \\
 &\quad \left. + 2 \sum_{g=h} \sqrt{N_g N_h}(\bar{Y}_g - \bar{\bar{Y}})(\bar{Y}_h - \bar{\bar{Y}}) \right) \\
 &= \frac{1}{2M(M-1)} \left(2(M-1)^2 {}_N\bar{S}_{yy} - 0 + 2(M-1) {}_N\bar{S}_{yy} \right) \\
 &= {}_N\bar{S}_{yy}
 \end{aligned}$$

□

Therefore,

Theorem 5.5.2. *Expectation of the mean of the weighted group level semivariogram is*

$$E({}_N\hat{\tilde{\Gamma}}) = E(S_{yy}) \left(1 + \frac{1}{M-1} \sum_g (N_g - 1) \bar{\rho}_g \right) \quad (5.117)$$

Proof. Applying (5.8) we have

$$\bar{y}_g = \sigma^2(1 - \bar{\rho}_g) \quad (5.118)$$

Substituting (5.118) into (4.23), we have

$$\begin{aligned}
 E({}_N\hat{\tilde{\Gamma}}) &= E({}_N\bar{S}_{yy}) \\
 &= \frac{1}{M-1} \left((N-1)\bar{y} - \sum_g (N_g - 1)\bar{y}_g \right) \\
 &= \frac{1}{M-1} \left((N-1)E(S_{yy}) - \sum_g (N_g - 1)\sigma^2 \cdot (1 - \bar{\rho}_g) \right)
 \end{aligned}$$

Assume the σ^2 is equal to the $E(S_{yy})$ (see section 5.4, then

$$E({}_N\hat{\tilde{\Gamma}}) = \frac{1}{M-1} \left((N-1)E(S_{yy}) - E(S_{yy}) \sum_g (N_g - 1)(1 - \bar{\rho}_g) \right)$$

Simplifying this will complete the proof

□

Corollary 5.5.3. *Assume spatial autocorrelation within the groups to be constant at $\bar{\rho}$, then*

$$E({}_N\tilde{\Gamma}) = E(S_{yy}) \left(1 + (\bar{N} - 1)\bar{\rho} \frac{M}{M-1} \right) \quad (5.119)$$

So we expect that the average of the weighted group level semivariogram to be above the individual level semivariogram provided the $\bar{\rho} > 0$. This expectation shows that the weighting factors should depend on the spatial correlation within the group. Consider using $w_{gh}(\bar{Y}_g - \bar{Y}_h)^2$ for some weights w_{gh} . Some idea of the appropriate weight can be obtained by considering what close estimate σ^2 when d_{gh} is large and we expect the group means to be uncorrelated. Hence we may develop the following theorem by assuming that d_{gh} is large, that is

Theorem 5.5.4. *The weighting factor in general is defined by*

$$w_{gh}^{-1} = \frac{1}{N_h}(1 - \bar{\rho}_h) + \frac{1}{N_g}(1 - \bar{\rho}_g) + \bar{\rho}_h + \bar{\rho}_g \quad (5.120)$$

Proof. Consider d_{gh} is large such that $C(\bar{Y}_g, \bar{Y}_h) = 0$, then

$$\begin{aligned} E(\bar{Y}_g - \bar{Y}_h)^2 &= V(\bar{Y}_g) + V(\bar{Y}_h) \\ &= \sigma^2 \left(\frac{1}{N_h}(1 + (N_h - 1)\bar{\rho}_h) + \frac{1}{N_g}(1 + (N_g - 1)\bar{\rho}_g) \right) \end{aligned}$$

Hence

$$E(w_{gh} \cdot (\bar{Y}_g - \bar{Y}_h)^2) = \sigma^2$$

for the w_{gh} ,

$$w_{gh}^{-1} = \left(\frac{1}{N_h}(1 + (N_h - 1)\bar{\rho}_h) + \frac{1}{N_g}(1 + (N_g - 1)\bar{\rho}_g) \right)$$

Modifying this will complete the proof □

Consider $\bar{\rho}_h = \bar{\rho}_g = 0$, then we have weighting factor

$$w_{gh} = \left(\frac{1}{N_g} + \frac{1}{N_h} \right)^{-1} \quad (5.121)$$

and hence

$$\widehat{\Gamma}_{gh}^{w2} = w_{gh} \cdot \widehat{\Gamma}_{gh} \quad (5.122)$$

If $\bar{\rho}_g = \bar{\rho}_h = 1$ then $w_{gh} = \frac{1}{2}$, as used in the unweighted group level semivariogram.

5.5.1 Illustration from simulated data

The simulation process is started by generating the individual observation according to a particular semivariogram model. The population will have a characteristic exponential semivariogram model with parameter nugget=15, sill=30, and range=20. This population has different parameters from the previous population, for the purpose of giving some variation in handling the estimation of the parameters. An other purpose was that to avoid the zero nugget. This is represented by the following model,

$$\gamma(d_{ij}) = 15 + 15 \left(1 - \exp \left[\frac{-3d_{ij}}{20} \right] \right) \tag{5.123}$$

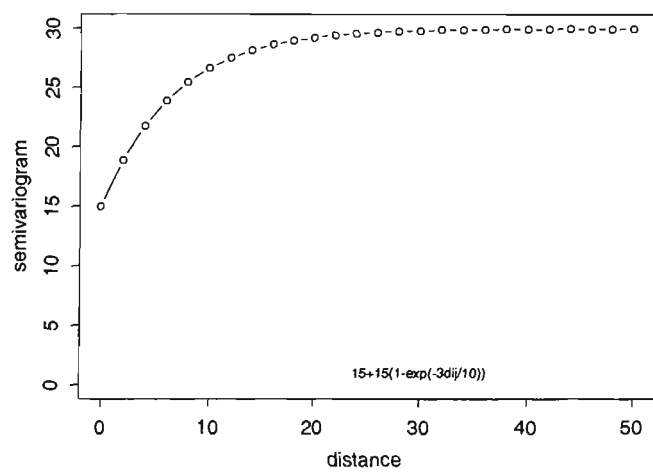


Figure 5.22. Exponential model of $\gamma(d_{ij}) = 15 + 15(1 - \exp(-3\frac{d_{ij}}{20}))$

The population consists of 2000 individuals, within a rectangular region with boundary (20,20) at lower left and (90,80) at upper right point. The second step is to partition the region into groups and creating the group level data according to their subregion definition. For this simulation purpose approximately 250 groups were created, hence $\bar{N} = \pm 8$. The impression of the simulation is shown in figure (5.23), which are a realization from two particular simulations.

Figure (5.23) shows two realization of the simulation. The figure indicates a comparison between the individual level semivariogram (o) and the unweighted group level semivariogram (●). The graph shows a turning point at a distance value around 20, which has meaning that the range is equal to 20. Table (5.14) presents the parameter estimates of the individual and unweighted group level semivariogram for those two

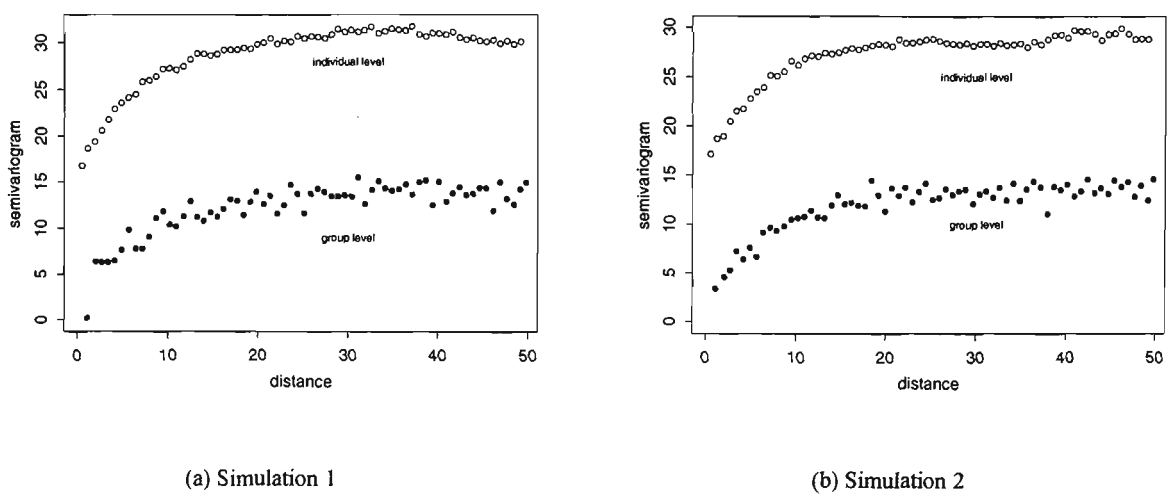


Figure 5.23. The first two simulation results. Note : \circ =individual level and \bullet = group level

simulations respectively. Again we see an evidence of negative bias for the nugget and sill and a positive bias for the range, in using the unweighted group level semivariogram.

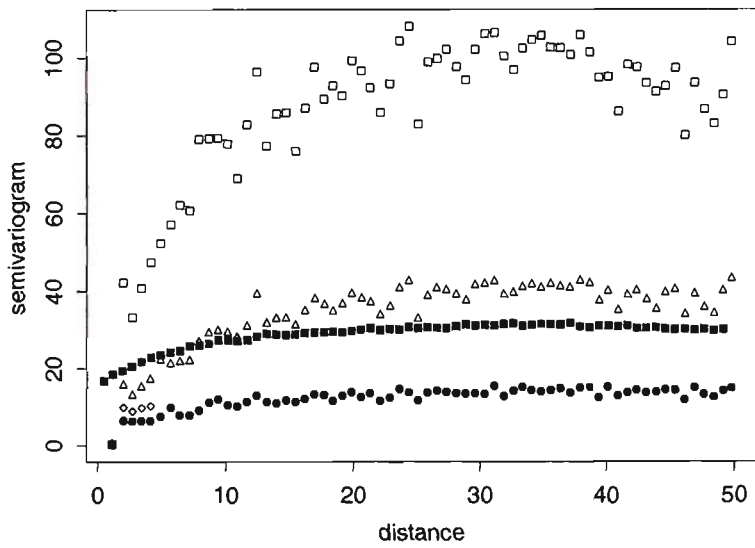
Table 5.14. The estimated parameter of the model (5.123) of individual and unweighted group level semi-variogram

| Simulation | Individual | | | Unweighted group | | |
|------------|-----------------------|---------------------|----------------------|-----------------------|---------------------|----------------------|
| | nugget (std. err.) | sill (std. err.) | range (std. err.) | nugget (std. err.) | sill (std. err.) | range (std. err.) |
| 1 | 16.253 (0.358) | 30.953 (0.064) | 22.010 (0.630) | 2.921 (0.942) | 14.106 (0.327) | 26.485 (3.341) |
| 2 | 16.110 (0.563) | 28.872 (0.095) | 19.082 (0.901) | 2.015 (0.943) | 13.480 (0.331) | 21.551 (3.427) |

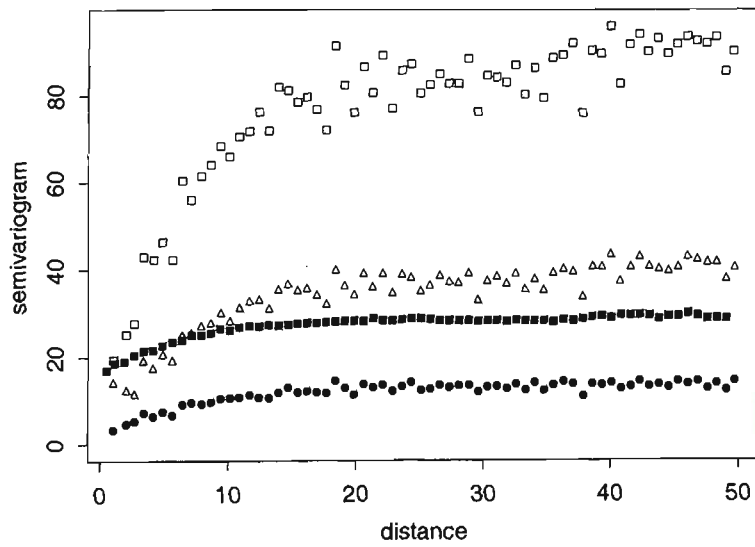
Two formulations of weighting factor of the group level semivariogram are considered as shown in table (5.15). Figure (5.24) shows the result of these weighting factors from the two simulations. They are compared with the unweighted group level semivariogram and individual semivariogram.

Table 5.15. The weighting factor

| Description | formulation | weighting factor |
|-----------------------------|--|---|
| 1. $N\hat{\Gamma}_{gh}$ | $\frac{1}{2} \left(\sqrt{N_g}(\bar{Y}_g - \bar{\bar{Y}}) - \sqrt{N_h}(\bar{Y}_h - \bar{\bar{Y}}) \right)^2$ | — |
| 2. $\hat{\Gamma}_{gh}^{w2}$ | $\left(\frac{1}{N_g} + \frac{1}{N_h} \right)^{-1} \cdot \hat{\Gamma}_{gh}$ | $\left(\frac{1}{N_g} + \frac{1}{N_h} \right)^{-1}$ |



(a) Simulation 1



(b) Simulation 2

Figure 5.24. The first two simulation results of the weighting factors. Note : ■ = individual level and • = unweighted group level, □ = ${}_N\hat{\Gamma}_{gh}$, $\Delta = \hat{\Gamma}_{gh}^{w2}$.

The simulations are repeated 110 times, which is determined arbitrarily. But it is expected to capture a lower standard error of the estimate. The result shows clearly that the unweighted group level semivariogram is located below the individual level semivariogram. The nugget and sill of the unweighted group level semivariogram are below, but the range of the unweighted group level semivariogram is above the individual level, see figure (5.25). Some statistics of figure (5.25) are shown in table (5.16). The table shows that the aggregation process may increase the parameter range, but reduce the nugget and sill parameters.

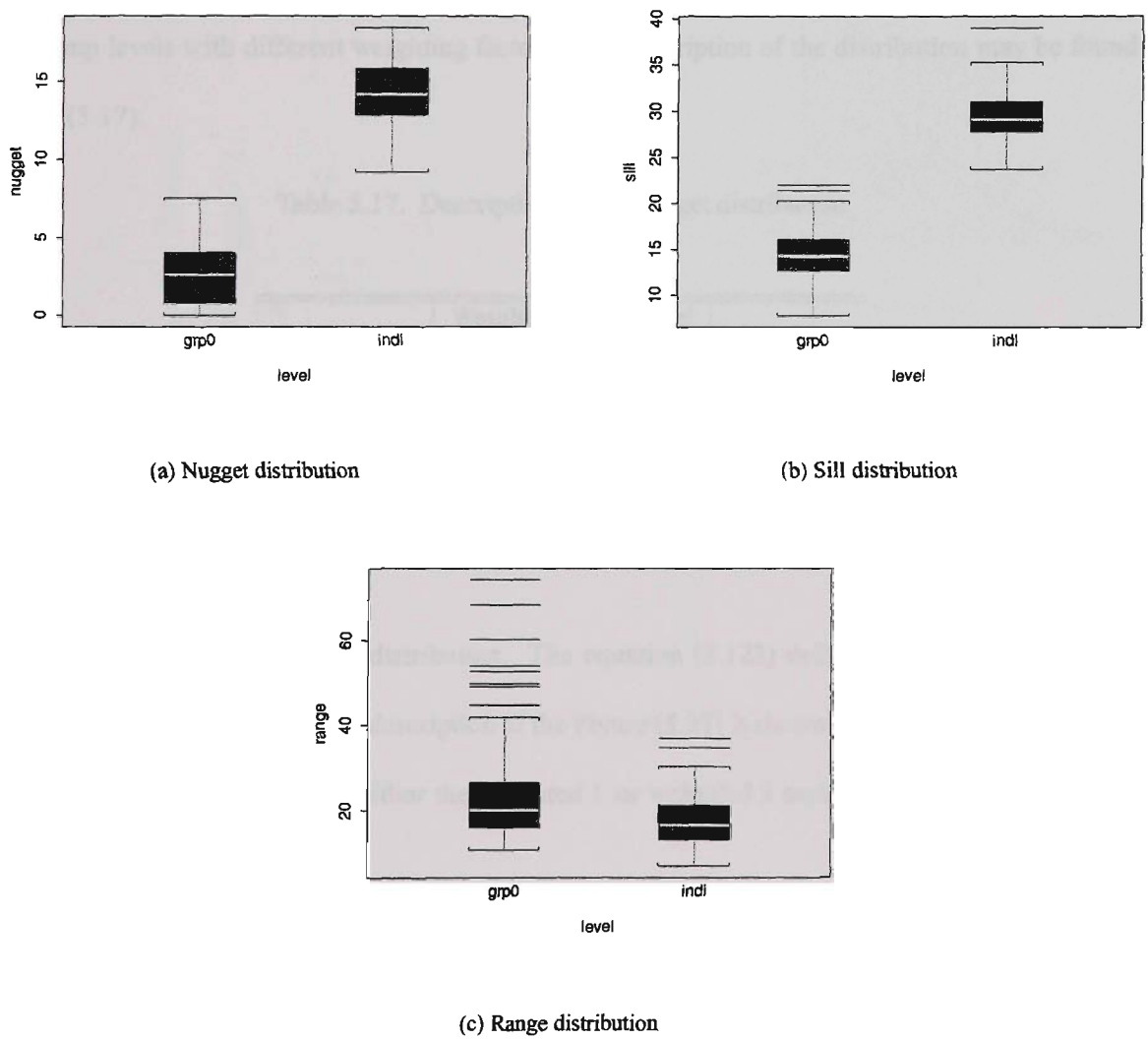


Figure 5.25. Semivariogram model parameter estimator distribution

Figures (5.26) show the nugget distribution for the individual level, unweighted and the weighted group level semivariogram from 110 simulations. Figure (5.26a) shows the nugget distribution of the individual and weighted group level. Figures (5.26b-f) indicate the nugget distribution for individual level, and each

Table 5.16. Description of the nugget, sill, and range distribution, of the individual and unweighted group level semivariogram model.

| Description | Nugget (15) | | Sill (30) | | Range (20) | |
|-------------|-------------|-------------|-----------|-------------|------------|-------------|
| | Indiv. | Unwght grp. | Indiv. | Unwght grp. | Indiv. | Unwght grp. |
| Mean | 14.2749 | 2.6534 | 29.4662 | 14.3085 | 17.9203 | 23.9069 |
| Median | 14.1760 | 2.5970 | 29.1105 | 14.2405 | 16.5565 | 20.0290 |
| Min. | 9.1920 | 0.0 | 23.6970 | 7.7740 | 6.9630 | 10.7190 |
| Max. | 18.4370 | 7.5440 | 38.9870 | 22.0190 | 36.9040 | 74.0210 |
| Std. err. | 1.9127 | 2.0576 | 2.5849 | 2.6206 | 5.9954 | 12.0998 |

of the group levels with different weighting factors. The description of the distribution may be found in the Table (5.17).

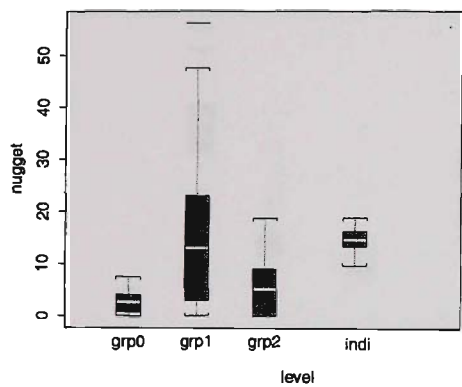
Table 5.17. Description of the nugget distribution

| Description | Weighted group level | | individual |
|-------------|----------------------|----------|------------|
| | weight 1 | weight 2 | |
| Mean | 14.6518 | 5.4985 | 14.2749 |
| Median | 13.0270 | 5.1835 | 14.1760 |
| Minimum | 0.0 | 0.0 | 9.1920 |
| Maximum | 56.2570 | 18.8380 | 18.4370 |
| Std. err. | 13.3537 | 5.0315 | 1.9127 |

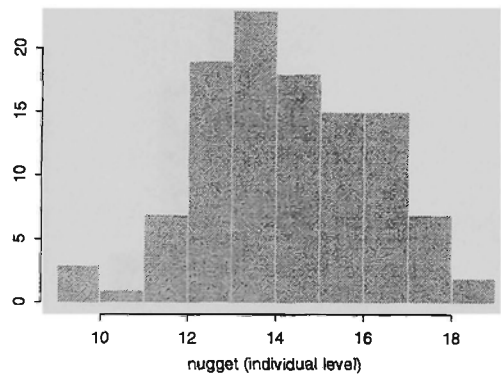
Figure (5.27) indicates the sill distribution. The equation (5.123) defines the sill parameter of the semivariogram model to be 30. The description of the Figure (5.27) is shown in the Table (5.18). The table shows that the estimated sill from either the weighted 1 or weighted 2 analysis are above the individual level sill.

Table 5.18. Description of the sill distribution

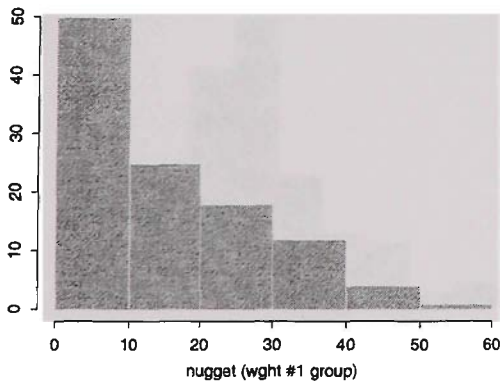
| Description | Weighted group level | | individual |
|-------------|----------------------|----------|------------|
| | weight 1 | weight 2 | |
| Mean | 94.9768 | 41.9799 | 29.4662 |
| Median | 91.3440 | 40.7360 | 29.1105 |
| Min. | 59.6860 | 24.6060 | 23.6970 |
| Max. | 142.7850 | 87.1430 | 38.9870 |
| Std. err. | 17.7087 | 8.4010 | 2.5849 |



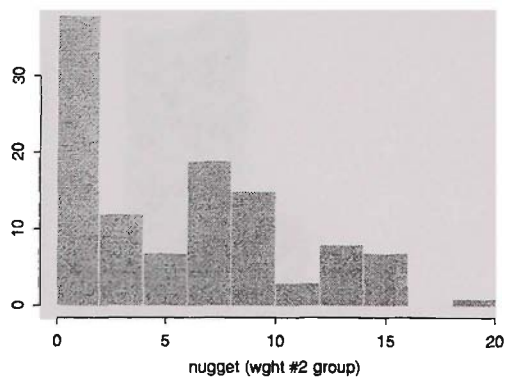
(a) Individual and weighted group



(b) Individual level

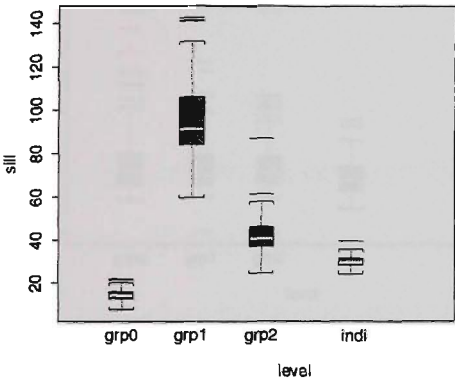


(c) Weighted group - 1

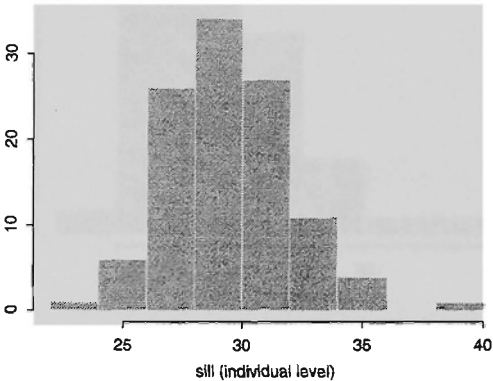


(d) Weighted group - 2

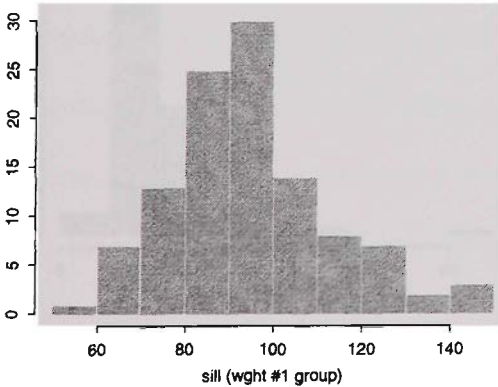
Figure 5.26. Nugget Distribution of individual level, unweighted and weighted group level



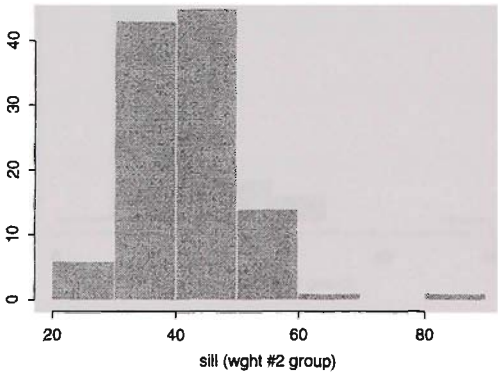
(a) Individual and weighted group



(b) Individual level



(c) Weighted group - 1



(d) Weighted group - 2

Figure 5.27. Sill Distribution of the individual level, unweighted and weighted group level

Figure (5.27b) shows the sill distribution of the individual level data. Table (5.18) shows that it has a mean 29.4662 with standard error 2.5849. Applying weighting factor 1 and 2 give a higher result of the sill value (mean 94.9768 and 41.9799, respectively). Graphically, the facts are showing in Figure (5.27a, 5.27c, 5.27d).

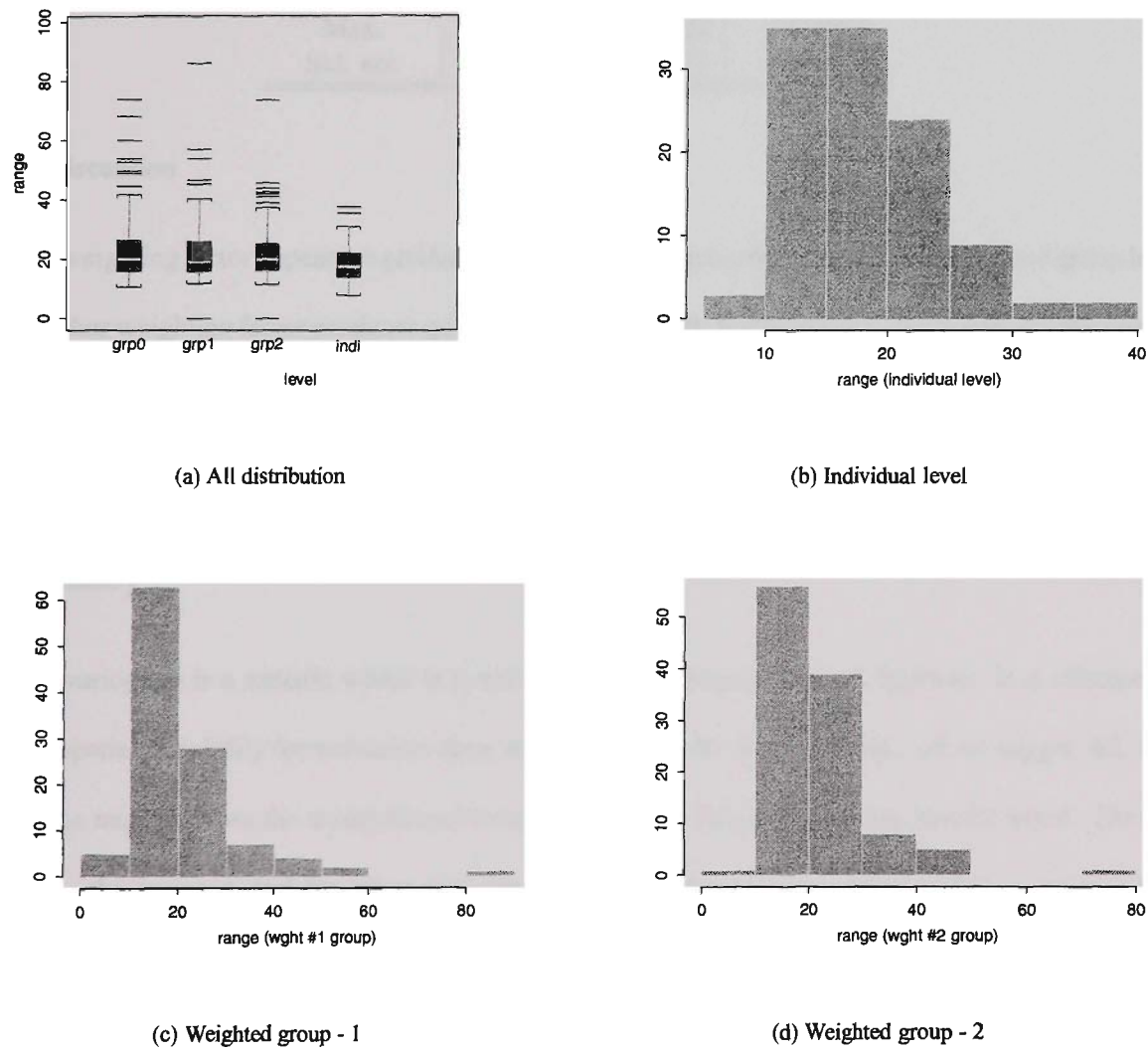


Figure 5.28. Range Distribution of the individual level, unweighted and weighted group level

Figure (5.28) displays the distribution of the range parameter of the semivariogram model. The model (equation 5.123) showed the true range equal to 20. The simulation result indicates the range distribution is close to the true parameter value (20) for all those cases, and it is shown in Figure (5.28a-b). The description of the Figure (5.28) is tabulated in Table (5.19).

Table 5.19. Description of the range distribution

| Description | Weighted group level | | individual |
|-------------|----------------------|----------|------------|
| | weight 1 | weight 2 | |
| Mean | 21.6842 | 22.2380 | 17.9203 |
| Median | 19.1490 | 19.8655 | 16.5565 |
| Min. | 0.0 | 0.0 | 6.9630 |
| Max. | 86.2180 | 73.7520 | 36.9040 |
| Std. err. | 11.5323 | 9.3177 | 5.9954 |

5.5.2 Discussion

The first weighting factor appears to produce better nugget estimate, than using the unweighted group level data. Neither weighting factor produces good estimate of the sill. In this example the range was reasonably estimated from all the group level semivariogram. Improvements in the weighting approach would need to obtain some estimates of $\bar{\rho}_g$ and $\bar{\rho}_h$ to use in (5.120).

5.6 Summary

The semivariogram is a statistic which is usually modeled through a distance function. It is effective in showing spatial variability for univariate data, which is shown by its parameters such as nugget, sill, and range. The nugget shows the measurement error at very close distance or at one location point. The sill is equal to the variance, and the range indicates a distance when the observations become independent between each other.

The group level semivariogram, which is derived from the group level data, may produce a bias compared with their corresponding individual level semivariogram. The bias is studied through a decomposition of group level semivariogram using a Taylor’s series expansion. The bias depends on the mean and variance of distance between pair of the points from two different groups, and within a group. These quantities depend on the size and shape of the groups, and approximations were developed by assuming the circle shape.

Estimating the individual level semivariogram parameters from the group level data gave an initial success. The main problem encountered was in determination of initial value for the estimated parameter.

Two situation were considered, such as the individual sample data was available, and the individual sample data was not available. The availability of individual sample data gave advantage in estimating the correct sill. When the individual sample data was not available then the estimated group level semivariogram parameter give an effective initial value of the estimated individual level semivariogram parameters.

Chapter 6

Cross-Semivariogram

Chapter 5 discussed univariate techniques for analysing social data based on semivariogram analysis. In investigating social phenomena one characteristic may be affected by other characteristics, or it may influence other characteristics. Therefore inter-relationships between characteristics may exist, for example in the bivariate case such as exhibited on figure (3.2). For example in the bivariate case the following model is sometimes used,

$$\mathbf{Y}_a = \alpha + \beta \cdot \mathbf{Y}_b + \mathbf{u} \tag{6.1}$$

where \mathbf{Y}_a is $N \times 1$ vector of observations of the dependent variable, \mathbf{Y}_b is the $N \times 1$ vector of values of the independent variable. The α and β are regression coefficients, and \mathbf{u} is the $N \times 1$ vector of errors. If the data are assumed to be IID then

$$E(\mathbf{u}\mathbf{u}^T) = \sigma^2 \cdot \mathbf{I} \tag{6.2}$$

but when there is dependency among observations, then we may have

$$E(\mathbf{u}\mathbf{u}^T) = \sigma^2 \cdot \mathbf{V} \tag{6.3}$$

where \mathbf{V} is a non-diagonal variance-covariance $N \times N$ matrix. When \mathbf{V} depends on the spatial location of the individuals, spatial variation exists and spatial analysis approaches might be applied.

The above would be an example of bivariate case which shows a model of relationship between independent and dependent variables. In the following discussion, we will consider a bivariate case which may portray a level of relationship between variables, hence the definition of independent and dependent are not needed. This chapter will develop methods for analysing relationships between characteristics using

group level social data. The characteristics are assumed to be spatially dependent each other. Methods will be developed based on cross-semivariogram analysis. The group level cross-semivariogram is investigated and some implications for the aggregation effect are discussed. The bivariate case will be the main focus of this discussion, though the method may be applied for the multivariate case. The results and methods obtained generalize those obtained for the univariate case in chapter 4 and 5.

6.1 Introduction

Consider a bivariate population such as defined in (4.1), that is $p = 2$. Hence we have $\mathbf{Y}_a = [Y_{a_1}, \dots, Y_{a_N}]^T$ and $\mathbf{Y}_b = [Y_{b_1}, \dots, Y_{b_N}]^T$, for example the employment status and the level of education. In another case, these variables can be categorized as the independent variable or dependent variable. We are also interested in analysis conditional on the locations \mathbf{L} . We assume that there exists the first and second moment structure for the conditional distribution of $\mathbf{Y}_a|\mathbf{L}$ and $\mathbf{Y}_b|\mathbf{L}$

$$\begin{aligned} (i) \quad & E(Y_{a_i}|\mathbf{L}) = \mu_{a_i}(\mathbf{L}) \\ (ii) \quad & Cov(Y_{a_i}; Y_{a_j} | \mathbf{L}) = \Delta_{a_{ij}}(\mathbf{L}) \\ (iii) \quad & Cov(Y_{a_i}; Y_{b_i} | \mathbf{L}) = \Sigma_{ab_i}(\mathbf{L}) \\ (iv) \quad & Cov(Y_{a_i}; Y_{b_j} | \mathbf{L}) = \Delta_{ab_{ij}}(\mathbf{L}), \quad \forall i \neq j \end{aligned} \tag{6.4}$$

The first two of (6.4) are the assumptions used in the univariate case, and (iii) and (iv) are the extensions for the bivariate case. The parameter Σ_{ab_i} reflects the covariance between variables a and b for the same individual, and is the ordinary covariance between two variables. The parameter $\Delta_{ab_{ij}}$ indicates covariance between two variables at two different locations, which Wackernagel (1998) called the cross-covariance. At this stage, we assume that the mean, covariance, and cross-covariance depend on the locations of all individuals in the population. For convenience we will omit the "L" from the expectation and parameters for the rest of this chapter.

In social phenomena some characteristics are usually related to some degree to other characteristics. This relationship is often presented in terms of the correlation coefficient of the variables for the same individual, that is

$$\rho_{ab_i} = \frac{\Sigma_{ab_i}}{\sqrt{\Sigma_{a_i} \cdot \Sigma_{b_i}}} \tag{6.5}$$

The IID assumption implies that ρ_{ab_i} are constant for every pairs individual. From a spatial perspective the values ρ_{ab_i} are no longer constant because the observations are not IID. Wackernagel (1988) noted that analysis of multivariate spatial data should take account of the relation between observations due to their geographical position and relation between characteristics or variables due to their partial redundancy. For

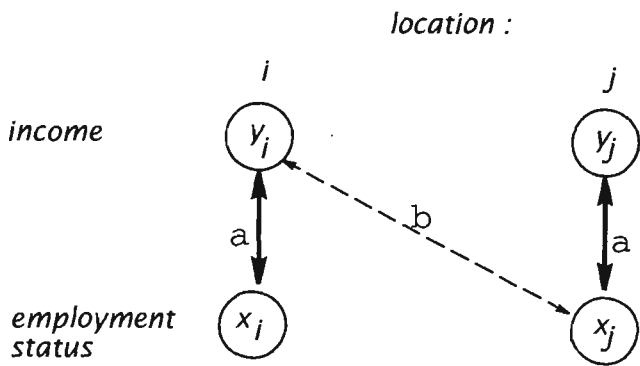


Figure 6.1. A diagram of relationship between income and employment status

example, consider the relationship between income and employment status. This relationship is shown graphically in figure (6.1). The figure shows two variables, income and employment status, which have two observations (y_i, x_i) and (y_j, x_j) , at two location ℓ_i , and ℓ_j , respectively. The income observation y_i at ℓ_i may be affected by its employment status x_i or its neighbor employment status x_j at ℓ_j .

Hence this figure shows two types of relationships between income and employment status. The correlation of type (a) is a conventional correlation that we consider when the IID assumption apply. But the correlation of type (b) occurs when there is dependency among observations based on their geographical locations. This correlation may show, for example, that high income for individual i may be associated with the employment status of individual i and also the employment status of individual j if the distance between location i and j are reasonably close. The association may be due to economic relationship between the locations, such as trade flows.

Hepple (1996, 1976) considered these issues by introducing a linear model. Wackernagel (1998, 1988), Ver Hoef and Cressie (1993) applied cross-semivariogram methods to investigate cross correlation between variables.

Myers (1988) considered multivariate analysis incorporating spatial correlation. He proposed the method of interpolation of points within the geographic region, that is a kriging in the univariate case and cokriging in the multivariate case. He illustrated the use of a multivariate technique such as principal component applied in conjunction with kriging method. Principal component analysis was introduced to simplify or to avoid the used of cokriging.

Wackernagel (1988) discussed some approaches for exploring the structure of spatially distributed multivariate data, using a combination of variogram modeling, principal component analysis, and cross-semivariogram modeling .

These authors have given the foundation of further analysis of multivariate spatial interpolation or prediction, as discussed in Ver Hoef and Cressie (1993). Prediction in the univariate case is known as kriging, and it is generalized by considering multivariate prediction in the form of cokriging. The corresponding statistics needed for these two prediction methods are variogram and cross variogram, respectively.

In this chapter, the discussion will be focused on the role of the cross-semivariogram in analysing the aggregation effect of social data.

6.2 Basic theorems of covariance

Define the matrix S be a variance-covariance matrix of the observations calculated from individual level data, which is defined by

$$S = Y^T \cdot A \cdot Y \quad (6.6)$$

where A is a $N \times N$ symmetric matrix, defined below. The Y is the $N \times p$ matrix of variables, which contains Y_a and Y_b . This equation is a generalization of (4.9) and (4.17).

6.2.1 Individual level covariances

The elements of A are

$$a_{ii} = \frac{1}{N}, \quad a_{ij} = \frac{-1}{N(N-1)} \quad (6.7)$$

The elements of \mathbf{S} are S_{aa} and S_{ab} , the variance of variable a and covariance between variable a and b , respectively. The statistic S_{aa} was defined in (4.17) and discussed in chapter (4). The statistic S_{ab} is defined as

$$\begin{aligned} S_{ab} &= \mathbf{Y}_a^T \cdot \mathbf{A} \cdot \mathbf{Y}_b \\ &= \frac{1}{N-1} \sum_{i \in \mathcal{U}} (Y_{a_i} - \bar{Y}_a) \cdot (Y_{b_i} - \bar{Y}_b) \end{aligned} \quad (6.8)$$

Using the same approach as in lemma (4.1.3), gives

$$E(S_{ab}) = \sum_{i \in \mathcal{U}} a_{ii} \text{Cov}(Y_{a_i}; Y_{b_i}) + \sum_{i \neq j \in \mathcal{U}} a_{ij} \text{Cov}(Y_{a_i}; Y_{b_j}) + S_{ab_{\mu\mu}} \quad (6.9)$$

where

$$S_{ab_{\mu\mu}} = \sum_{i \in \mathcal{U}} a_{ii} \cdot \mu_{a_i} \mu_{b_i} + \sum_{i \neq j \in \mathcal{U}} a_{ij} \cdot \mu_{a_i} \mu_{b_j}$$

Theorem 6.2.1. *The expectation of the individual level covariances between variables a and b is*

$$E(S_{ab}) = \bar{\Sigma}_{ab} - \bar{\Delta}_{ab} + S_{ab_{\mu\mu}} \quad (6.10)$$

where

$$\bar{\Sigma}_{ab} = \frac{1}{N} \sum_{i \in \mathcal{U}} \Sigma_{ab_i} \quad \text{and} \quad \bar{\Delta}_{ab} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \Delta_{ab_{ij}}$$

Proof. Substituting (6.7) into (6.9) gives

$$E(S_{ab}) = \sum_{i \in \mathcal{U}} \frac{1}{N} \Sigma_{ab_i} + \sum_{i \neq j \in \mathcal{U}} \frac{-1}{N(N-1)} \Delta_{ab_{ij}} + S_{ab_{\mu\mu}}$$

Simplifying it will give

$$E(S_{ab}) = \bar{\Sigma}_{ab} - \bar{\Delta}_{ab} + S_{ab_{\mu\mu}}$$

□

Theorem (6.2.1) shows that the expectation the covariance between variable a and b has three components, these are the average covariance ($\bar{\Sigma}_{ab}$) and the average cross-covariance ($\bar{\Delta}_{ab}$) and the covariance of the expectation ($S_{ab_{\mu\mu}}$). The term *cross-covariance* indicates the covariance between two variables for two different individuals. When the observations follow the IID assumptions then the $\bar{\Delta}_{ab}$ and $S_{ab_{\mu\mu}}$ will be zero. But in general the $\bar{\Delta}_{ab}$ will not be zero. This theorem is a starting point in developing further theorems concerning the properties of aggregated data.

6.2.2 Group level covariance

Based on theorem (4.1.2), the group level means, $\{\bar{Y}_{a_g}, \bar{Y}_{b_g}\}$, where

$$\bar{Y}_{a_g} = \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} Y_{a_i}; \quad \text{and} \quad \bar{Y}_{b_g} = \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} Y_{b_i}$$

have the moment structure which can be derived as follows,

$$\begin{aligned} (i) \quad E(\bar{Y}_{a_g}) &= \bar{\mu}_{a_g} \\ (ii) \quad \text{Cov}(\bar{Y}_{a_g}, \bar{Y}_{b_g}) &= \frac{1}{N_g} (\bar{\Sigma}_{ab_g} + (N_g - 1)\bar{\Delta}_{ab_g}) \\ (iii) \quad \text{Cov}(\bar{Y}_{a_g}, \bar{Y}_{b_h}) &= \bar{\Delta}_{ab_{gh}}; \quad g \neq h \end{aligned} \quad (6.11)$$

where

$$\begin{aligned} \bar{\mu}_{a_g} &= \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} \mu_{a_i} & \bar{\Sigma}_{ab_g} &= \frac{1}{N_g} \sum_{i \in \mathcal{U}_g} \Sigma_{ab_i} \\ \bar{\Delta}_{ab_g} &= \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \Delta_{ab_{ij}} & \bar{\Delta}_{ab_{gh}} &= \frac{1}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \Delta_{ab_{ij}} \end{aligned}$$

In the same way as (4.28) and (4.29), we can define $\bar{\Delta}_{ab_B}$ and $\bar{\Delta}_{ab_W}$, the average of the between group covariances, and the average of the within group covariances respectively. They are formulated,

$$\bar{\Delta}_{ab_B} = \frac{\sum_{g \neq h} N_g N_h \bar{\Delta}_{ab_{gh}}}{\sum_{g \neq h} N_g N_h}; \quad \text{and} \quad \bar{\Delta}_{ab_W} = \frac{\sum_g N_g(N_g - 1) \bar{\Delta}_{ab_g}}{\sum_g N_g(N_g - 1)} \quad (6.12)$$

and the unweighted versions can be defined as

$$\tilde{\Delta}_{ab_B} = \frac{1}{M(M-1)} \sum_{g \neq h} \bar{\Delta}_{ab_{gh}}; \quad \text{and} \quad \tilde{\Delta}_{ab_W} = \frac{1}{M} \sum_g \bar{\Delta}_{ab_g} \quad (6.13)$$

The $\tilde{\Sigma}_{ab}$ and $\bar{\Sigma}_{ab}$ are defined respectively by

$$\tilde{\Sigma}_{ab} = \frac{1}{M} \sum_g \bar{\Sigma}_{ab_g}; \quad \text{and} \quad \bar{\Sigma}_{ab} = \frac{1}{N} \sum_g N_g \bar{\Sigma}_{ab_g} \quad (6.14)$$

Theorem 6.2.2. *The average of the individual level cross-covariances, $\tilde{\Delta}_{ab}$, can be decomposed into components $\bar{\Delta}_{ab_W}$ and $\bar{\Delta}_{ab_B}$, that is*

$$\bar{\Delta} = \frac{1}{N-1} \left\{ [\bar{N}(1+C^2) - 1] \bar{\Delta}_{ab_W} + [N - \bar{N}(1+C^2)] \bar{\Delta}_{ab_B} \right\} \quad (6.15)$$

where the C^2 was defined in (4.31).

Proof. See the proof of theorem (4.2.3). □

Corollary 6.2.3. *If the N_g are constant at \bar{N} , then $\bar{\Delta}_{ab_W} = \tilde{\Delta}_{ab_W}$, $\bar{\Delta}_{ab_B} = \tilde{\Delta}_{ab_B}$, and $C^2 = 0$, and (6.15) can be expressed as*

$$\bar{\Delta}_{ab} = \frac{\bar{N} - 1}{N - 1} \tilde{\Delta}_{ab_W} + \frac{\bar{N}(M - 1)}{N - 1} \tilde{\Delta}_{ab_B} \quad (6.16)$$

Relative covariance of N_g and $\bar{\Delta}_{ab_g}$

Define $\bar{S}_{ab_{N\bar{\Delta}}}$ as the covariance between N_g and $\bar{\Delta}_{ab_g}$,

$$\bar{S}_{ab_{N\bar{\Delta}}} = \frac{1}{M - 1} \sum_g (N_g - \bar{N}) \cdot (\bar{\Delta}_{ab_g} - \tilde{\Delta}_{ab_W}) \quad (6.17)$$

Then

$$\sum_g N_g \bar{\Delta}_{ab_g} = (M - 1) \bar{S}_{ab_{N\bar{\Delta}}} + M \bar{N} \tilde{\Delta}_{ab_W} \quad (6.18)$$

Define $\bar{C}_{ab_{N\bar{\Delta}}}$ as the relative covariance between N_g and $\bar{\Delta}_{ab_g}$,

$$\bar{C}_{ab_{N\bar{\Delta}}} = \frac{\bar{S}_{ab_{N\bar{\Delta}}}}{\bar{N} \tilde{\Delta}_{ab_W}} \quad (6.19)$$

so that

$$\bar{S}_{ab_{N\bar{\Delta}}} = \bar{N} \tilde{\Delta}_{ab_W} \bar{C}_{ab_{N\bar{\Delta}}}$$

We can define the relative covariance between N_g and $\bar{\Sigma}_{ab_g}$,

$$\bar{C}_{ab_{N\bar{\Sigma}}} = \frac{\bar{S}_{ab_{N\bar{\Sigma}}}}{\bar{N} \tilde{\Sigma}_{ab}} \quad (6.20)$$

where

$$\bar{S}_{ab_{N\bar{\Sigma}}} = \frac{1}{M - 1} \sum_g (N_g - \bar{N}) \cdot (\bar{\Sigma}_{ab_g} - \tilde{\Sigma}_{ab})$$

and so

$$\sum_g N_g \bar{\Sigma}_{ab_g} = (M - 1) \bar{S}_{ab_{N\bar{\Sigma}}} + M \bar{N} \tilde{\Sigma}_{ab} \quad (6.21)$$

By (6.14) and rearranging equation (6.21) gives

$$(\tilde{\Sigma}_{ab} - \bar{\Sigma}_{ab}) = -\frac{M-1}{N} \bar{S}_{ab_{N\bar{\Sigma}}} \quad (6.22)$$

Substituting (6.20) gives

$$(\tilde{\Sigma}_{ab} - \bar{\Sigma}_{ab}) = -\left(\frac{1}{\bar{N}} - \frac{1}{N}\right) \bar{N} \tilde{\Sigma}_{ab} \bar{C}_{ab_{N\bar{\Sigma}}} = -\left(1 - \frac{1}{M}\right) \tilde{\Sigma}_{ab} \cdot \bar{C}_{ab_{N\bar{\Sigma}}} \quad (6.23)$$

Expectation of the unweighted and weighted group level covariance

Consider $\bar{\mathbf{S}}$ the variance-covariance matrix of the group level data, which is calculated using the group level data $\bar{\mathbf{Y}}_a = (\bar{Y}_{a_1}, \dots, \bar{Y}_{a_M})^\top$ and $\bar{\mathbf{Y}}_b = (\bar{Y}_{b_1}, \dots, \bar{Y}_{b_M})^\top$.

The elements of $\bar{\mathbf{S}}$ are \bar{S}_{aa} and \bar{S}_{ab} , the group level variance of the variable a and the group level covariance of the variables a and b , respectively. The covariance \bar{S}_{ab} can be expressed as a quadratic form,

$$\bar{S}_{ab} = \bar{\mathbf{Y}}_a^\top \cdot \mathbf{A} \cdot \bar{\mathbf{Y}}_b \quad (6.24)$$

where \mathbf{A} is a symmetric matrix. The elements of \mathbf{A} are those defined in (4.10) or (4.12).

The unweighted group level covariance, ${}_1\bar{S}_{ab}$, can be defined by substituting (4.10) as elements of \mathbf{A} , giving

$${}_1\bar{S}_{ab} = \frac{1}{M-1} \sum_g (\bar{Y}_{a_g} - {}_1\bar{Y}_a) \cdot (\bar{Y}_{b_g} - {}_1\bar{Y}_b) \quad (6.25)$$

where ${}_1\bar{Y}_a = \frac{1}{M} \sum_g \bar{Y}_{a_g}$ and ${}_1\bar{Y}_b = \frac{1}{M} \sum_g \bar{Y}_{b_g}$.

The weighted group level covariance, ${}_N\bar{S}_{ab}$, is determined by using (4.12) as the elements of \mathbf{A} , that is

$${}_N\bar{S}_{ab} = \frac{1}{M-1} \sum_g N_g (\bar{Y}_{a_g} - \bar{Y}_a) \cdot (\bar{Y}_{b_g} - \bar{Y}_b) \quad (6.26)$$

where

$$\bar{Y}_a = \frac{1}{N} \sum_g N_g \bar{Y}_{a_g} = \frac{1}{N} \sum_{i \in \mathcal{U}} Y_{a_i}; \quad \text{and} \quad \bar{Y}_b = \frac{1}{N} \sum_g N_g \bar{Y}_{b_g} = \frac{1}{N} \sum_{i \in \mathcal{U}} Y_{b_i}$$

Theorem 6.2.4. *Expectation of the group level covariances is*

$$E(\bar{S}_{ab}) = \sum_g a_{gg} \text{Cov}(\bar{Y}_{a_g}; \bar{Y}_{b_g}) + \sum_{g \neq h} a_{gh} \text{Cov}(\bar{Y}_{a_g}; \bar{Y}_{b_h}) + \bar{S}_{ab_{\mu\mu}} \quad (6.27)$$

where

$$\bar{S}_{ab\mu\mu} = \sum_g a_{gg} \bar{\mu}_{a_g} \bar{\mu}_{b_g} + \sum_{g \neq h} a_{gh} \bar{\mu}_{a_g} \bar{\mu}_{b_h}$$

Proof. Applying Lemma (4.1.3) for the covariances. \square

Theorem 6.2.5. *Expectation of the unweighted covariances between variable a and b is*

$$E({}_1\bar{S}_{ab}) = \frac{1}{M} \left(\sum_g \frac{1}{N_g} [\bar{\Sigma}_{ab_g} + (N_g - 1)\bar{\Delta}_{ab_g}] - \sum_{g \neq h} \frac{\bar{\Delta}_{ab_{gh}}}{M-1} \right) + {}_1\bar{S}_{ab\mu\mu} \quad (6.28)$$

Proof. Using a_{gg} and a_{gh} as defined in (4.10) and by (6.11), then rearranging equation (6.27) completes the proof \square

Theorem 6.2.6. *Expectation of the weighted covariances between variables a and b is*

$$E({}_N\bar{S}_{ab}) = \frac{1}{M-1} \left(\sum_g \left[1 - \frac{N_g}{N} \right] (\bar{\Sigma}_{ab_g} + (N_g - 1)\bar{\Delta}_{ab_g}) - \sum_{g \neq h} \frac{N_g N_h}{N} \bar{\Delta}_{ab_{gh}} \right) + {}_N\bar{S}_{ab\mu\mu} \quad (6.29)$$

Proof. The proof is done by substituting equation (4.12) and (6.11) into equation (6.27). \square

From now on, we assume that the means are constant over the population, hence the component of ${}_N\bar{S}_{ab\mu\mu}$ in (6.29) is zero. It implies to equation (6.29) which can be simplified into

$$E({}_N\bar{S}_{ab}) = \frac{1}{M-1} \left(M\bar{\Sigma}_{ab} - \bar{\Sigma}_{ab} \right) - \left(\frac{N-1}{M-1} \right) \bar{\Delta}_{ab} + \frac{1}{M-1} \left(\sum_g N_g \bar{\Delta}_{ab_g} - M\bar{\Delta}_{ab_W} \right) \quad (6.30)$$

Corollary 6.2.7. *The expectation of the weighted covariance can be modified to*

$$E({}_N\bar{S}_{ab}) = \bar{\Sigma}_{ab} \left(1 - \frac{\bar{C}_{ab_{N\bar{\Sigma}}}}{M} \right) - \left(\frac{N-1}{M-1} \right) \bar{\Delta}_{ab} + \bar{\Delta}_{ab_W} \bar{N} \left(\frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} + \bar{C}_{ab_{N\bar{\Delta}}} \right) \quad (6.31)$$

Proof. Equation (6.30) gives

$$E({}_N\bar{S}_{ab}) = \bar{\Sigma}_{ab} + \frac{1}{M-1} (\bar{\Sigma}_{ab} - \bar{\Sigma}_{ab}) - \left(\frac{N-1}{M-1} \right) \bar{\Delta}_{ab} + \frac{M}{M-1} \bar{\Delta}_{ab_W} (\bar{N}-1) + \bar{S}_{ab_{N\bar{\Delta}}}$$

Substituting (6.23) and (6.19)

$$\begin{aligned} E({}_N\bar{S}_{ab}) &= \bar{\Sigma}_{ab} + \frac{1}{M-1} \left[- \left(1 - \frac{1}{M} \right) \bar{\Sigma}_{ab} \bar{C}_{ab_{N\bar{\Sigma}}} \right] - \left(\frac{N-1}{M-1} \right) \bar{\Delta}_{ab} \\ &\quad + \frac{M}{M-1} \bar{\Delta}_{ab_W} (\bar{N}-1) + \bar{N} \bar{\Delta}_{ab_W} \bar{C}_{ab_{N\bar{\Delta}}} \\ &= \bar{\Sigma}_{ab} \left(1 - \frac{1}{M} \bar{C}_{ab_{N\bar{\Sigma}}} \right) - \left(\frac{N-1}{M-1} \right) \bar{\Delta}_{ab} + \left(\frac{M}{M-1} (\bar{N}-1) + \bar{N} \bar{C}_{ab_{N\bar{\Delta}}} \right) \bar{\Delta}_{ab_W} \\ &= \bar{\Sigma}_{ab} \left(1 - \frac{\bar{C}_{ab_{N\bar{\Sigma}}}}{M} \right) - \left(\frac{N-1}{M-1} \right) \bar{\Delta}_{ab} + \bar{\Delta}_{ab_W} \bar{N} \left(\frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} + \bar{C}_{ab_{N\bar{\Delta}}} \right) \end{aligned}$$

□

This corollary corresponds to theorem (4.2.6) for the variances. This corollary (6.2.7) shows the role of $\bar{C}_{ab_{N\bar{\Sigma}}}$ and $\bar{C}_{ab_{N\bar{\Delta}}}$ in the expectation of the weighted group level covariance. Some cases can be noted such as,

Corollary 6.2.8. *If M is large, so that $\frac{1}{M} \approx 0$ and $\frac{N-M}{M-1} \approx \bar{N} - 1$ then*

$$E(N\bar{S}_{ab}) \approx \bar{\Sigma}_{ab} - \bar{\Delta}_{ab} - (\bar{N} - 1)(\bar{\Delta}_{ab} - \tilde{\Delta}_{ab_W}) + \tilde{\Delta}_{ab_W} \bar{N} \bar{C}_{ab_{N\bar{\Delta}}} \quad (6.32)$$

Corollary 6.2.9. *If N_g is constant, then $\tilde{\Sigma}_{ab} = \bar{\Sigma}_{ab}$, and (6.31) becomes*

$$E(N\bar{S}_{ab}) \approx \bar{\Sigma}_{ab} - \bar{\Delta}_{ab} + \frac{N-M}{M-1}(\bar{\Delta}_{ab} - \tilde{\Delta}_{ab_W}) \quad (6.33)$$

Proof. If N_g is constant then $\bar{C}_{ab_{N\bar{\Sigma}}} = 0$ and $\bar{C}_{ab_{N\bar{\Delta}}} = 0$ □

This corollary shows that the weighted group level variance in case of constant $N_g = \bar{N}$ depends on three factors, such as average individual level covariance ($\bar{\Sigma}_{ab}$), average individual level cross covariance ($\bar{\Delta}_{ab}$), and within group average individual level cross covariance ($\tilde{\Delta}_{ab_W}$). The parameters $\bar{\Sigma}_{ab}$ and $\bar{\Delta}_{ab}$ are unaffected by the grouping, but the parameter $\tilde{\Delta}_{ab_W}$ is affected by the groups used.

6.2.3 Relationship between individual and group level covariance

The relationship between covariances calculated using individual level data and those calculated from group level data can be developed in the same way as the theorem (4.2.1).

Lemma 6.2.10. *The total cross product of variable a and b at the individual level data can be partitioned into*

$$\sum_{i \in \mathcal{U}} (Y_{a_i} - \bar{Y}_a) \cdot (Y_{b_i} - \bar{Y}_b) = \sum_g N_g (\bar{Y}_{a_g} - \bar{Y}_a) \cdot (\bar{Y}_{b_g} - \bar{Y}_b) + \sum_g \sum_{i \in \mathcal{U}_g} (Y_{a_i} - \bar{Y}_a) \cdot (Y_{b_i} - \bar{Y}_b) \quad (6.34)$$

and it can be formulated in term of covariance as

$$(N-1)S_{ab} = (M-1)\bar{N}\bar{S}_{ab} + (N-M)S_{ab}^{<W>} \quad (6.35)$$

where

$$S_{ab}^{<W>} = \frac{1}{N-M} \sum_g (N_g - 1)S_{ab}^{<g>}; \quad \text{and} \quad S_{ab}^{<g>} = \frac{1}{N_g - 1} \sum_{i \in \mathcal{U}_g} (Y_{a_i} - \bar{Y}_a) \cdot (Y_{b_i} - \bar{Y}_b) \quad (6.36)$$

Corollary 6.2.11.

$$E(S_{ab}^{<g>}) = \bar{\Sigma}_{ab_g} - \bar{\Delta}_{ab_g} \quad (6.37)$$

Theorem 6.2.12. *Expectation of the $S_{ab}^{<W>}$ is given by*

$$E(S_{ab}^{<W>}) = \tilde{\Sigma}_{ab} \left(1 + \frac{N - \bar{N}}{N - M} \bar{C}_{ab_{N\bar{\Sigma}}} \right) - \tilde{\Delta}_{ab_W} \left(1 + \frac{N - \bar{N}}{N - M} \bar{C}_{ab_{N\bar{\Delta}}} \right) \quad (6.38)$$

Proof. Expectation of the $S_{ab}^{<W>}$ can be obtained by substituting (6.37) into expectation of (6.36), that is

$$\begin{aligned} (N - M)E(S_{ab}^{<W>}) &= \sum_g (N_g - 1)(\bar{\Sigma}_{ab_g} - \bar{\Delta}_{ab_g}) \\ &= \sum_g N_g \bar{\Sigma}_{ab_g} - \sum_g N_g \bar{\Delta}_{ab_g} - \sum_g \bar{\Sigma}_{ab_g} + \sum_g \bar{\Delta}_{ab_g} \end{aligned}$$

Substituting (6.18) and (6.21) gives

$$(N - M)E(S_{ab}^{<W>}) = (M - 1)(\bar{\Sigma}_{ab_{N\bar{\Sigma}}} - \bar{\Sigma}_{ab_{N\bar{\Delta}}}) + M(\bar{N} - 1)(\tilde{\Sigma}_{ab} - \tilde{\Delta}_{ab_W})$$

and substituting (6.19) and (6.20) gives

$$E(S_{ab}^{<W>}) = \tilde{\Sigma}_{ab} \left(1 + \frac{N - \bar{N}}{N - M} \bar{C}_{ab_{N\bar{\Sigma}}} \right) - \tilde{\Delta}_{ab_W} \left(1 + \frac{N - \bar{N}}{N - M} \bar{C}_{ab_{N\bar{\Delta}}} \right)$$

□

6.2.4 Aggregation effect

Assume that the both variables have constant means, i.e. $\mu_{a_i} = \mu_a$, so that ${}_N\bar{S}_{ab_{\mu\mu}}$ and $S_{ab_{\mu\mu}}$ are zero, then the aggregation effect is defined as follows.

Theorem 6.2.13. *Expectation of the difference between group level and individual level covariance of variable a and b is*

$$\begin{aligned} E({}_N\bar{S}_{ab} - S_{ab}) &= -\tilde{\Sigma}_{ab} \cdot \bar{C}_{ab_{N\bar{\Sigma}}} - \frac{M(\bar{N} - 1)}{M - 1} \bar{\Delta}_{ab} + \tilde{\Delta}_{ab_W} \bar{N} \left(\frac{M(\bar{N} - 1)}{(M - 1)\bar{N}} + \bar{C}_{ab_{N\bar{\Delta}}} \right) \\ &= -\tilde{\Sigma}_{ab} \cdot \bar{C}_{ab_{N\bar{\Sigma}}} - \frac{M(\bar{N} - 1)}{M - 1} (\bar{\Delta}_{ab} - \tilde{\Delta}_{ab_W}) + \tilde{\Delta}_{ab_W} \bar{N} \bar{C}_{ab_{N\bar{\Delta}}} \end{aligned} \quad (6.39)$$

Proof. Use (6.10), (6.23) and (6.31).

□

There are several factors affecting the aggregation effect, as $\bar{C}_{ab_{N\bar{\Sigma}}}$, $\bar{C}_{ab_{N\bar{\Delta}}}$, $\bar{\Delta}_{ab}$, $\bar{\Sigma}_{ab}$, and $\bar{\Delta}_{ab_W}$. The first two factors, $\bar{C}_{ab_{N\bar{\Sigma}}}$ and $\bar{C}_{ab_{N\bar{\Delta}}}$, are affected by the N_g and $\bar{\Sigma}_{ab_g}$ or $\bar{\Delta}_{ab_g}$. The $\bar{\Delta}_{ab}$ and $\bar{\Sigma}_{ab}$ are unaffected by groups, but the factor, $\bar{\Delta}_{ab_W}$, is affected by groups. The key factor is likely to be $(\bar{\Delta}_{ab} - \bar{\Delta}_{ab_W})$.

Corollary 6.2.14. *If N_g is constant, then we have aggregation effect*

$$\begin{aligned} E(N\bar{S}_{ab} - S_{ab}) &= -\frac{M(\bar{N} - 1)}{M - 1}(\bar{\Delta}_{ab} - \bar{\Delta}_{ab_W}) \\ &\approx -(\bar{N} - 1)(\bar{\Delta}_{ab} - \bar{\Delta}_{ab_W}) \quad \text{for } M \text{ large} \end{aligned} \quad (6.40)$$

6.3 Cross-semivariogram

The relationship between variables can be defined not only between variables measured on the same individual but also can be extended for pairs of individuals separated by a distance d . In measuring the relationship between variables we need to calculate two components, those are the covariances and variances of the variables. In terms of the spatial perspective the two components correspond to the covariogram and semivariogram in univariate case, and cross covariogram and cross-semivariogram in the multivariate case.

6.3.1 Assumptions and definitions

Recall Y_{a_i} and Y_{b_i} as the realization of the random process $i = 1, \dots, N$. The \mathbf{L} is the matrix of the coordinates of the geographical points within the region \mathcal{D} of \mathcal{R}^2 . The intrinsic stationarity assumptions described in (5.1.1) are generalized as

$$\begin{aligned} (i) \quad &E(Y_{a_i} - Y_{a_j}) = 0 \\ (ii) \quad &\gamma_{ab}(ij) = \gamma(\ell_i - \ell_j) \end{aligned} \quad (6.41)$$

where $\ell_i - \ell_j$ is called the increment. The expression $2\gamma_{ab}(\cdot)$ is called cross variogram and $\gamma_{ab}(\cdot)$ is called cross-semivariogram. Stronger assumptions are those of second order stationarity (5.4), that is

$$\begin{aligned} (i) \quad &E(\mathbf{Y}_a) = \mu_a \\ (ii) \quad &\text{Cov}(Y_{a_i}; Y_{b_j}) = C_{ab}(\ell_i - \ell_j), \text{ for } i \neq j \\ (iii) \quad &\text{Cov}(Y_{a_i}; Y_{b_j}) = \sigma_{ab}, \text{ for } i = j \end{aligned} \quad (6.42)$$

where the $C_{ab}(\cdot)$ represents a cross covariance between variables a and b at locations ℓ_i and ℓ_j , respectively. This value can be interpreted as a covariance between variables a and b at two geographical location.

The expression σ_{ab} is analog with the Σ_{ab_i} of (6.4), which is covariance between variable a and b for the same individual. Two different individuals at one location results in the increment $\ell_i - \ell_j = 0$, and the covariance between Y_{a_i} and Y_{b_j} is then $C_{ab}(0)$. In general $C_{ab}(0) \neq \sigma_{ab}$.

Wackernagel (1998) mentioned that the relationship of cross covariance function with the spatial correlation function in the intrinsic correlation model, that is

$$C_{ab}(\ell_i - \ell_j) = \sigma_{ab} \cdot \rho(\ell_i - \ell_j) \quad (6.43)$$

This relationship shows that if we consider the situation of one individual at one location, it implies $\rho(0) = 1$, then in this special case $C_{ab}(0) = \sigma_{ab}$.

Wackernagel (1998) noted that in general

$$C_{ab}(\ell_i - \ell_j) \neq C_{ba}(\ell_i - \ell_j) \quad \text{and} \quad C_{ab}(\ell_j - \ell_i) \neq C_{ab}(\ell_i - \ell_j) \quad (6.44)$$

but it is true that

$$C_{ab}(\ell_i - \ell_j) = C_{ba}(\ell_j - \ell_i) \quad (6.45)$$

If $\ell_i - \ell_j = d_{ij}$ then $d_{ji} = -d_{ij}$, therefore $C_{ab}(d_{ij}) = C_{ba}(d_{ji})$. Wackernagel (1998,p146) defined this condition as the delay effect (page 146). This condition shows that a changing of order of the observations will change the value of the cross- covariance. Wackernagel (1998, p148) gave an example from a physical experiment comparing two time series of the fluctuation of a gas input into a furnace and the output rate of CO_2 from the furnace. The experimental cross covariance function reveals a delay of 45 seconds between fluctuation in the gas input and a subsequent effect on the rate of carbon dioxide measured at the output of the system.

We will assume that the spatial order of observations does not affect the analysis of social data. Individuals interact with their social environments, in terms of their behaviours, their actions, their opinions, etc, without concerns with any directions. Hence we focus on the absolute value of the increment $\ell_i - \ell_j$, which leads to the isotropic condition, in which the relevant difference between two locations of individuals is measured as an absolute distance. Some empirical works on univariate cases were also relevant to support this claim, such as Griffith et al. (1994), Carrat and Valleron (1992), . Therefore the cross-semivariogram

can be defined as a function of distance d_{ij} , that is

$$\gamma_{ab}(ij) = \gamma_{ab}(d_{ij}) \quad (6.46)$$

Also related to equation (6.45), we will assume that

$$C_{ab}(\ell_i - \ell_j) = C_{ab}(\|\ell_i - \ell_j\|) = C_{ab}(d_{ij}) \quad (6.47)$$

Under these conditions the cross-semivariogram is the half of a covariance of the difference between variables a and b at two particular geographical locations. That is

$$\gamma_{ab}(ij) = \frac{1}{2} \text{Cov}((Y_{a_i} - Y_{a_j}), (Y_{b_i} - Y_{b_j})) \quad (6.48)$$

Theorem 6.3.1. *The intrinsic stationarity assumption (6.41) implies*

$$\gamma_{ab}(ij) = \frac{1}{2} E((Y_{a_i} - Y_{a_j}) \cdot (Y_{b_i} - Y_{b_j})) \quad (6.49)$$

Proof. By definition

$$\text{Cov}((Y_{a_i} - Y_{a_j}), (Y_{b_i} - Y_{b_j})) = E((Y_{a_i} - Y_{a_j}) \cdot (Y_{b_i} - Y_{b_j})) - E(Y_{a_i} - Y_{a_j}) \cdot E(Y_{b_i} - Y_{b_j})$$

The proof is done by applying (6.41). □

Corollary 6.3.2. *The relationship between cross-semivariogram and cross-covariogram is*

$$\gamma_{ab}(ij) = \sigma_{ab} - C_{ab}(d_{ij}) \quad (6.50)$$

Corollary 6.3.3. *The average of the cross-semivariogram is*

$$\bar{\gamma}_{ab} = \sigma_{ab} - \bar{C}_{ab} \quad (6.51)$$

where

$$\bar{C}_{ab} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} C_{ab}(d_{ij}) \quad (6.52)$$

6.3.2 Relationship between cross-semivariogram and spatial correlation

Figure (6.1) illustrated two types of correlation between variables a and b , these being a non spatial correlation (type **a**) and spatial correlation (type **b**). Define spatial correlation $\rho_{ab}(d_{ij})$

$$\rho_{ab}(d_{ij}) = \frac{C_{ab}(d_{ij})}{[\sigma_a^2 \cdot \sigma_b^2]^{\frac{1}{2}}} \quad (6.53)$$

Theorem 6.3.4. *Relationship between cross-semivariogram and spatial correlation is formulated by*

$$\gamma_{ab}(d_{ij}) = \left[\sigma_a^2 \cdot \sigma_b^2 \right]^{\frac{1}{2}} (\rho_{ab} - \rho_{ab}(d_{ij})) \quad (6.54)$$

Proof. Equation (6.50) can be modified

$$\begin{aligned} \gamma_{ab}(d_{ij}) &= \left[\sigma_a^2 \cdot \sigma_b^2 \right]^{\frac{1}{2}} \left(\frac{\sigma_{ab}}{\left[\sigma_a^2 \cdot \sigma_b^2 \right]^{\frac{1}{2}}} - \frac{C_{ab}(d_{ij})}{\left[\sigma_a^2 \cdot \sigma_b^2 \right]^{\frac{1}{2}}} \right) \\ &= \left[\sigma_a^2 \cdot \sigma_b^2 \right]^{\frac{1}{2}} \cdot (\rho_{ab} - \rho_{ab}(d_{ij})) \end{aligned}$$

□

The ρ_{ab} is an ordinary correlation between variable a and b (type **a** in figure 6.1).

Wackernagel (1998) defined the codispersion coefficients, as the ratio of cross-semivariogram and square root of semivariogram of each variable.

$$\rho_{ab}^{\gamma}(d_{ij}) = \frac{\gamma_{ab}(d_{ij})}{\left[\gamma_a(d_{ij}) \cdot \gamma_b(d_{ij}) \right]^{\frac{1}{2}}} \quad (6.55)$$

He noted that the codispersion coefficients can be used to identify the existence of spatial correlation between variables. If this coefficient is constant then the correlation of the variables does not depend on the spatial scale.

6.3.3 The empirical cross-semivariogram

In practice the cross-semivariogram is modeled using the same set of techniques as used for the semivariogram modeling (see section 5.1.3). The processes are initiated by computing the empirical cross-semivariogram, developing categorized cross semivariograms using a particular distance classification, and then developing the cross-semivariogram model as a function of distance.

Theorem 6.3.5. *For an intrinsically stationary process an unbiased estimator of the cross-semivariogram*

$\gamma_{ab}(ij)$ is

$$\hat{\gamma}_{ab}(ij) = \frac{1}{2} ((Y_{a_i} - Y_{a_j}) \cdot (Y_{b_i} - Y_{b_j})) \quad (6.56)$$

Proof. By theorem (6.3.1)

$$\begin{aligned} E(\hat{\gamma}_{ab}(ij)) &= \frac{1}{2} E((Y_{a_i} - Y_{a_j}) \cdot (Y_{b_i} - Y_{b_j})) \\ &= \gamma_{ab}(ij) \end{aligned}$$

□

This cross-semivariogram estimator will be the basic unit of the empirical cross-semivariogram. The categorized version of the empirical cross-semivariogram is obtained by computing the average of the basic unit empirical cross-semivariogram within a particular distance class. The distance classification is produced in the same manner as the distance classification in variogram analysis. The classical categorized estimator of the cross semivariogram can then be obtained as in (5.16).

Models for the cross-semivariogram can be formulated in the same way as semivariogram model (see section 5.1.2). The main parameters of the cross-semivariogram model are nugget (n_{ab}), sill (s_{ab}), and range (r_{ab}). The sill (s_{ab}) may indicate the covariance between variable a and b .

6.3.4 Relationship between cross-semivariogram and semivariogram

The cross-semivariogram of variables a and b , $\gamma_{ab}(d_{ij})$, is bounded by the product of the corresponding semivariogram of each variable (Wackernagel, 1998),

$$\|\gamma_{ab}(d_{ij})\|^2 \leq \gamma_a(d_{ij}) \cdot \gamma_b(d_{ij}) \quad (6.57)$$

Myers (1982) noted an alternative relationship, which is formulated by

$$\gamma_{ab}(d_{ij}) = \frac{1}{2} \cdot (\gamma_{ab}^+(d_{ij}) - \gamma_a(d_{ij}) - \gamma_b(d_{ij})) \quad (6.58)$$

where $\gamma_{ab}^+(d_{ij})$ is the semivariogram of the linear combination $\mathbf{Y}_a + \mathbf{Y}_b$.

The relationship (6.58) is a useful tool in checking the validity of a cross-semivariogram for variables a and b . The inequality (6.57) is important for determining constraints on the parameter values of the cross-semivariogram in terms of the parameter values of the corresponding semivariogram. Consider the exponential semivariogram model with parameters (n_a , s_a , r_a) for variable a and (n_b , s_b , r_b) for variable b . The parameters of the cross-semivariogram with exponential models (n_{ab} , s_{ab} , r_{ab}) are bounded by the value of (n_a , s_a , r_a) and (n_b , s_b , r_b).

Corollary 6.3.6. *The valid nugget and sill value of the exponential cross-semivariogram model are*

$$\begin{aligned}\|n_{ab}\| &\leq [n_a \cdot n_b]^{\frac{1}{2}} \\ \|s_{ab}\| &\leq [s_a \cdot s_b]^{\frac{1}{2}}\end{aligned}\tag{6.59}$$

Proof. There are a specific conditions for the nugget and sill, those are

- Consider $d_{ij} = 0$, then $\gamma_a(d_{ij}) = n_a$ and $\gamma_b(d_{ij}) = n_b$. Evaluating equation (6.57) gives

$$\|n_{ab}\| \leq [n_a \cdot n_b]^{\frac{1}{2}}$$

- Consider $d_{ij} = \infty$, then $\gamma_a(d_{ij}) = s_a$ and $\gamma_b(d_{ij}) = n_b$. Evaluating equation (6.57) gives

$$\|s_{ab}\| \leq [s_a \cdot s_b]^{\frac{1}{2}}$$

□

Constraints on the r_{ab} cannot be defined easily, but trial and error method can be used. The upper bound limits of the cross-semivariogram can be obtained from (6.57). For example suppose that the parameters of exponential semivariogram models of variables a and b are respectively ($n_a = 9$, $s_a = 16$, $r_a = 20$) and ($n_b = 4$, $s_b = 25$, $r_b = 5$). Consider the equality of (6.57), then the absolute of upper limit of the cross-semivariogram can be obtained, as shown in figure (6.2) with (·) symbols.

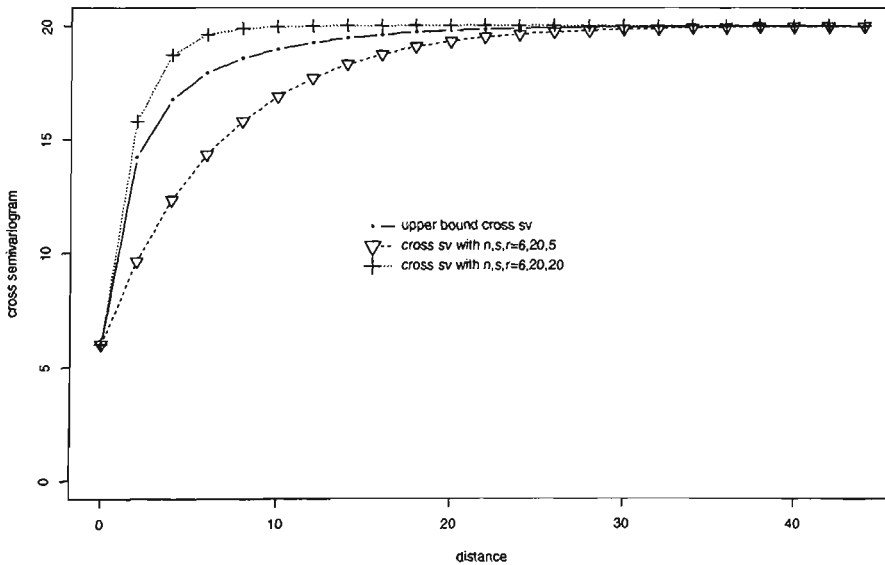


Figure 6.2. Cross-semivariogram and their corresponding semivariogram

According to the corollary (6.3.6) the nugget and sill of the cross-semivariogram are bounded by $-6 \leq n_{ab} \leq 6$ and $-20 \leq s_{ab} \leq 20$. For simplicity consider the $n_{ab} = 6$, and $s_{ab} = 20$. Two different values of the range are drawn in figure (6.2). The first cross-semivariogram model is obtained by setting the range equal to the minimum range of the semivariogram models of variables a and b , that is 5 (∇). Figure (6.2) shows this cross semivariogram that the absolute value of the valid cross semivariogram models lie below the absolute upper bound of the cross-semivariogram model (the dash line with ∇). The second cross-semivariogram model is obtained by setting the range equal to the maximum range of the semivariogram models of variables a and b , that is 20 (+). This cross-semivariogram model with + sign lie above the absolute upper bound of the the cross-semivariogram model, hence this cross-semivariogram is not valid. This figure illustrates that choosing the minimum range from the semivariogram models of variables a and b as the range of the cross-semivariogram of variables a and b gives a valid cross-semivariogram according to the inequality (6.57).

6.3.5 Generating random observations based of a cross-semivariogram model

Suppose we want to generate observations consistent with a given semivariogram model. Consider the random variables \mathbf{Y} which is defined by

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (6.60)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are defined by

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}; \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \quad (6.61)$$

The matrix Σ_{aa} and Σ_{bb} are determined by applying relation (5.1.1) of the semivariogram model of variables a and b respectively (see section 5.1.8). Meanwhile the Σ_{ab} or Σ_{ba} are determined by using relationship in (6.50). Having defined the $\boldsymbol{\Sigma}$ then the simulation procedure described in section (5.1.8) can be applied.

For example, consider the exponential model of the semivariogram and cross-semivariogram. The semivariogram model's parameters for the variables a and b are respectively ($n_a = 4$, $s_a = 25$, $r_a = 10$)

and $(n_b = 9, s_b = 16, r_b = 12)$. The valid cross-semivariogram model's parameters could be $(n_{ab} = 5, s_{ab} = 10, r_{ab} = 10)$. The result of simulation of the cross-semivariogram model from 1500 observations are presented in figure (6.3).

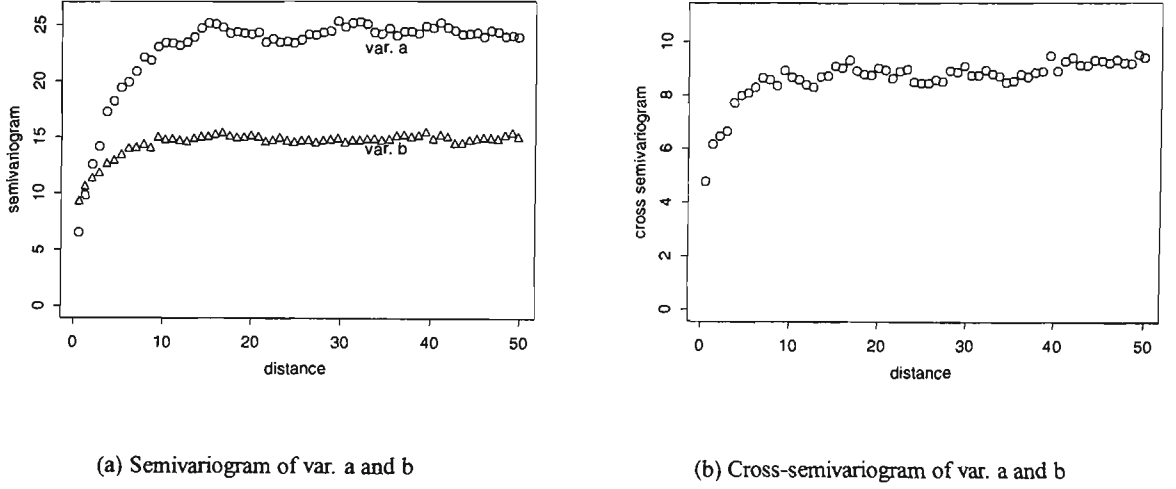


Figure 6.3. Simulation result of the cross-semivariogram with $(n_{ab} = 5, s_{ab} = 8, r_{ab} = 10)$.

6.4 Group level cross-semivariogram

Given the group level data $\{\bar{Y}_{a_1}, \bar{Y}_{a_g}, \dots, \bar{Y}_{a_M}\}, \{\bar{Y}_{b_1}, \bar{Y}_{b_g}, \dots, \bar{Y}_{b_M}\}$ obtained from the individual data for which the intrinsically stationary assumptions apply,

$$E(\bar{Y}_{a_g} - \bar{Y}_{a_h}) = 0 \quad (6.62)$$

Define

$$\Gamma_{ab}(gh) = \frac{1}{2} \text{Cov}((\bar{Y}_{a_g} - \bar{Y}_{a_h}); (\bar{Y}_{b_g} - \bar{Y}_{b_h}) | \mathbf{L}) \quad (6.63)$$

If the second order stationarity assumptions apply to the individual data, then

$$\begin{aligned} (i) \quad & E(\bar{Y}_{a_g}) = \mu_a \\ (ii) \quad & \text{Cov}(\bar{Y}_{a_g}; \bar{Y}_{b_h}) = \frac{1}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} C_{ab}(d_{ij}) \end{aligned} \quad (6.64)$$

Theorem 6.4.1. *The group level cross-semivariogram is, for a stationary process*

$$\Gamma_{ab}(gh) = \frac{1}{2} E \left((\bar{Y}_{a_g} - \bar{Y}_{a_h}) \cdot (\bar{Y}_{b_g} - \bar{Y}_{b_h}) \right) \quad (6.65)$$

Theorem 6.4.2. *An unbiased estimator of the group level cross-semivariogram $\Gamma_{ab}(gh)$ is*

$$\hat{\Gamma}_{ab}(gh) = \frac{1}{2} \left((\bar{Y}_{a_g} - \bar{Y}_{a_h}) \cdot (\bar{Y}_{b_g} - \bar{Y}_{b_h}) \right) \quad (6.66)$$

Theorem 6.4.3. *Relationship between group level cross-semivariogram and individual level cross-semivariogram is*

$$\Gamma_{ab}(gh) = \bar{\gamma}_{ab_{gh}} - \frac{1}{2} \frac{N_g - 1}{N_g} \bar{\gamma}_{ab_g} - \frac{1}{2} \frac{N_h - 1}{N_h} \bar{\gamma}_{ab_h} \quad (6.67)$$

where

$$\begin{aligned} \bar{\gamma}_{ab_{gh}} &= \frac{1}{N_g N_h} \sum_{i \in \mathcal{U}_g} \sum_{j \in \mathcal{U}_h} \gamma_{ab}(d_{ij}); \\ \bar{\gamma}_{ab_g} &= \frac{1}{N_g(N_g - 1)} \sum_{i \neq j \in \mathcal{U}_g} \gamma_{ab}(d_{ij}); \\ \bar{\gamma}_{ab_h} &= \frac{1}{N_h(N_h - 1)} \sum_{i \neq j \in \mathcal{U}_h} \gamma_{ab}(d_{ij}) \end{aligned}$$

Proof. Expanding equation (6.65) of theorem (6.4.1), we have

$$2\Gamma_{ab}(gh) = \text{Cov}(\bar{Y}_{a_g}, \bar{Y}_{b_g}) + \text{Cov}(\bar{Y}_{a_h}, \bar{Y}_{b_h}) - \text{Cov}(\bar{Y}_{a_g}, \bar{Y}_{b_h}) - \text{Cov}(\bar{Y}_{a_h}, \bar{Y}_{b_g})$$

Substituting (6.64 -ii) and (6.50)

$$2\Gamma_{ab}(gh) = 2\bar{\gamma}_{ab_{gh}} - \frac{N_g - 1}{N_g} \bar{\gamma}_{ab_g} - \frac{N_h - 1}{N_h} \bar{\gamma}_{ab_h}$$

□

6.5 Relationship between sample covariances and cross-semivariogram

The relationship is started by defining the following theorem, that is

Theorem 6.5.1. *The expectation of the individual level sample covariances between variables a and b is equal the average of the individual level cross-semivariogram between variables a and b ,*

$$E(S_{ab}) = \bar{\gamma}_{ab} + S_{ab_{\mu\mu}} \quad (6.68)$$

where

$$\bar{\gamma}_{ab} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathcal{U}} \gamma_{ab}(d_{ij})$$

Proof. Use (6.10) and (6.51). □

Equation (6.35) shows that the individual covariance, S_{ab} , is a linear combination of weighted group level covariance and the average cross product within group, $S_{ab}^{<W>}$.

Corollary 6.5.2. *Expectation of the individual cross product of variables a and b within the g th group is*

$$E(S_{ab}^{<g>}) = \bar{\gamma}_{ab_g} + S_{ab_{\mu\mu}}^{<g>} \quad (6.69)$$

Proof. The individual cross product within the group is defined in (6.36), which is analog with the individual population covariance, S_{ab} . Hence applying theorem (6.5.1) completes the proof. □

Corollary 6.5.3. *The group level covariance and cross covariance between variable a and b are*

$$\begin{aligned} (i) \quad \text{Cov}(\bar{Y}_{a_g}; \bar{Y}_{b_g}) &= \bar{\Sigma}_{ab_g} - \frac{N_g - 1}{N_g} \bar{\gamma}_{ab_g} \\ (ii) \quad \text{Cov}(\bar{Y}_{a_g}; \bar{Y}_{b_h}) &= \frac{1}{2} (\bar{\Sigma}_{ab_g} + \bar{\Sigma}_{ab_h}) - \bar{\gamma}_{ab_{gh}} \end{aligned} \quad (6.70)$$

Proof. See theorem (4.4.4). □

Define two components, which are the average between group cross-semivariogram ($\bar{\gamma}_{ab_B}$) and the average within group cross-semivariogram ($\bar{\gamma}_{ab_W}$).

$$\bar{\gamma}_{ab_B} = \frac{\sum_{g \neq h} N_g N_h \bar{\gamma}_{ab_{gh}}}{\sum_{g \neq h} N_g N_h} \quad ; \text{ and } \quad \bar{\gamma}_{ab_W} = \frac{\sum_g N_g (N_g - 1) \bar{\gamma}_{ab_g}}{\sum_g N_g (N_g - 1)} \quad (6.71)$$

and the unweighted versions

$$\tilde{\gamma}_{ab_B} = \frac{1}{M(M-1)} \sum_{g \neq h} \tilde{\gamma}_{ab_{gh}}; \quad \text{and} \quad \tilde{\gamma}_{ab_W} = \frac{1}{M} \sum_g \tilde{\gamma}_{ab_g} \quad (6.72)$$

Theorem 6.5.4. *The $\bar{\gamma}_{ab}$ can be expressed as a function of $\bar{\gamma}_{ab_B}$ and $\bar{\gamma}_{ab_W}$.*

$$\bar{\gamma}_{ab} = \frac{1}{N-1} \left\{ [\bar{N}(1+C^2) - 1] \bar{\gamma}_{ab_W} + [N - \bar{N}(1+C^2)] \bar{\gamma}_{ab_B} \right\} \quad (6.73)$$

where C^2 is defined in (4.31).

Proof. See the proof of theorem (4.4.5). □

Corollary 6.5.5. *If N_g is constant, then $C^2 = 0$ and (6.73) becomes*

$$\tilde{\gamma}_{ab} = \frac{\tilde{N} - 1}{N - 1} \tilde{\gamma}_{ab_W} + \frac{\tilde{N}(M - 1)}{N - 1} \tilde{\gamma}_{ab_B} \quad (6.74)$$

6.5.1 Relative covariance of N_g and $\tilde{\gamma}_{ab_g}$

Define $\bar{S}_{ab_{N\tilde{\gamma}}}$ as the covariance between N_g and $\tilde{\gamma}_{ab_g}$,

$$\bar{S}_{ab_{N\tilde{\gamma}}} = \frac{1}{M - 1} \sum_g (N_g - \tilde{N}) \cdot (\tilde{\gamma}_{ab_g} - \tilde{\gamma}_{ab_W}) \quad (6.75)$$

Rearranging this equation, gives

$$\sum_g N_g \tilde{\gamma}_{ab_g} = (M - 1) \bar{S}_{ab_{N\tilde{\gamma}}} + M \tilde{N} \tilde{\gamma}_{ab_W} \quad (6.76)$$

The relative covariance of N_g and $\tilde{\gamma}_{ab_g}$ can be defined by

$$\bar{C}_{ab_{N\tilde{\gamma}}} = \frac{\bar{S}_{ab_{N\tilde{\gamma}}}}{\tilde{N} \tilde{\gamma}_{ab_W}} \quad (6.77)$$

Hence

$$\bar{S}_{ab_{N\tilde{\gamma}}} = \tilde{N} \cdot \tilde{\gamma}_{ab_W} \cdot \bar{C}_{ab_{N\tilde{\gamma}}} \quad (6.78)$$

The quantity of $\bar{C}_{ab_{N\tilde{\gamma}}}$ is useful when we try to evaluate the effect of group size in the group level cross-semivariogram.

6.5.2 Expectation of the $S_{ab}^{<W>}$, \bar{S}_{ab} , and $N\bar{S}_{ab}$

Corollary 6.5.6. *Expectation of the average individual cross product within the g th group is defined by*

$$E(S_{ab}^{<W>}) = \frac{M - 1}{M(\tilde{N} - 1)} \left(\bar{S}_{ab_{N\tilde{\gamma}}} + \frac{M - 1}{M(\tilde{N} - 1)} \tilde{\gamma}_{ab_W} \right) + S_{ab_{\mu\mu}}^{<W>} \quad (6.79)$$

where

$$S_{ab_{\mu\mu}}^{<W>} = \frac{1}{N - M} \sum_g (N_g - 1) S_{ab_{\mu\mu}}^{<g>}$$

Proof. Expectation of (6.36) can be defined as

$$E(S_{ab}^{<W>}) = \frac{1}{N - M} \sum_g (N_g - 1) E(S_{ab}^{<g>})$$

Substituting (6.69)

$$\begin{aligned} E(S_{ab}^{<W>}) &= \frac{1}{N-M} \sum_g (N_g - 1) (\tilde{\gamma}_{abg} + S_{ab\mu\mu}^{<g>}) \\ &= \frac{1}{N-M} \sum_g N_g \tilde{\gamma}_{abg} - \frac{1}{N-M} \sum_g \tilde{\gamma}_{abg} + S_{ab\mu\mu}^{<W>} \end{aligned}$$

Substituting (6.76)

$$\begin{aligned} E(S_{ab}^{<W>}) &= \frac{1}{N-M} ((M-1)\bar{S}_{abN\bar{\gamma}} + M\bar{N}\tilde{\gamma}_{abW}) - \frac{1}{N-M} \cdot M\tilde{\gamma}_{abW} + S_{ab\mu\mu}^{<W>} \\ &= \frac{M-1}{M(\bar{N}-1)} \bar{S}_{abN\bar{\gamma}} + \tilde{\gamma}_{abW} \left(\frac{M\bar{N}}{M(\bar{N}-1)} - \frac{1}{\bar{N}-1} \right) + S_{ab\mu\mu}^{<W>} \\ &= \frac{M-1}{M(\bar{N}-1)} \bar{S}_{abN\bar{\gamma}} + \tilde{\gamma}_{abW} + S_{ab\mu\mu}^{<W>} \\ &= \frac{M-1}{M(\bar{N}-1)} \left(\bar{S}_{abN\bar{\gamma}} + \frac{M(\bar{N}-1)}{M-1} \tilde{\gamma}_{abW} \right) + S_{ab\mu\mu}^{<W>} \end{aligned}$$

□

Theorem 6.5.7. *Expectation of the unweighted group level covariance is*

$$E({}_1\bar{S}_{ab}) = \tilde{\gamma}_{abB} - \tilde{\gamma}_{abW} + \frac{1}{M} \sum_g \frac{\bar{\gamma}_{abg}}{N_g} + {}_1\bar{S}_{ab\mu\mu} \quad (6.80)$$

Proof. See proof of theorem (4.4.7).

□

Theorem 6.5.8. *Expectation of the weighted group level covariance is*

$$\begin{aligned} E({}_N\bar{S}_{ab}) &= \frac{N-1}{M-1} \tilde{\gamma}_{ab} - \bar{N}\tilde{\gamma}_{abW} \left(\bar{C}_{abN\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \right) \\ &\quad + {}_N\bar{S}_{ab\mu\mu} - \frac{N-1}{M-1} S_{ab\mu\mu} \end{aligned} \quad (6.81)$$

Proof. See proof of theorem (4.4.8).

□

Corollary 6.5.9. *Assume the constant mean and*

$$\frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}} \approx 1; \quad \text{and} \quad \frac{N-1}{M-1} \approx \bar{N}$$

then the expectation of the weighted covariance can be approximated

$$E({}_N\bar{S}_{ab}) \approx \bar{N} (\tilde{\gamma}_{ab} - \tilde{\gamma}_{abW} [\bar{C}_{abN\bar{\gamma}} + 1]) \quad (6.82)$$

This corollary shows that the factors $\tilde{\gamma}_{abW}$ and $\bar{C}_{abN\bar{\gamma}}$ play key roles in determining the expectation of the group level covariance.

6.6 Aggregation effect in terms of cross-semivariogram

The aggregation effect on the covariance between variables a and b can be related to the cross-semivariogram.

Theorem 6.6.1. *Expectation of the aggregation effect of the weighted group level covariance and its individual covariance is*

$$E(N\bar{S}_{ab} - S_{ab}) = \frac{M(\bar{N} - 1)}{M - 1} (\bar{\gamma}_{ab} - \bar{\gamma}_{abw}) - \bar{N}\bar{\gamma}_{abw}\bar{C}_{abN\bar{\gamma}} + N\bar{S}_{ab\mu\mu} - S_{ab\mu\mu} \quad (6.83)$$

Proof. The proof is immediately from theorem (6.5.8) and (6.5.1). \square

Theorem 6.6.2. *Expectation of the aggregation effect in term of the ratio of the weighted group level covariance and its individual covariance is, for a process with constant mean*

$$\frac{E(N\bar{S}_{ab})}{E(S_{ab})} = \frac{N - 1}{M - 1} - \frac{\bar{\gamma}_{abw}}{\bar{\gamma}_{ab}} \bar{N} \left(\frac{M}{M - 1} \cdot \frac{\bar{N} - 1}{\bar{N}} + \bar{C}_{abN\bar{\gamma}} \right) \quad (6.84)$$

Proof. The proof is immediately from theorem (6.5.8) and (6.5.1). \square

Corollary 6.6.3. *Consider a case that N_g is constant at \bar{N} , implies $\bar{C}_{abN\bar{\gamma}} = 0$, then*

$$E(N\bar{S}_{ab} - S_{ab}) = \frac{M(\bar{N} - 1)}{M - 1} (\bar{\gamma}_{ab} - \bar{\gamma}_{abw}) \quad (6.85)$$

and

$$\frac{E(N\bar{S}_{ab})}{E(S_{ab})} = \frac{N - 1}{M - 1} - \frac{\bar{\gamma}_{abw}}{\bar{\gamma}_{ab}} \frac{M(\bar{N} - 1)}{M - 1} \quad (6.86)$$

The other consequence is

Corollary 6.6.4. *If data are available to estimate S_{ab} and $N\bar{S}_{ab}$ then an estimator of $\bar{\gamma}_{abw}$ can be developed as*

$$\hat{\gamma}_{abw} = \frac{\left(S_{ab} \cdot \frac{N-1}{M-1} - N\bar{S}_{ab} \right)}{\left(\frac{M(\bar{N}-1)}{M-1} - \bar{N}\bar{C}_{abN\bar{\gamma}} \right)} \quad (6.87)$$

Proof. Exploring (6.83) can be represented into

$$-\hat{\gamma}_{abw} \left(\frac{M(\bar{N} - 1)}{M - 1} - \bar{N}\bar{C}_{abN\bar{\gamma}} \right) = N\bar{S}_{ab} - S_{ab} - \frac{N - M}{M - 1} S_{ab}$$

This can be simplified into

$$-\hat{\gamma}_{abw} \left(\frac{M(\bar{N} - 1)}{M - 1} - \bar{N}\bar{C}_{abN\bar{\gamma}} \right) = N\bar{S}_{ab} - S_{ab} \frac{N - 1}{M - 1}$$

or

$$\tilde{\gamma}_{abw} = \frac{\left(S_{ab} \cdot \frac{N-1}{M-1} - N\bar{S}_{ab}\right)}{\left(\frac{M(\tilde{N}-1)}{M-1} - \tilde{N}\bar{C}_{abN\tilde{\gamma}}\right)}$$

□

6.7 The weighting factors of the group level cross-semivariogram

This use of weighting factors to adjust the group level cross-semivariogram so that it is closer to the individual level cross-semivariogram, as we discussed in section (5.5), are considered here. The previous theoretical results of the cross-semivariogram show the same basic results as for the semivariogram. Therefore, the weighting factors for the cross-semivariogram will be the same as we discussed in the semivariogram. They are defined in table (6.1).

Table 6.1. The weighting factor

| Description | formulation | weighting factor |
|---------------------------------|--|---|
| 1. $\hat{N}\Gamma_{ab}(gh)$ | $\frac{1}{2} \left(\left[\sqrt{N_g}(\bar{Y}_{a_g} - \bar{Y}_a) - \sqrt{N_h}(\bar{Y}_{a_h} - \bar{Y}_a) \right] \cdot \left[\sqrt{N_g}(\bar{Y}_{b_g} - \bar{Y}_b) - \sqrt{N_h}(\bar{Y}_{b_h} - \bar{Y}_b) \right] \right)$ | – |
| 2. $\hat{\Gamma}_{ab}^{w2}(gh)$ | $\left(\frac{1}{N_g} + \frac{1}{N_h} \right)^{-1} \cdot \hat{\Gamma}_{ab}(gh)$ | $\left(\frac{1}{N_g} + \frac{1}{N_h} \right)^{-1}$ |

We illustrate the use of these weighting factors based on the simulation of a population with a exponential cross-semivariogram model. The parameters of the exponential model are defined as nugget (n_{ab})=10, sill (s_{ab})=15, and range (r_{ab})=25. The population size is 1500, and the simulations were repeated 100 times. The simulations were done with limited repetition since the time constraint of generating bivariate population. This parameter estimates obtained using individual data are shown in figure (6.4) and described in table (6.2). The figure and table show that the individual population was generated with the expected parameter values.

The parameters are estimated using the categorized version of the empirical cross-semivariogram by the weighted least squares methods (Cressie, 1985). The weight was defined to be number of pairs involved in a category over the square of the model value.

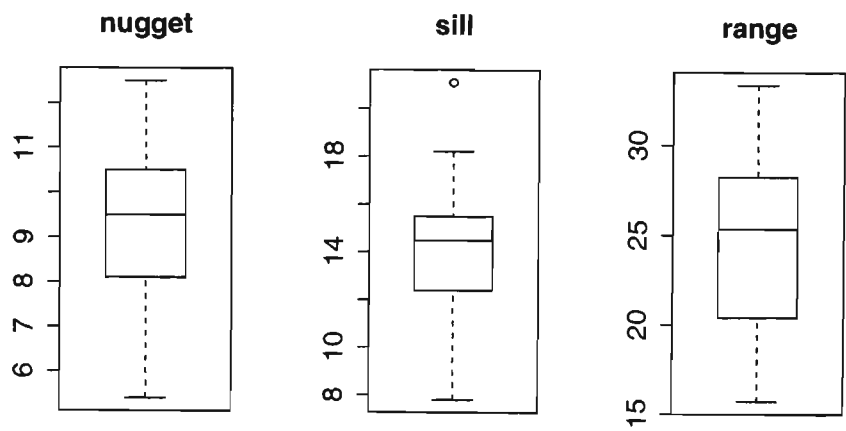


Figure 6.4. The distribution of the estimated individual population parameter from the 100 simulations

Table 6.2. The description of the estimated parameter of the individual population

| Parameter | Mean | Median | Min. | Max. | std. err. |
|-----------|-------|--------|------|------|-----------|
| nugget | 9.34 | 9.5 | 5.4 | 12.5 | 1.76 |
| sill | 13.83 | 14.5 | 7.8 | 21.1 | 2.54 |
| range | 24.90 | 25.35 | 15.7 | 33.4 | 4.33 |

True value $n_{ab} = 10, s_{ab} = 15, r_{ab} = 25$

The simulation process was initiated by generating the individual population under a particular parameter nugget, sill, and range. Then the grouping process was applied to create group level data. The grouping was done by creating small uniform rectangles across the region. The region is bounded by the points (20,30) at the lower left corner and (90,100) at the upper right corner. The region is divided into 150 small uniform rectangles by dimension 10 x 15. The group level data were obtained by calculating the mean of each group. After that the cross-semivariogram is calculated based on the individual level and group level data. The two different weighted cross-semivariograms are also calculated as presented in table (6.1).

Figure (6.5), (6.6), and (6.7) give an illustration of the estimated parameters from the group level cross-semivariogram, which can be compared with the estimated individual level cross-semivariogram parameters. Figure (6.5) shows the estimated group level nugget, both unweighted and weighted. The far left boxplot shows the unweighted group level nugget. The unweighted nugget is far below the individual level nugget as the discussion of the effect of aggregation on the estimated nugget in chapter 5 suggested.

But the nugget estimated from the weighted group level analysis shows some adjustment which increases the estimated group level nugget to beyond or around the individual level nugget. The first weighting factor produced too much adjustment to the group level nugget, hence its estimated values were larger than the individual level. Figure (6.5) shows that the second weighting factor gives better adjustment. Table (6.3) confirms this situation, though it seems that the averages are lower than the expected parameter.

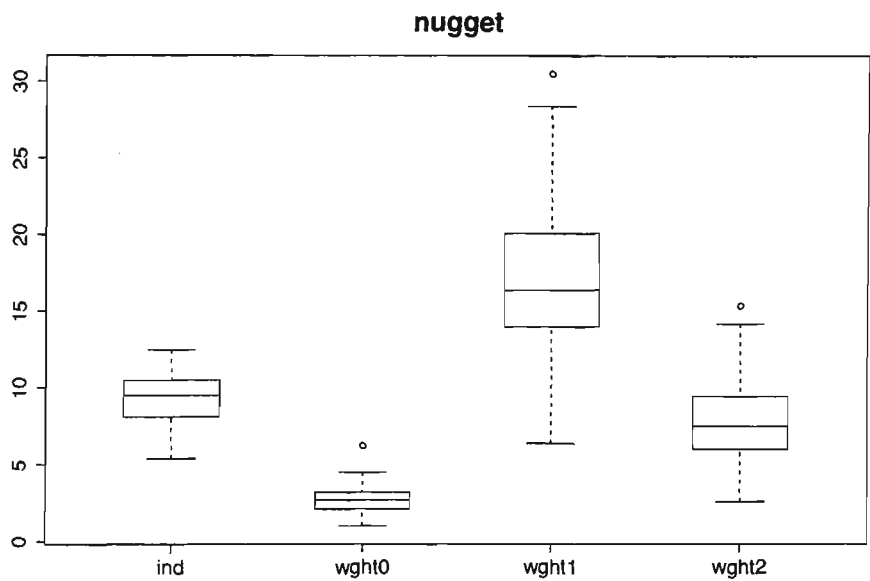


Figure 6.5. Distribution of the estimated nugget from the individual level, group level, and the two different weighted group level cross-semivariogram

Figure (6.6) shows the estimated group level sill compared with the individual level sill. The un-weighted group level sill is below the individual level sill as the discussion in chapter 5 suggested. The first weighting factor adjusts group level sill to be higher than the individual level sill. The second weighting factor gives a better adjustments to the group level sill. Their estimated parameters are around the individual level sill.

Figure (6.7) illustrates the comparison of the estimated group level range and the individual level range. The boxplots show that the estimated ranges are all close to each others, either the group level or the individual level. Tables (6.3) and (6.2) show the value of the estimated values fell within the value 22.0 to 27.0. The weighting factors also did not give any significance adjustment to the estimated group level range.

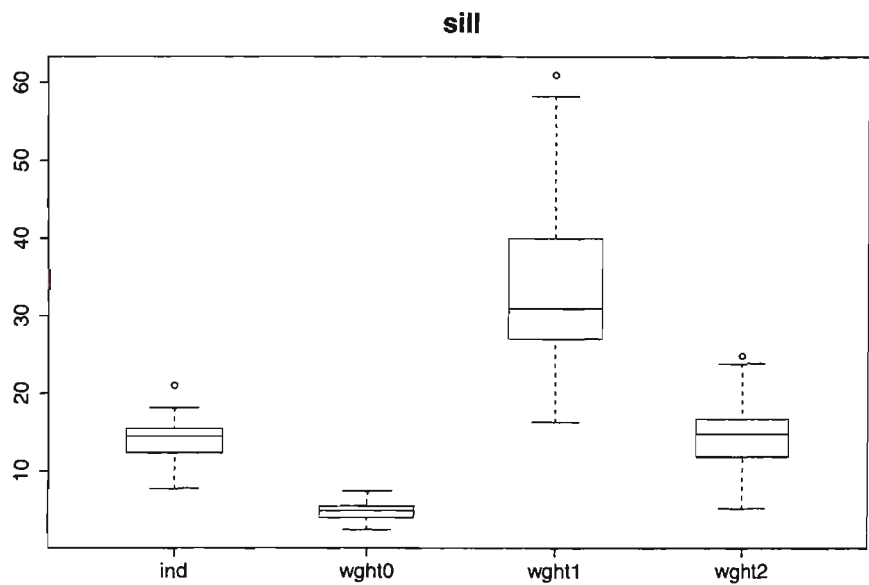


Figure 6.6. Distribution of the estimated sill from the individual level, group level, and the two different weighted group level cross-semivariogram

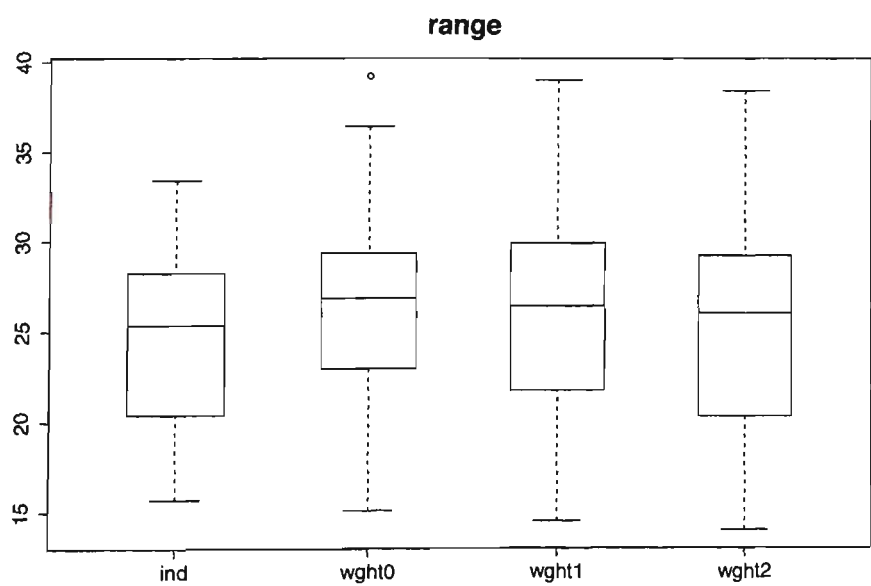


Figure 6.7. Distribution of the estimated range from the individual level, group level, and the two different weighted group level cross-semivariogram

Table 6.3. The description of the estimated parameter of the group level cross-semivariogram, unweighted and weighted

| Parameter | Mean | Median | Min. | Max. | std. err. |
|------------------|-------|--------|------|------|-----------|
| Individual level | | | | | |
| nugget | 9.34 | 9.5 | 5.4 | 12.5 | 1.33 |
| sill | 13.83 | 14.5 | 7.8 | 21.1 | 1.59 |
| range | 24.90 | 25.35 | 15.7 | 33.4 | 2.08 |
| Unweighted | | | | | |
| nugget | 2.74 | 2.7 | 1.0 | 6.2 | 0.93 |
| sill | 4.92 | 5.0 | 2.5 | 7.5 | 1.03 |
| range | 26.16 | 26.9 | 15.1 | 39.2 | 2.23 |
| weighted #1 | | | | | |
| nugget | 16.92 | 16.4 | 6.4 | 30.5 | 2.18 |
| sill | 33.32 | 31.0 | 16.4 | 61.0 | 3.09 |
| range | 25.66 | 26.4 | 14.5 | 39.0 | 2.35 |
| weighted #2 | | | | | |
| nugget | 7.83 | 7.5 | 1.6 | 15.4 | 1.62 |
| sill | 14.69 | 14.9 | 5.3 | 25.0 | 2.00 |
| range | 25.12 | 26.0 | 14.0 | 38.4 | 2.32 |

Using these weighting factors may provide a way of determining the initial value of the estimation of the individual level cross semivariogram parameter from the group level one. The second weighting factor gives better adjustment to the estimated group level cross-semivariogram parameters. Hence their values may be used as the initial values of the estimation procedure of the individual level cross-semivariogram parameters which will be discussed in the next section (6.8).

6.8 Estimation of individual level cross-semivariogram parameters from the group level cross-semivariogram

This section is an extension of the methods discussed in section (5.4) for the individual semivariogram estimation to the bivariate case. Consider the situation when there are available the group level data of (\bar{Y}_a, \bar{Y}_b) , where $\bar{Y}_a^T = [\bar{Y}_{a_1}, \dots, \bar{Y}_{a_M}]$, $\bar{Y}_b^T = [\bar{Y}_{b_0}, \dots, \bar{Y}_{b_M}]$ and the groups of centroid's locations, $[\ell_1, \dots, \ell_M]$.

Define the empirical unit group level cross-semivariogram as

$$\hat{\Gamma}_{ab}(gh) = \frac{1}{2} ((\bar{Y}_{a_g} - \bar{Y}_{a_h}) \cdot (\bar{Y}_{b_g} - \bar{Y}_{b_h}))$$

(6.88)

In practice, estimation of the parameters of the cross semivariogram model is usually done using the categorized version of the empirical group level cross-semivariogram. The categorization is done by taking

the average of all pairs which belong to the defined category,

$$\hat{\Gamma}_{ab}(k) = \frac{1}{|M_k|} \sum_{d_{gh} \in D_k}^{|M_k|} \hat{\Gamma}_{ab}(gh), \quad k = 1, \dots, K$$

where $|M_k|$ is number of pair in the class distance D_k , and $\hat{\Gamma}_{ab}(d_{gh})$ is the empirical unit cross-semivariogram.

See section (5.1.3) for more detail of categorization process.

The data $(\hat{\Gamma}_{ab}(k), \bar{d}_k)$ of $k = 1, \dots, K$ be used to fit a particular cross-semivariogram model, where

$$\bar{d}_k = \frac{1}{|M_k|} \sum_{d_{gh} \in D_k}^{|M_k|} d_{gh}, \quad k = 1, \dots, K \quad (6.89)$$

Consider the exponential model then

$$\hat{\Gamma}_{ab}(gh) = \hat{n}_{ab} + (\hat{s}_{ab} - \hat{n}_{ab}) \left(1 - \exp \left[\frac{-3d_{gh}}{\hat{r}_{ab}} \right] \right) \quad (6.90)$$

where \hat{n}_{ab} , \hat{s}_{ab} , and \hat{r}_{ab} are the group level estimators of the nugget, sill, and range, respectively. The weighted least square method may be used to estimate these parameters, with the weight as defined in (5.18).

The basic model for development of this approach is the relationship between the group level cross-semivariogram and the individual level cross-semivariogram, such as defined in theorem (6.4.3). Applying the second order assumptions and the intrinsic stationary assumptions, the approximation can be developed by using the Taylor series expansion, that is

$$\Gamma_{ab}(gh) \approx \gamma_{ab}(d_{gh}) - \frac{1}{2} \frac{N_g - 1}{N_g} \gamma_{ab}(\bar{d}_g) - \frac{1}{2} \frac{N_h - 1}{N_h} \gamma_{ab}(\bar{d}_h) \quad (6.91)$$

Equation (6.91) is similar to equation (5.93) concerning the approximation group level semivariogram with the individual level semivariogram.

Assume that the individual level cross-semivariogram is the exponential model, with parameter n_{ab} , s_{ab} , and r_{ab} . Hence the extension of equation (5.92) into the cross-semivariogram can be rewritten as

$$\begin{aligned} \Gamma_{ab}(gh) \approx & s_{ab} - (s_{ab} - n_{ab}) \cdot \exp \left[\frac{-3d_{gh}}{r_{ab}} \right] - s_{ab} \left\{ 1 - \frac{1}{2N_g} - \frac{1}{2N_h} \right\} \\ & + \frac{N_g - 1}{2N_g} (s_{ab} - n_{ab}) \cdot \exp \left[\frac{-3\bar{d}_g}{r_{ab}} \right] + \frac{N_h - 1}{2N_h} (s_{ab} - n_{ab}) \cdot \exp \left[\frac{-3\bar{d}_h}{r_{ab}} \right] \end{aligned} \quad (6.92)$$

In the same way as in section (5.4), we may extend the methods of solving equation (6.92) given the value $\widehat{\Gamma}_{ab}(gh)$, N_g , N_h , d_{gh} , \bar{d}_g , \bar{d}_h . The non-linear least square method can be developed to estimate the three parameters η_{ab} , s_{ab} , and r_{ab} . Estimating these three parameters provides an inference concerning the spatial structure in the individual level population. A major issue in applying the non-linear least square method found in section (5.4) is the need to determine good initial value of those three parameters. Incorrect determination of the initial value may result in the estimation process not converging or may result in a very large value of estimator. In setting the initial value, we may consider three different situations, those are

- the situation when the \widehat{S}_{aa} , \widehat{S}_{bb} , and \widehat{S}_{ab} are available from some source, such as a random sample.
- the situation when \widehat{S}_{aa} , and \widehat{S}_{bb} only are available.
- the situation when no individual level information is available.

The first and second situations apply when the individual level statistics are available from some sources. For example in the case of census data, these can be available from micro-sample data, which is a random sample of the census data. But the random sample usually excludes the locations of the individual, because of confidentiality reasons. In these two situations, the initial values of the parameters can be determined by following the methods discussed in section (5.4.1). In the first situation the initial value s_{ab} is equal to the \widehat{S}_{ab} . In the second situation then the upper limit of corollary (6.3.6) may be applied, that is $s_{ab} = \sqrt{(\widehat{S}_{aa} \cdot \widehat{S}_{bb})}$. The third situation applies when only the group level data are available. It is assumed that the group level data also contain the location of the group, i.e. the centroid of the group.

In situations when the individual sample was available, we can compute the individual level covariance (\widehat{S}_{ab}), which becomes the initial value of the sill. If we have only information about \widehat{S}_{aa} , and \widehat{S}_{bb} available from published data, then we can define $s_{ab} = \sqrt{(\widehat{S}_{aa} \cdot \widehat{S}_{bb})}$ as initial value of the sill. The initial value of nugget is defined equal to zero (see section 5.4.2). Meanwhile the initial value for the range can be chosen from the estimated weighted group level cross-semivariogram by the second weighting factor as discussed in the previous section.

In the situation when no individual sample was available, then initial values of the nugget, sill, and range are determined from the estimated parameters of the weighted group level cross semivariogram by using the second weighting factors. Therefore no information of the individual sample is needed.

For an illustration the simulated data as discussed in section (6.7) are considered. The estimation procedures are done in SAS, which are similar with the SAS codes in appendix (E) for the semivariogram. They are the extension of the similar procedures which are discuss theoretically in section (5.4). The results of estimation are shown in figure (6.8), (6.8), and (6.10). These distributions can be compared with figure (6.4) relating to the estimates obtained from individual level data. The descriptions of the estimated parameters are tabulated in table (6.4).

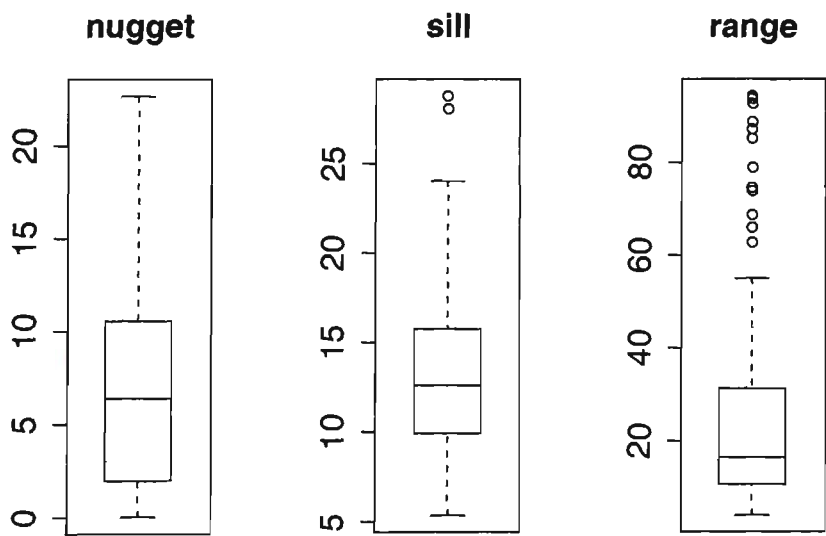


Figure 6.8. Distribution of the estimated individual level parameter nugget, sill, and range, when \hat{S}_{aa} , \hat{S}_{bb} , and \hat{S}_{ab} are available.

The figures and table show that estimation of the individual level parameters from group level data in the these situations give very similar results. Although variation of the estimated parameters are still high compared with the individual level estimates, the average of the individual level estimates (table 6.2) are within the interval of the first and third quartile of the estimated nugget, sill, and range (figure 6.8, 6.9, and 6.10). The standard deviation of the estimated parameters show a similar results for the three situations.

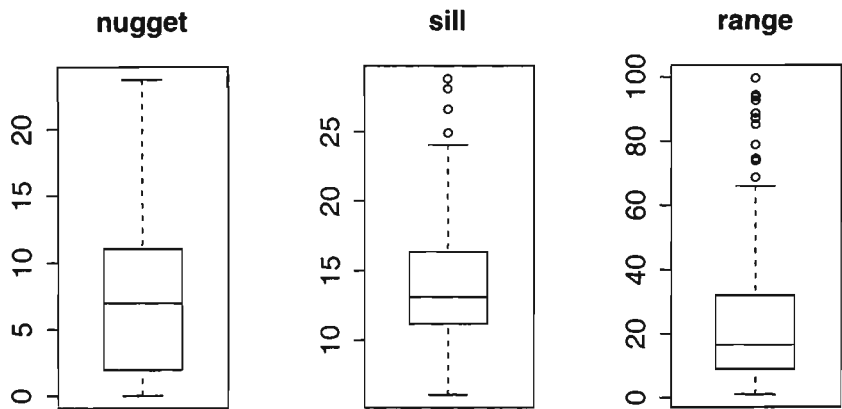


Figure 6.9. Distribution of the estimated individual level parameter nugget, sill, and range, when only \hat{S}_{aa} and \hat{S}_{bb} are available.

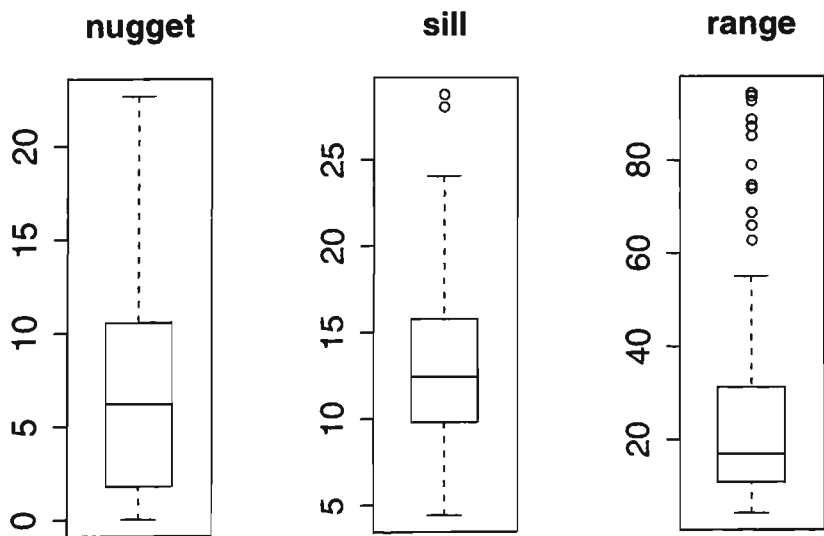


Figure 6.10. Distribution of the estimated individual level parameter nugget, sill, and range, when individual sample was not available

Table 6.4. Estimated individual level parameters of cross semivariogram as shown in figure (6.8), (6.9), and (6.10)

| Parameter | mean | median | min. | max. | std.err. |
|--|-------|--------|------|-------|----------|
| Individual level | | | | | |
| nugget | 9.34 | 9.5 | 5.4 | 12.5 | 1.76 |
| sill | 13.83 | 14.5 | 7.8 | 21.1 | 2.54 |
| range | 24.90 | 25.35 | 15.7 | 33.4 | 4.33 |
| Individual sample available ($\hat{S}_{aa}, \hat{S}_{bb}, \hat{S}_{ab}$) | | | | | |
| nugget | 6.90 | 6.38 | 0.0 | 22.66 | 5.58 |
| sill | 12.96 | 12.60 | 5.35 | 28.79 | 4.28 |
| range | 24.34 | 16.51 | 4.06 | 94.31 | 20.75 |
| Individual sample available ($\hat{S}_{aa}, \hat{S}_{bb}$) | | | | | |
| nugget | 7.33 | 6.99 | 0.0 | 23.71 | 5.97 |
| sill | 14.08 | 13.11 | 6.03 | 28.79 | 4.30 |
| range | 24.75 | 16.51 | 0.91 | 99.58 | 22.44 |
| No individual sample | | | | | |
| nugget | 6.83 | 6.18 | 0.0 | 22.66 | 5.60 |
| sill | 12.89 | 12.44 | 4.40 | 28.79 | 4.35 |
| range | 24.44 | 16.88 | 4.16 | 94.38 | 20.71 |
| True value $n_{ab} = 10, s_{ab} = 15, r_{ab} = 25$ | | | | | |

Compared with table (6.3), the non-linear regression approach using group level data has considerably reduced the bias in the estimation of the nugget and produced sill estimates close to those obtained using individual level data. While the unadjusted group level analysis did not produce seriously biased estimates of the range, the use of the non-linear regression approach did produce range estimates close to the individual level value.

6.9 Summary

The study of aggregation effects on the cross-semivariogram shows that it shares the same derivation with the semivariogram. It is shown that the cross-semivariogram is a bivariate case of the semivariogram. Hence the theorems in semivariogram are also applied for the cross-semivariogram.

The cross-semivariogram graphs gives an effective description of the spatial correlation between variables, as indicated by its parameters nugget, sill, and range. The nugget indicates co-variations between variables due to measurement error. The sill shows the covariance of the variables, and the range indicates the distance in which covariance become constant, so that the observations are independence each other.

Estimation of the individual level cross-semivariogram parameters using the group level data also gives an initial success, although some problems occur in determining initial values of the parameters. Applying the estimated unadjusted group level cross-semivariogram parameters as the initial value give a similar estimate as using the adjusted group level or the individual sample statistics such as \hat{S}_{aa} , \hat{S}_{bb} , or \hat{S}_{ab}

Generating individual observations with a particular cross-semivariogram is constrained by computer time, since the simulation process is extremely time consuming. The number of individuals which can be generated are reduced by a half compared with the univariate case, such as in the semivariogram.

Chapter 7

The MAUP as a tool in Semivariogram Analysis

7.1 Introduction

Social data often come from a census, sample survey or administrative source. The data from a census or administrative system are often available in terms of aggregated data at one or more scales for particular zoning schemes. For example, the 1991 Australian Census of Housing and Population provide data in the forms of tables at the collection district level.

In general the aggregation process is done by creating spatial groups or areal units at some scale for a zoning scheme. This process can be done by partitioning the study area into mutually exclusive groups. The scale refers to the number of groups, and the zoning refers to boundaries of those groups.

Cressie (1996) noted that analysis of aggregated spatial data will be limited to the level of scale used. Furthermore Holt et al. (1996) noted that the results of the analysis also depends on the zoning being used. These two factors potentially affect all spatially aggregated data and are referred to as the modifiable areal unit problem or MAUP. The scaling and zoning aspects affect some statistics such as variance, correlation coefficient (Openshaw & Taylor, 1981), but do not affect the mean. This chapter is intended to demonstrate, theoretically and empirically, that the MAUP can be used as a tool in analysing spatial data, especially for variogram analysis.

7.2 Development of the study on the MAUP

Holt et al. (1996) showed that the MAUP is caused by the failure to incorporate area or spatial effects into the analysis. They argued that the MAUP can be explained by incorporating the area effects into the model

underpinning the analysis. Further discussion is found in Steel and Holt (1996a) and Wrigley, Holt, Steel, and Tranmer (1996).

Cressie (1996) discussed an analogy to the MAUP in analysing mining data, which he referred to as the change of spatial support problem. The ore samples from a few cubic feet in volume are used to predict the average ore grade of mining units in the whole study region. Biased prediction may result from the analysis which did not take account this change of spatial support.

One approach to the MAUP was discussed early by Openshaw (1977) who proposed transforming the problems into the automatic zoning problem. A particular solution of the MAUP in term of ecological fallacy was proposed by Steel and Holt (1996a).

In most studies, it seems that the MAUP was viewed as a problem in analysing aggregated data. It is a challenge to explore the MAUP further, and consider how the MAUP can be used as a tool in analysing spatially aggregated data.

7.3 Evidence of the MAUP

The evidence of the MAUP have been studied by many researchers. It was originally identified by Gehlke and Biehl (1934), when they discussed the difference between results obtained using data for areal units from a different sets of zones. Openshaw and Taylor (1979) showed that the correlation, between the percentage vote for Republican candidates in the congressional election of 1968 and the percentage of population over sixty years old, tended to increase when the zones were grouped together. They showed that this effect was related to spatial correlation. Later this fact was investigated by Arbia (1989a). Openshaw and Taylor (1981) discussed the nature of the modifiable areal unit problem, and mentioned that the problem exists because the zoning systems are arbitrary and modifiable.

7.3.1 Theoretical evidence

Cressie (1996) noted that one of the disturbing features about a change in the level of aggregation is that inferences about the relationship within or between processes may vary as a function of the aggregation. Openshaw (1977) argued that the scale and zoning are a unique characteristics of the aggregated spatial

data and cannot be removed but can be controlled in some ways. Steel and Holt (1996a) introduced an adjustment factor which attempted to eliminate the aggregation effect in the ecological fallacy.

Steel and Holt (1996b) stated that the variance may contain key information relating to the aggregation effect. The aggregation effect can be formulated by the difference between the variance of the aggregated data and the original data, as defined by Amrhein and Reynolds (1996). If the original data are at the individual level data, then the aggregation effect is the difference of the variance of the group level data and the variance of the individual level data.

Chapter (4) has discussed in detail the theoretical aspect of the aggregation effect. The discussion was focused on the expectation of the difference between the weighted group level variance and its individual level variance (see theorem 4.2.7). Assume that the mean term of (4.44) is zero, then this equation can be rewritten as

$$E(N\bar{S}_{yy} - S_{yy}) = \frac{M(\bar{N} - 1)}{M - 1}(\tilde{\Delta}_W - \bar{\Delta}) - \tilde{\Sigma}\bar{C}_{N\bar{\Sigma}} + \tilde{\Delta}_W\bar{N}\bar{C}_{N\bar{\Delta}} \quad (7.1)$$

where $\bar{C}_{N\bar{\Sigma}}$ is the relative covariance of N_g and $\bar{\Sigma}_g$, and $\bar{C}_{N\bar{\Delta}}$ is the relative covariance of N_g and $\bar{\Delta}_g$. The relative covariances, $\bar{C}_{N\bar{\Sigma}}$ and $\bar{C}_{N\bar{\Delta}}$, reflect the relationship between N_g and $\bar{\Sigma}_g$ and $\bar{\Delta}_g$, respectively. These values may indicate the effect of scale as the group size N_g changes.

Equation (7.1) indicates that the expectation of the difference between group and individual level of variance can be decomposed into three components. The first component involves \bar{N} , $\tilde{\Delta}_W$, and $\bar{\Delta}$. The factor \bar{N} is determined by the scale and is usually known, and is the same for all zoning at a given scale. The factor $\bar{\Delta}$ is a parameter of the individuals in the population and is not affected by either scale or zoning. The key factor is $\tilde{\Delta}_W$ which may be affected by both the scale and zoning used. The second component is affected by the $\bar{C}_{N\bar{\Sigma}}$ factor, which may be significant when \bar{N} is not constant. If there was a negative relation between N_g and $\bar{\Sigma}_g$ then this component may give a positive effect, but it will give a negative effect otherwise. The last component contains two main factors, $\tilde{\Delta}_W$ and $\bar{C}_{N\bar{\Delta}}$. These factors can be affected by scale or zoning effect. Later these factors will be looked at in terms of $\tilde{\gamma}_W$ and $\bar{C}_{N\tilde{\gamma}}$, which are expected to be the main factors of aggregation effect.

7.3.2 The empirical evidence

Various empirical studies of the MAUP have been reported. For example Amrhein and Reynolds (1996) discussed the result of exploratory analysis examining the correlation between aggregation effect and a set of spatial statistics. Green and Flowerdew (1996) presented a discussion based on the result of correlation and regression analysis of variables undertaken on the basis of random, systematic, and spatial aggregation. The authors argued that the spatial regression may provide an explanation of the cause of the MAUP, a prediction of how the result of a standard regression analysis should change as the level of aggregation is increased, and an approach that may alleviate the MAUP. A similar discussion may be found in the article by Wong (1996) and Curtis and MacPherson (1996).

Here some empirical evidence will be presented using data from the 1991 Australian Census of Population and Housing. The Adelaide region will be chosen as the study region. Figure (8.1a) shows the region is divided into non-overlapping collection districts. Each collection district is characterized by their centroid (Fig. 8.1b). There are 1713 collection district within this region and 767 030 people over 15 years old were counted by the census.

There are three groupings, which the CD level data can be readily aggregated to. The groupings are SSC, DPC, and LGA (ABS & MapInfo, 1993). The SSC is a term that refers to a collection district derived suburb. It is composed of one or more collection districts that lie wholly within a suburb. If the CD is split across two or more suburb boundaries, then the CD is allocated to the most appropriate suburb. The DPC is the 1991 census derived from Australian Post Postcode boundaries. It may contain one or more collection districts. The LGA is the legal local government area, and it may contain one or more collection districts.

Table 7.1. Number of groups for each grouping factor

| Grouping | CD | SSC | DPC | LGA |
|----------|------|-----|-----|-----|
| M | 1713 | 313 | 102 | 27 |

Five variables will be considered, these are employment rate, unemployment rate, labor participation rate, formal education rate, and informal education rate (see section 8.2). Some statistics are tabulated in

table (7.2), and boxplot is used to give some descriptions of the group mean of the characteristics for the different groupings (Fig. 7.1).

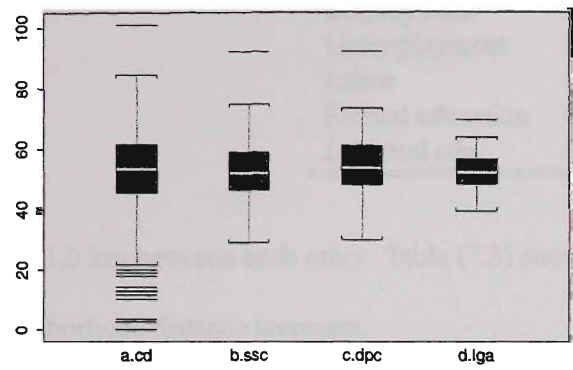
Table 7.2. Some statistics of the variables at different grouping factors

| Characteristics | | a.cd | b.ssc | c.dpc | d.lga |
|-------------------|---------------|----------|----------|----------|----------|
| Employment rate | mean | 0.530147 | 0.525477 | 0.541936 | 0.526310 |
| | unwght. var. | 0.013365 | 0.008666 | 0.008 | 0.004651 |
| | weighted var. | 5.890811 | 21.75529 | 53.14686 | 132.8199 |
| Unemployment rt. | mean | 0.072689 | 0.072653 | 0.073276 | 0.072657 |
| | unwght. var. | 0.001281 | 0.001019 | 0.000705 | 0.000591 |
| | weighted var. | 0.532484 | 1.888799 | 4.715055 | 12.62434 |
| Labor part. rt. | mean | 0.602836 | 0.598131 | 0.615212 | 0.598966 |
| | unwght. var. | 0.011157 | 0.006252 | 0.006053 | 0.003122 |
| | weighted var. | 4.920792 | 16.56473 | 39.93089 | 101.8567 |
| Formal edu. rt. | mean | 0.127972 | 0.140581 | 0.126042 | 0.152750 |
| | unwght. var. | 0.007285 | 0.007932 | 0.005761 | 0.005992 |
| | weighted var. | 3.126571 | 15.63804 | 42.33584 | 168.8433 |
| Informal edu. rt. | mean | 0.133475 | 0.127839 | 0.136465 | 0.125443 |
| | unwght. var. | 0.001504 | 0.001085 | 0.000855 | 0.000581 |
| | weighted var. | 0.660828 | 2.569285 | 5.941580 | 18.63357 |

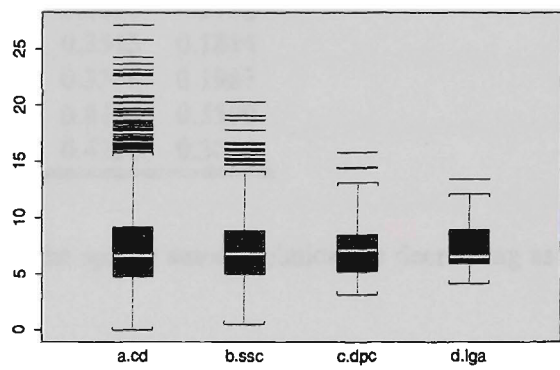
Table (7.2) shows that the mean of the characteristics are not affected much by the grouping. But the unweighted and weighted variance of the variables are affected by the grouping. Table (7.2) shows that the weighted variance, which would equal the individual level variance for IID data, increases with the scale. Figure (7.1) shows a boxplot of each variable for the different groupings and indicates the variation of the observations are getting smaller as the grouping goes from CD level to the LGA level. The effect of grouping on the variance may also affect other analysis which involve the variance, such as correlation or regression analysis.

In a particular level of grouping, the MAUP can be investigated by looking at the presence of the spatial autocorrelation (Arbia, 1989a). Consider the Adelaide data available at the CD level, and consider some neighborhood distance 1.0 km, 2.0 km, and 5.0 km. The connectivity matrix can be created for each neighborhood distance, then the Moran coefficient can be calculated (Ding & Fotheringham, 1992). The minimum distance between CD is 0.11 km.

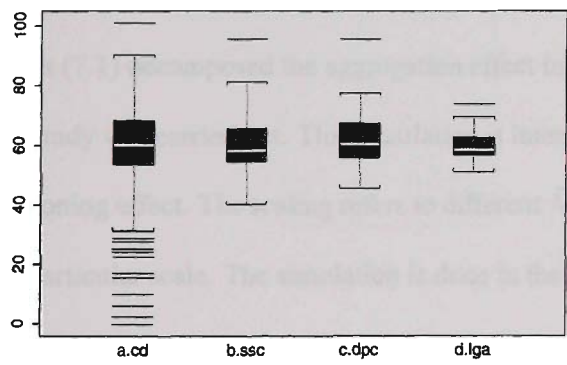
The different neighborhood distances can be looked at as different groupings of the collection district. For example, the neighborhood distance 1.0 km will aggregate the CDs which have distance of at



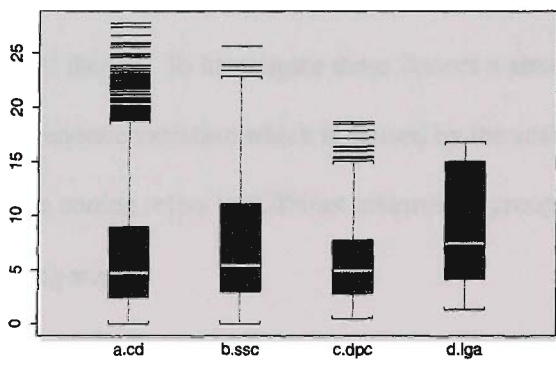
(a) Employment rate (%)



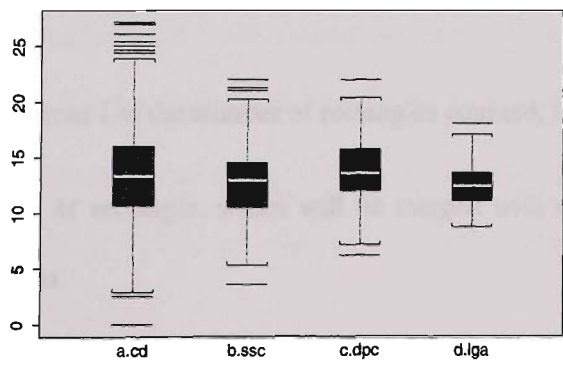
(b) Unemployment rate (%)



(c) Labor participation rate (%)



(d) Formal education rate (%)



(e) Informal education rate (%)

Figure 7.1. Boxplot at different grouping factors, ranging from CD level to LGA level.

Table 7.3. Moran autocorrelation coefficient of Adelaide CD data at different neighborhood distances

| Distance (km) | 1.0 | 2.0 | 5.0 |
|------------------|--------|--------|--------|
| Employment | 0.5241 | 0.3690 | 0.2132 |
| Unemployment | 0.5074 | 0.3513 | 0.1816 |
| Labor | 0.4711 | 0.3374 | 0.1987 |
| Formal education | 0.9406 | 0.8106 | 0.5590 |
| Informal edu. | 0.5703 | 0.4253 | 0.3033 |

least 1.0 km between each other. Table (7.3) shows that the spatial autocorrelation are decreasing as the neighborhood distance increases.

7.4 Empirical illustration of the MAUP

7.4.1 Simulation process

Result (7.1) decomposed the aggregation effect into several factors. To investigate these factors a simulation study was carried out. This simulation is intended to generate variation which is caused by the scaling and zoning effect. The scaling refers to different \bar{N} , and the zoning refers to different schemes of grouping at a particular scale. The simulation is done in the following steps :

step 0 Specify the dimensions of the rectangular region, number of groups required (M), and number of repetitions of the simulation. The dimension of the rectangular region for example can be 10x10, 15x10, etc, denote this as G . The region is partitioned into a regular grid resulting in equal size rectangles.

step 1 Number the rectangles from 1 to the number of rectangles counted, i.e. G .

step 2 Randomly choose $G - M$ rectangle, which will be merged with other rectangles to create the required number of groups.

step 3 Assign each rectangle chosen for merging to one of the two different merging processes, (a) merge to the left or right of the adjacent grid, and (b) merge to the top or bottom of the adjacent grid. Assignment of the merging process was done randomly between the alternatives.

step 4 List all the members of the defined groups, and calculate the groups level data.

- step 5 Compute some group level statistics.
- step 6 Repeat step 1 to 5 according to the number of repetitions.

For example, suppose the region is divided into a 7x8 grid producing 56 equal size rectangles and 44 groups are to be created. There are 12 ($56 - 44 = 12$) rectangles chosen randomly and assigned the merging process. Therefore in the region 44 groups ($56 - 12$) are created. One realization is drawn in figure (7.2).

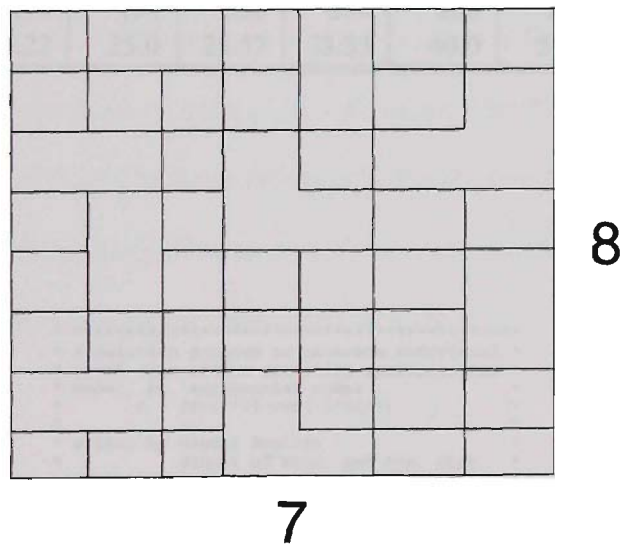


Figure 7.2. The grouping process of the region with 7×8 grids.

The individual level data are generated according to the exponential model semivariogram with the following parameters, nugget=5, sill=20, and range=15. There are 10000 individual points in the population within the region of 60 by 80 length unit and 4800 square area units. The population was generated by using the Choleski decomposition methods (section 5.1.8). Some statistics of the population and the individual level empirical semivariogram plot are shown in figure (7.3). Twenty different number of groups (scales) and sixty repetitions (zoning) were applied for each scale in the simulation. The twenty different number of groups are shown in table (7.4).

Table 7.4. Twenty different number of groups (scales)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| grid | 50x50 | 50x50 | 40x50 | 40x40 | 40x40 | 40x35 | 40x35 | 30x35 | 30x30 | 30x30 |
| M | 2000 | 1800 | 1600 | 1400 | 1200 | 1000 | 900 | 800 | 700 | 600 |
| N̄ | 5.0 | 5.56 | 6.25 | 7.14 | 8.33 | 10.0 | 11.11 | 12.50 | 14.29 | 16.67 |

| | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| grid | 25x30 | 25x25 | 25x25 | 25x20 | 20x20 | 20x20 | 20x20 | 20x15 | 20x15 | 20x15 |
| M | 500 | 450 | 400 | 350 | 300 | 250 | 200 | 150 | 100 | 50 |
| N̄ | 20.0 | 22.22 | 25.0 | 28.57 | 33.33 | 40.0 | 50.0 | 66.67 | 100.0 | 200.0 |

```
*****
* Simulation program to generate individual *
* level data with a specified semivariogram *
* model, eg. exponential model :          *
*   n + (s-n)*(1-exp(-3*d/r))             *
*                                           *
* written by Gandhi Pawitan                *
*   School of Math. and App. Stat.         *
*****
Semivariogram model : n + s Exp(r)
n,s,r      :    5.000  20.000  15.000
Number of points : 10000
Region shape : rectangle
X-min,X-max   :   20.000  80.000
Y-min,Y-max   :   10.000  90.000
-----
Performance of the generated points :
           Min      Max      Mean      Var.
easting    :  20.000  79.992  49.855  298.042
northing   :  10.012  89.990  49.867  539.949
original z  :           0.004    1.003
generated z : -17.113  16.818    0.758  18.397
```

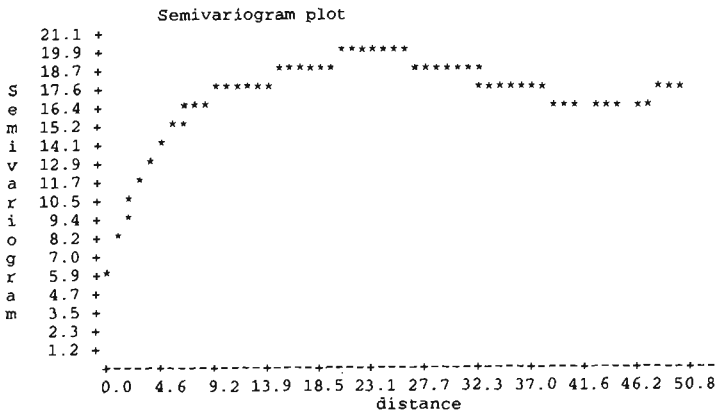


Figure 7.3. The semivariogram plot of the individual population

7.4.2 The theoretical background

Theorem (4.3.8) shows the relationship between the weighted group level variance and the individual level empirical semivariogram. Furthermore equation (4.71) can be decomposed into three important parts,

$$\begin{aligned} {}_N\bar{S}_{yy}^{(1)} &= \frac{N-1}{M-1} \bar{\gamma} \\ {}_N\bar{S}_{yy}^{(2)} &= \bar{N} \tilde{\gamma}_w \bar{C}_{N\tilde{\gamma}} \\ {}_N\bar{S}_{yy}^{(3)} &= \tilde{\gamma}_w \frac{M}{M-1} (\bar{N} - 1) \end{aligned} \quad (7.2)$$

In equation (4.71) the $\bar{\gamma}$ is not affected by either the scale nor zoning effect. But the factors \bar{N} , $\tilde{\gamma}_w$, and $\bar{C}_{N\tilde{\gamma}}$ are affected by scale or zoning effect. The scale effect will affect all three factors, but the zoning will affect only the $\tilde{\gamma}_w$, and $\bar{C}_{N\tilde{\gamma}}$. The factor \bar{N} is usually known, however the other factors may not be.

Define two types of aggregation effect as \hat{D} and \hat{R} , which are the difference and the ratio of the weighted group level variance and the individual level variance,

$$\hat{D} = {}_N\bar{S}_{yy} - S_{yy}; \quad \text{and} \quad \hat{R} = \frac{{}_N\bar{S}_{yy}}{S_{yy}} \quad (7.3)$$

Results for these aggregation effects were given by (4.72) and (4.73), respectively. These show that the variation in the aggregation effects, either in terms of the difference or ratio, can be related to variation in the three factors, \bar{N} , $\tilde{\gamma}_w$, and $\bar{C}_{N\tilde{\gamma}}$. The relationship of \bar{N} with the \hat{D} or \hat{R} might give the effect of the scale. Since \bar{N} is usually known, we might look at the relationship of \bar{N} with $\tilde{\gamma}_w$ and $\bar{C}_{N\tilde{\gamma}}$. We speculate that $\tilde{\gamma}_w$ is a more important factor in explaining aggregation effects than $\bar{C}_{N\tilde{\gamma}}$. This will be investigated through a graphical approach.

In the same way as (7.2), we can separate the ((4.72) into three different parts, that is

$$\begin{aligned} \hat{D}^{(1)} &= \frac{N-M}{M-1} \bar{\gamma} \\ \hat{D}^{(2)} &= \bar{N} \tilde{\gamma}_w \bar{C}_{N\tilde{\gamma}} \\ \hat{D}^{(3)} &= \tilde{\gamma}_w \frac{M}{M-1} (\bar{N} - 1) \end{aligned} \quad (7.4)$$

so $\hat{D} = \hat{D}^{(1)} - \hat{D}^{(2)} - \hat{D}^{(3)}$. And also for the (4.73),

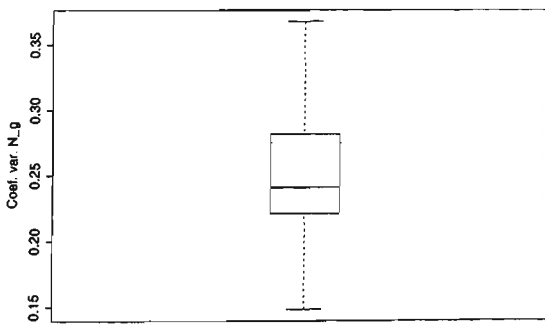
$$\begin{aligned}\hat{R}^{(1)} &= \frac{N-1}{M-1} \\ \hat{R}^{(2)} &= \bar{N} \frac{\tilde{\gamma}_w}{\tilde{\gamma}} \bar{C}_{N\tilde{\gamma}} \\ \hat{R}^{(3)} &= \frac{\tilde{\gamma}_w}{\tilde{\gamma}} \frac{M}{M-1} (\bar{N}-1)\end{aligned}\quad (7.5)$$

so $\hat{R} = \hat{R}^{(1)} - \hat{R}^{(2)} - \hat{R}^{(3)}$.

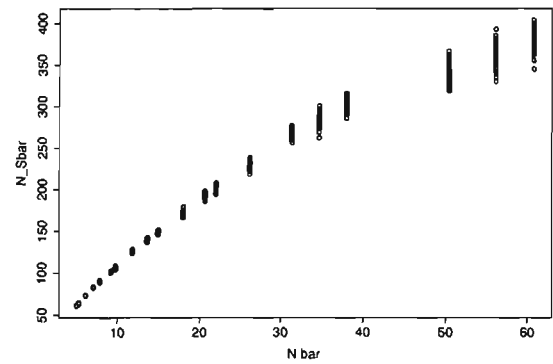
7.4.3 Some illustration from the simulation

The factor \bar{N} plays an important role in exploring the MAUP, since the \bar{N} is the easiest factor to be measured. We will also show that the other factors can be explored mainly based on their relationship with the \bar{N} . Then in the next section (section 7.5) the factor $\tilde{\gamma}_w$ will be explored using semivariogram analysis. We are also going to show empirically that the factor $\bar{C}_{N\tilde{\gamma}}$ can be ignored.

The factor $\bar{C}_{N\tilde{\gamma}}$ will be influenced by the variation in the group population sizes, N_g . The coefficient variation of the group size N_g in the simulation varies from 0.15 to the 0.38. This can be compared with the coefficient of variation found in the CDs in the Illawarra and Adelaide data sets of 0.11. The distribution of this coefficient of variation of the simulated N_g is shown in figure (7.4-a).



(a) C^2 of N_g



(b) $N\bar{S}_{yy}$

Figure 7.4. Coefficient of variation of the N_g and the relationship of \bar{N} (scale) with the $N\bar{S}_{yy}$

The figure (7.4-b) shows the effect of scale and zoning on the calculation of the weighted group level variance, $N\bar{S}_{yy}$. The weighted group level variance increases non-linearly with the increasing scale. Equa-

tion (4.71) shows that $N\bar{S}_{yy} = N\bar{S}_{yy}^{(1)} - N\bar{S}_{yy}^{(2)} - N\bar{S}_{yy}^{(3)}$, where

$$\begin{aligned} N\bar{S}_{yy}^{(1)} &= \frac{N-1}{M-1} \bar{\gamma} \\ N\bar{S}_{yy}^{(2)} &= \bar{N} \bar{\gamma}_w \bar{C}_{N\bar{\gamma}} \\ N\bar{S}_{yy}^{(3)} &= \bar{\gamma}_w \frac{M}{M-1} (\bar{N} - 1) \end{aligned} \quad (7.6)$$

As $\bar{\gamma}$ does not depend on the scale or zoning and is a populations constant, the term $N\bar{S}_{yy}^{(1)}$ will increase linearly with \bar{N} . The other terms also include \bar{N} , and would be linear in \bar{N} if $\bar{\gamma}_w$ and $\bar{C}_{N\bar{\gamma}}$ were constant. However they are not constant, and these relationships with \bar{N} is what leads to the non-linearity in figure (7.4-b). Hence we can examine these factors in the following figure (7.5). The term $N\bar{S}_{yy}^{(2)}$ has little impact being about 20 times smaller than $N\bar{S}_{yy}^{(3)}$. The main effect is due to $N\bar{S}_{yy}^{(3)}$, which has a gentle upward curve, resulting in the observed non-linear relationship of $N\bar{S}_{yy}$ with \bar{N} .

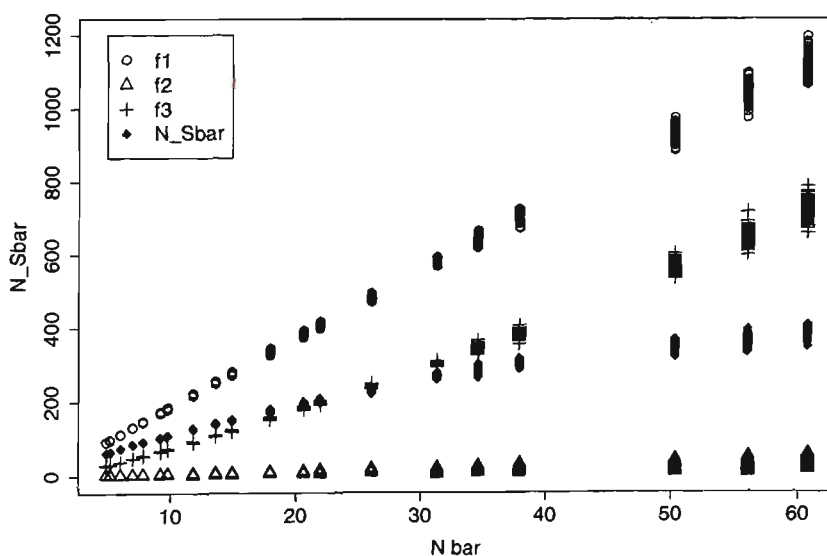


Figure 7.5. Relationship between \bar{N} and each factor of the $N\bar{S}_{yy}^{(1)}$ (\circ), $N\bar{S}_{yy}^{(2)}$ (Δ), and $N\bar{S}_{yy}^{(3)}$ ($+$). The \blacklozenge is the $N\bar{S}_{yy}$.

Ignoring the $N\bar{S}_{yy}^{(2)}$ and $N\bar{S}_{yy}^{(3)}$, the $N\bar{S}_{yy}$ is proportional to the S_{yy} (represented by $\bar{\gamma}$) with factor $\frac{N-1}{M-1} \approx \bar{N}$. But the $N\bar{S}_{yy}^{(2)}$ and $N\bar{S}_{yy}^{(3)}$ give a negative effect, hence the $N\bar{S}_{yy}$ is lower than $N\bar{S}_{yy}^{(1)}$. The separate plots of the $N\bar{S}_{yy}^{(2)}$ and $N\bar{S}_{yy}^{(3)}$ (Figure 7.6) give a clearer idea of the relationship between \bar{N} and these two factors. Both factors increases as the \bar{N} increases. The $N\bar{S}_{yy}^{(2)}$ factor shows a lot of variation

for each \bar{N} , which increases as \bar{N} increases, but this is much smaller than $N\bar{S}_{yy}^{(3)}$ which will be the main factors. The variation of $N\bar{S}_{yy}^{(3)}$ is much smaller for each \bar{N} .

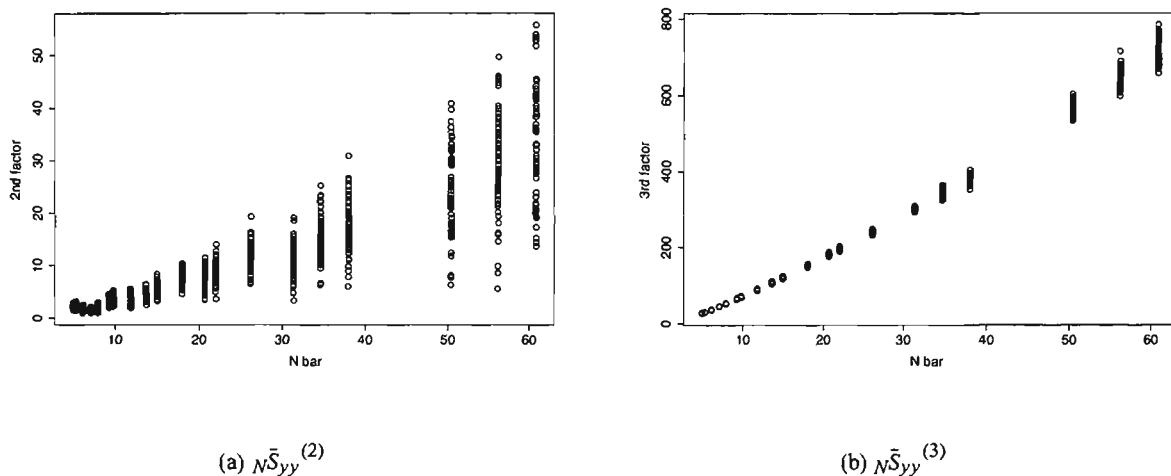


Figure 7.6. Separate plot between $N\bar{S}_{yy}^{(2)}$ and $N\bar{S}_{yy}^{(3)}$ versus \bar{N}

The factors $\tilde{\gamma}_w$ and $\bar{C}_{N\tilde{\gamma}}$ affect the value of $N\bar{S}_{yy}$. The factor $\tilde{\gamma}_w$ indicates the average of the $\tilde{\gamma}_g$, hence it is a measure of the average of the within group variance. Figure (7.7) shows the relationship between $\tilde{\gamma}_w$ and \bar{N} for this simulation study. The value of $\tilde{\gamma}_w$ increases as \bar{N} increases resulting in an increase in the weighted group level variance. The figure indicates that the $\tilde{\gamma}_w$ is proportional with the \bar{N} with a relation $a + \bar{N}^b$. In this simulation we got $\hat{a} = 4.988$ and $\hat{b} = 0.4687$ with $MSE = 0.0209$, and plotted it as a curve line in figure (7.7).

Figure (7.8) illustrates the behaviour of the covariance and correlation between N_g and $\tilde{\gamma}_g$ for this simulation. Figure (7.8-a) shows the scatter of $\bar{C}_{N\tilde{\gamma}}$ and little pattern is shown in the scatter points. The very small value of $\bar{C}_{N\tilde{\gamma}}$ shows why the term $N\bar{S}_{yy}^{(2)}$ has little impact on the value of $N\bar{S}_{yy}$. In (7.8-b), the points exhibit a positive association between $\bar{R}_{N\tilde{\gamma}}$ and N_g (with correlation 0.6719). Increasing \bar{N} tends to produce a larger correlation between N_g and $\tilde{\gamma}_g$.

In the same way, the aggregation effect, \hat{D} , can be illustrated by the figure (7.9). Figure (7.9-a) shows the aggregation effect in terms of difference (\blacklozenge), the \circ , \triangle , and $+$ are the first, second, and third factors respectively of the difference, $N\bar{S}_{yy} - S_{yy}$. The \blacklozenge is the result of the first, second, and third factors. The second and third factors produce a negative effect.

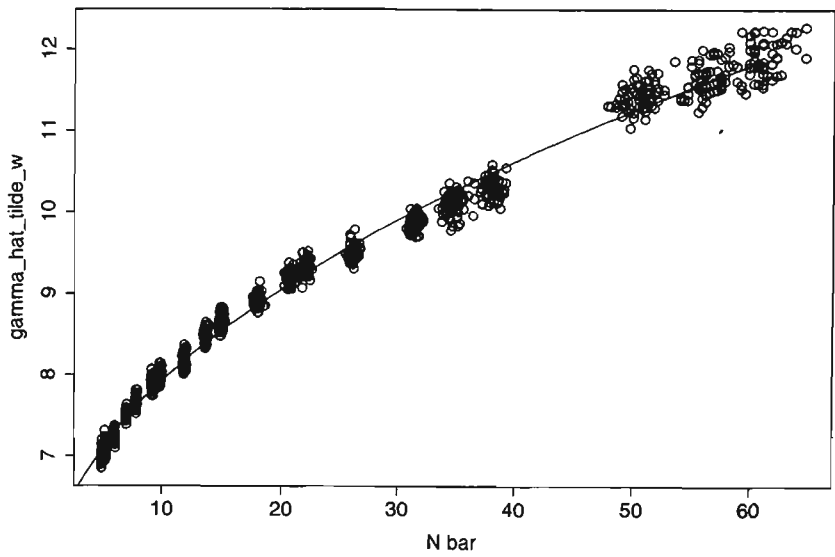


Figure 7.7. Relationship between \bar{N} and $\tilde{\gamma}_W$. The curve line indicates the relationship with the model $\hat{\gamma}_W = 4.988 + \bar{N}^{0.4687}$, with $MSE = 0.0209$.

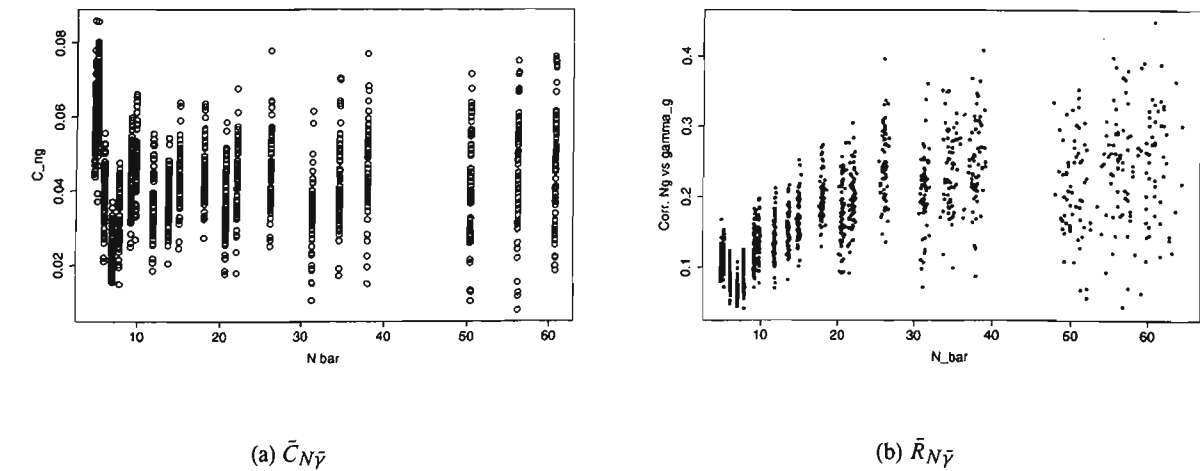
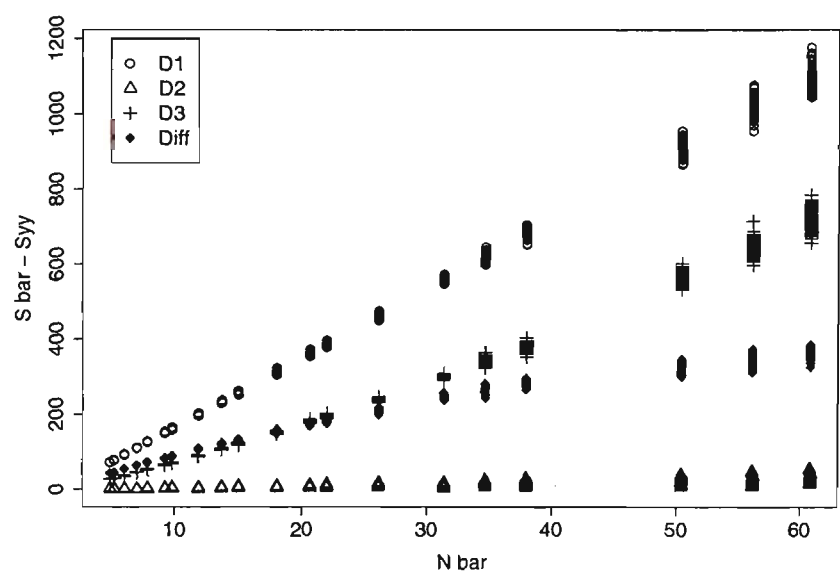
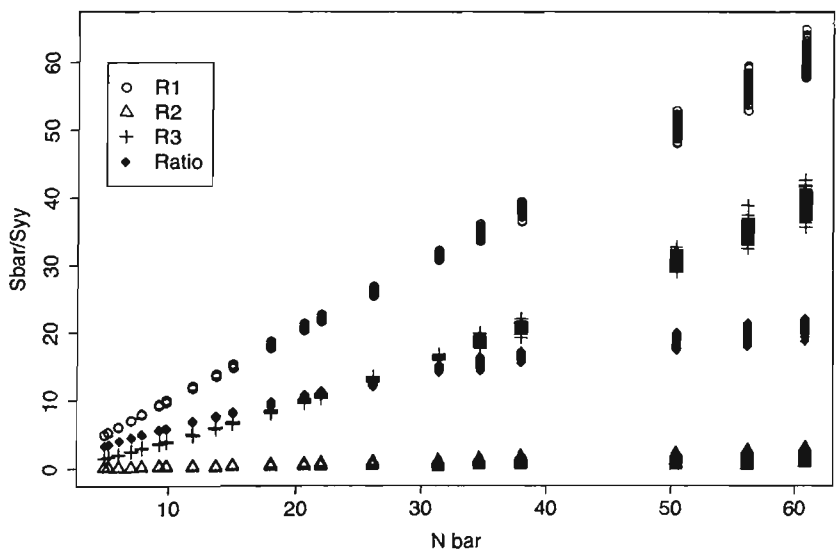


Figure 7.8. The scatter plot of $\bar{C}_{N\tilde{\gamma}}$ and $\bar{R}_{N\tilde{\gamma}}$



(a) difference

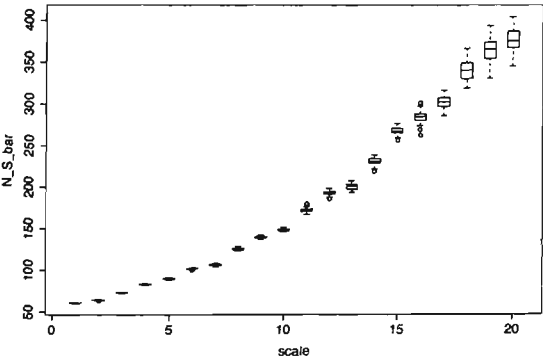


(b) ratio

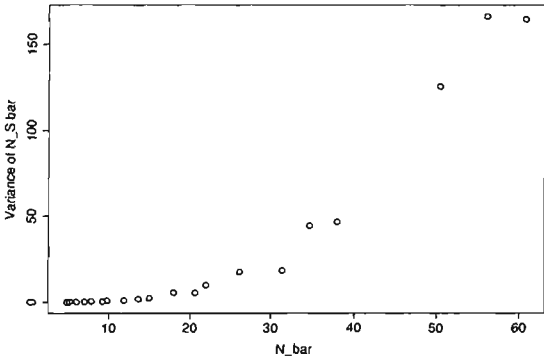
Figure 7.9. Relationship between \bar{N} with the aggregation effect, (a) in term of difference $\bar{N}\bar{S}_{yy} - S_{yy}$ and (b) in term of ratio $\frac{\bar{N}\bar{S}_{yy}}{S_{yy}}$.

Figure (7.9-b) shows the aggregation effect in terms of the ratio, \hat{R} , (\blacklozenge), and the \circ , \triangle , and $+$ are the first, second, and third factors respectively of the ratio, $\frac{N\bar{S}_{yy}}{\bar{S}_{yy}}$. The figure is similar to the figure of aggregation effect in terms of the difference (7.9-a).

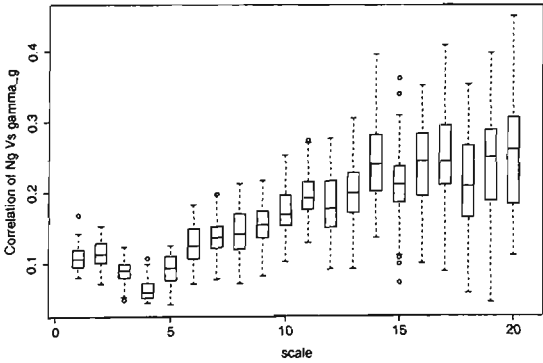
These figures are an illustration of the scale effect. Meanwhile the zoning effect may be observed in terms of the variation of the statistics at the same scale. For example the variation of the $N\bar{S}_{yy}$ and $\bar{R}_{N\bar{y}}$ at the same scale, can be illustrated in terms of boxplot, such as in figure (7.10). The first figure shows that the variation of the $N\bar{S}_{yy}$ at the lower scales is smaller than the variation at the larger scales (see figure 7.10-a and 7.10-b). Meanwhile the last figure (7.10-c) shows the same trend for the $\bar{R}_{N\bar{y}}$, which exhibits smaller variation at the lower scales than the larger scales.



(a)



(b)



(c)

Figure 7.10. (a) Boxplot of the $N\bar{S}_{yy}$ at each scale, (b) the scatter plot of the variation of $N\bar{S}_{yy}$ each scale versus \bar{N} , (c) boxplot of $\bar{R}_{N\bar{y}}$ at each scale.

Figure (7.11) shows the relationship between ${}_N\bar{S}_{yy}$ with $\tilde{\gamma}_w$ and the variation of $\tilde{\gamma}_w$ at each scale. The first plot indicates that the variation of $\tilde{\gamma}_w$ increases as ${}_N\bar{S}_{yy}$ increases. The curve line exhibits the relation between ${}_N\bar{S}_{yy}$ with $\tilde{\gamma}_w$ as explained by the model $\tilde{\gamma}_w = 2.2113 + {}_N\bar{S}_{yy}^{0.3732}$ ($MSE = 0.0814$). The second plot shows that variation of $\tilde{\gamma}_w$ versus the average ${}_N\bar{S}_{yy}$ at each scale level. The variation of $\tilde{\gamma}_w$ is relatively small compared with the variation of ${}_N\bar{S}_{yy}$ (see figure 7.10-b). The variance of $\tilde{\gamma}_w$ at each level increases extremely starting at ${}_N\bar{S}_{yy} = 250$ and onward. The third and fourth figures show the relationship of $\bar{C}_{N\tilde{\gamma}}$ with ${}_N\bar{S}_{yy}$. The third figure shows no clear pattern in the plot of $\bar{C}_{N\tilde{\gamma}}$ versus ${}_N\bar{S}_{yy}$. The variation of $\bar{C}_{N\tilde{\gamma}}$ at each scale increases as the ${}_N\bar{S}_{yy}$ increases. This is shown clearly in the fourth figure. But the variation of $\bar{C}_{N\tilde{\gamma}}$ is relatively smaller than the variation of $\tilde{\gamma}_w$. Hence these figures exhibit the key term of $\tilde{\gamma}_w$ compared with the $\bar{C}_{N\tilde{\gamma}}$.

The variation of the aggregation effect at each scale also shows the same situation, in terms of both the difference or ratio. Figure (7.12) shows the relationship of the variation of aggregation effect with the \bar{N} . The trend shows that the variations are very small for the small \bar{N} and get larger as the \bar{N} increase. The same trends are observed on both types of the aggregation effect. This is probably due to the number of groups decreasing as \bar{N} increases, since $M = \frac{N}{\bar{N}}$.

The scale factor has been observed so far is in terms of \bar{N} , which is a non-spatial scale factor. It is interesting to examine the scale factor in terms of the area of the group (A_g). Figure (7.13) shows the distribution of the coefficient variation of the group areas, which is spread from 0.1 to 0.3.

Figure (7.14-a) shows the relationship between the average group area size and ${}_N\bar{S}_{yy}$. The trend is similar to that exhibited in figure (7.4). The variation of the average group area for a particular \bar{N} is small, for example figure (7.14-b) shows the distribution of the ${}_N\bar{S}_{yy}$ over the the average groups area at the last \bar{N} . Figures (7.14-c) and (7.14-d) show the relationships of the average group area size with $\tilde{\gamma}_w$ and variance of the $\tilde{\gamma}_w$ at each scale. Figure (7.14-c) indicates that the relationship between average group area size with $\tilde{\gamma}_w$ can be explained by the model $\tilde{\gamma}_w = 5.6667 + \bar{A}^{0.5214}$ ($MSE = 0.0347$), where $\bar{A} = \frac{1}{M} \sum_g A_g$. Relationship of the variation of the $\tilde{\gamma}_w$ with the average group area is similar to the relationship of variation of the $\tilde{\gamma}_w$ with the average ${}_N\bar{S}_{yy}$ (see figure 7.11-b).

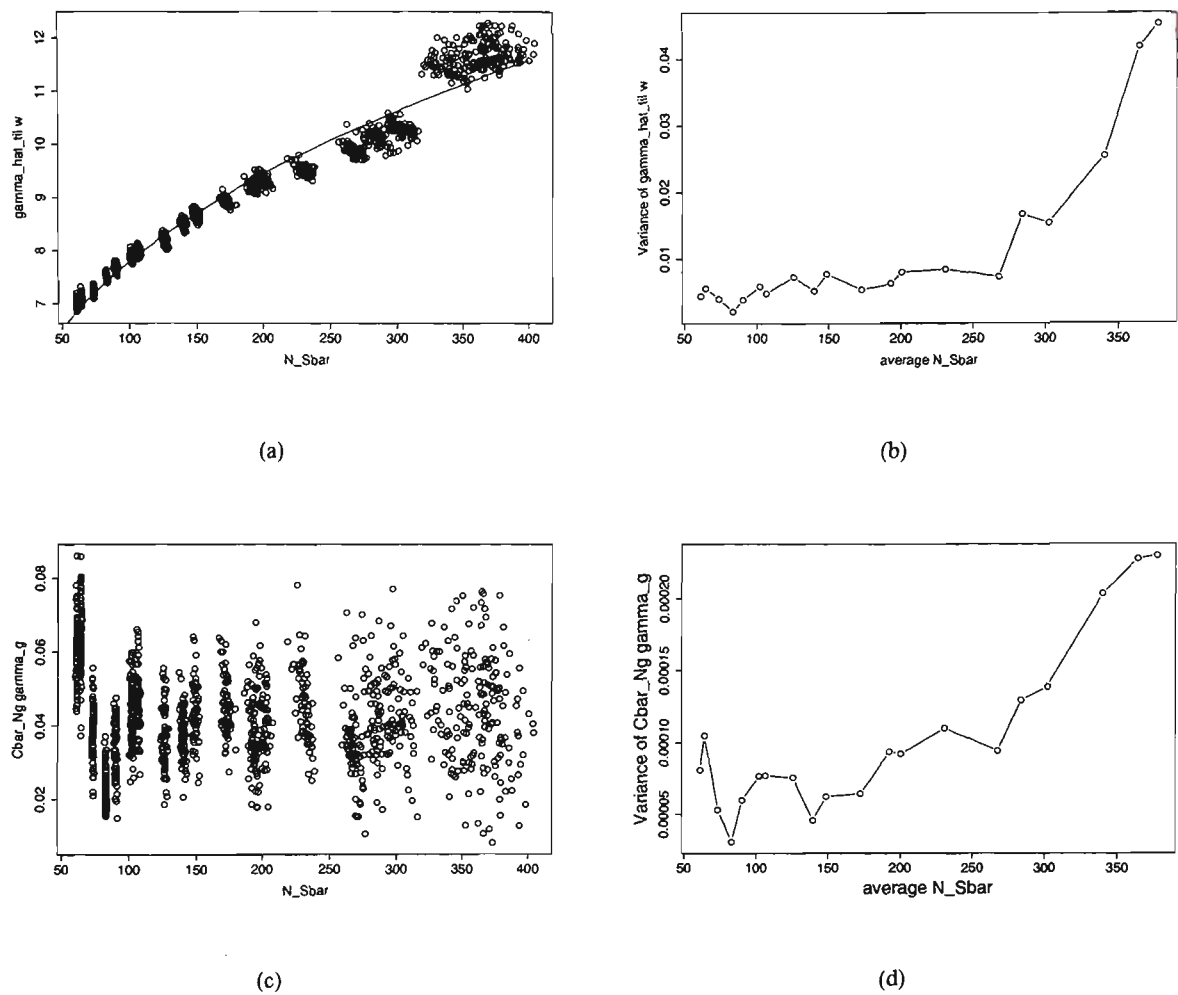
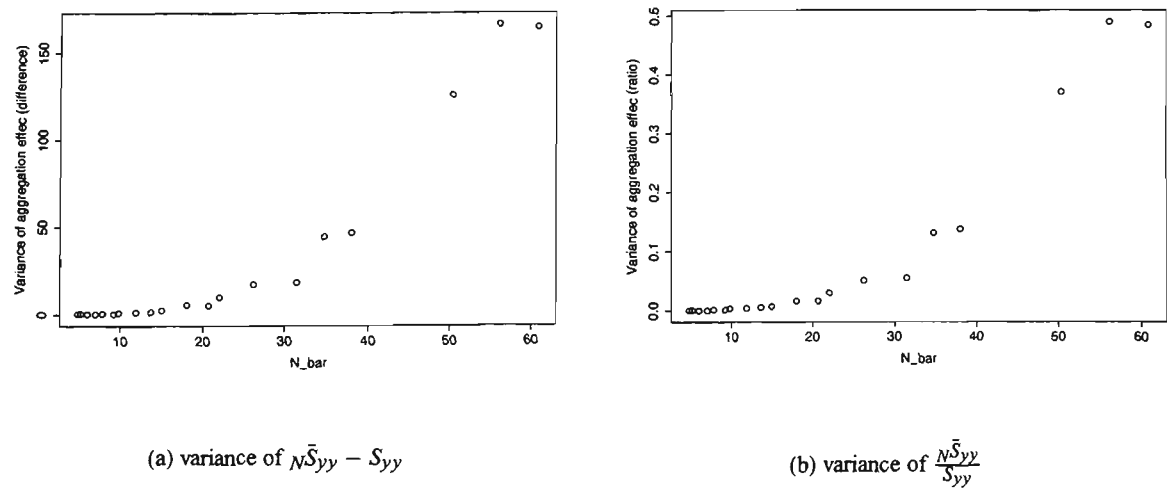


Figure 7.11. (a) Relationship between $\tilde{\gamma}_W$ with $N\bar{S}_{yy}$, (b) relationship of variation of $\tilde{\gamma}_W$ at each scale with the average of $N\bar{S}_{yy}$ at each scale, (c) relationship between $\bar{C}_{N\tilde{\gamma}}$ with $N\bar{S}_{yy}$, and (d) relationship of variation of $\bar{C}_{N\tilde{\gamma}}$ at each scale with the average of $N\bar{S}_{yy}$ at each scale.



(a) variance of $N\bar{S}_{yy} - S_{yy}$ (b) variance of $\frac{N\bar{S}_{yy}}{S_{yy}}$

Figure 7.12. Relationship between \bar{N} with the variance of the aggregation effect in term of difference and ratio at a particular \bar{N}

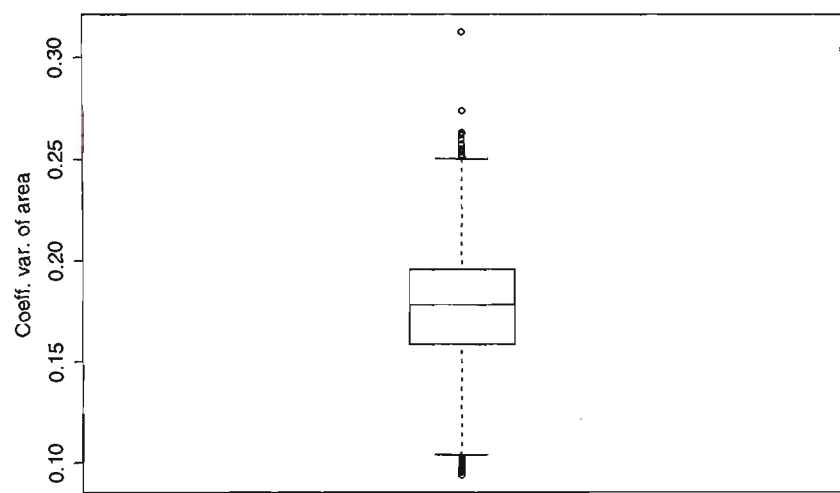


Figure 7.13. Coefficient of variation of the \mathcal{A}_g over the whole simulation

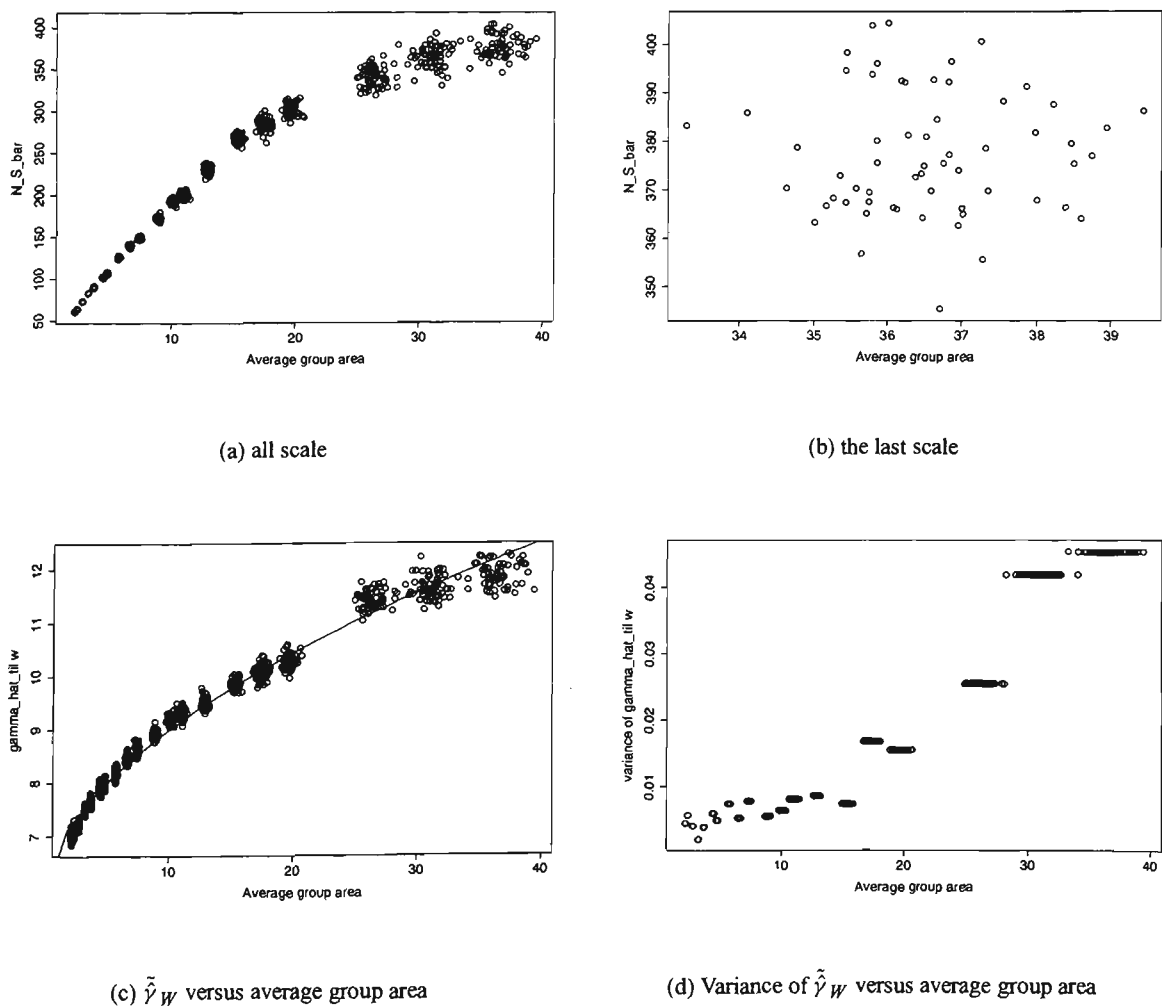


Figure 7.14. Relationship between \bar{N} with the average group area, (a) for whole scale and (b) only the last scale

7.5 The use of the MAUP as an analysis tool : spatial analysis development

The discussion in this chapter has shown how the aggregation effect for different scales and zoning can be related to $\tilde{\gamma}_w$. In this section the theory will be developed to show that the MAUP can be used to estimate the parameters of a semivariogram model. This theoretical development is based on the semivariogram theorems(section 4.3 and chapter 5). At this point, the univariate case will be considered. The semivariogram is defined under two assumptions, the intrinsic stationarity and the second order stationarity (see section 5.1).

The semivariogram can be defined as a statistic which measures the variance of pairs of the observations, and can be related to the distance between the pair of observations. The semivariogram becomes a measure of the spatial association for the data. Cliff and Ord (1981) also noted a connection between semivariogram and spatial autocorrelation. Table (7.3) shows the spatial autocorrelation from different groupings. We will investigate the connection of these differences with the semivariogram, and the possibilities for using it in semivariogram analysis.

Recall that the empirical semivariogram, $\hat{\gamma}_{ij}$, is defined as follow

$$\hat{\gamma}_{ij} = \frac{1}{2}(Y_i - Y_j)^2 \quad (7.7)$$

Theorem 7.5.1. *The ratio of the weighted group level variance over the individual level variance can be approximated by*

$$\frac{N\bar{S}_{yy}}{S_{yy}} \approx \frac{M}{M-1} \left(\bar{N} - (\bar{N} - 1) \frac{\tilde{\gamma}_w}{S_{yy}} \right) \quad (7.8)$$

Proof. Consider equation (4.73) $\bar{C}_{N\tilde{\gamma}} \approx 0$, then we have

$$\frac{N\bar{S}_{yy}}{S_{yy}} = \frac{N-1}{M-1} \left(1 - \frac{M(\bar{N}-1)}{N-1} \frac{\tilde{\gamma}_w}{\tilde{\gamma}} \right)$$

Assume that $\tilde{\gamma} = S_{yy}$ and large N , such that $\frac{N}{N-1} \approx 1$. Hence it can be approximated by

$$\begin{aligned} \frac{N\bar{S}_{yy}}{S_{yy}} &\approx \bar{N} \frac{M}{M-1} \left(1 - \frac{\bar{N}-1}{\bar{N}} \frac{\tilde{\gamma}_w}{S_{yy}} \right) \\ &= \frac{M}{M-1} \left(\bar{N} - (\bar{N} - 1) \frac{\tilde{\gamma}_w}{S_{yy}} \right) \end{aligned}$$

□

Theorem (7.5.1) suggests that if S_{yy} was known then this can be used to estimate $\tilde{\gamma}_w$, and we can then change scales and zoning to see how $\tilde{\gamma}_w$ changes with scale and zoning. If S_{yy} was not available, then trying different scale and zoning, can give an idea of how $\frac{\tilde{\gamma}_w}{S_{yy}}$ varies. The advantage of this approach is that we only have to consider within group distances, which can be approximated by (5.68). In the next section it will be shown how this can be used to obtain an estimate of the parameters of a semivariogram model.

7.5.1 Exponential model of semivariogram

Consider the exponential semivariogram model in (5.9) which can be rewritten as

$$\gamma(d_{ij}) = s - (s - n) \cdot \exp\left[\frac{-3d_{ij}}{r}\right] \quad (7.9)$$

where n , s , r are nugget, sill, and range respectively.

Matérn (1986) considered the variation in the group level variance for difference shapes of groups of unit area. He assumed that the points are uniformly distributed within the study region, and the study region was partitioned into non-overlapping groups. He concluded that the circle shape may produce the largest group level variance compared with other shapes such as rectangle, triangle, hexagon, and square.

For simplicity, let assume that variation in shape and size is not great. The previous work indicated that the second order term of the Taylor series expansion of $\gamma()$ is relatively small compared with $\gamma()$ and also the second order term of $\gamma()$ for the exponential model is approaching zero as the group area size increases (see figure 5.7-a). Hence ignoring the second order, we have

$$\bar{\gamma}_g = s - (s - n) \exp\left[\frac{-3\bar{d}_g}{r}\right] \quad (7.10)$$

Theorem 7.5.2. Using the Taylor series expansion, the $\exp\left[\frac{-3\bar{d}_g}{r}\right]$ term can be approximated by

$$\exp\left[\frac{-3\bar{d}_g}{r}\right] \approx 1 - \frac{3\bar{d}_g}{r} \quad (7.11)$$

Proof. The $\exp()$ term can be represented by Taylor series expansion as

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Hence

$$\exp\left[\frac{-3\bar{d}_g}{r}\right] = 1 + \frac{-3\bar{d}_g}{r} + \frac{9\bar{d}_g^2}{2r^2} + \frac{-27\bar{d}_g^3}{6r^3} + \dots$$

or it can be rewritten

$$\exp\left[\frac{-3\bar{d}_g}{r}\right] = 1 - \frac{3\bar{d}_g}{r} \cdot \left(1 - \frac{3\bar{d}_g}{2r} + \frac{9\bar{d}_g^2}{6r^2} + \dots\right)$$

Consider the component within the bracket will approach the unit (one) then approximation of the $\exp\left[\frac{-3\bar{d}_g}{r}\right]$

is

$$\exp\left[\frac{-3\bar{d}_g}{r}\right] \approx 1 - \frac{3\bar{d}_g}{r}$$

□

Theorem (7.5.2) is applicable in situations where $\frac{3\bar{d}_g}{r}$ is close to zero, that is the case when groups are smaller than r . Applying theorem (7.5.2) into (7.10), we have an approximation for the $\bar{\gamma}_g$, that is

$$\bar{\gamma}_g = s - (s - n) \cdot \left(1 - \frac{3\bar{d}_g}{r}\right) \quad (7.12)$$

Substituting (5.68) into (7.12) gives

$$\bar{\gamma}_g \approx n + (s - n) \cdot 3 \cdot \frac{k_1 \sqrt{A_g}}{r} \quad (7.13)$$

where $\bar{d}_g \approx k_1 \sqrt{A_g}$, for A_g is the area of the g th group (Matérn, 1986).

Corollary 7.5.3. *The $\tilde{\gamma}_W$ can be approximated as*

$$\tilde{\gamma}_W \approx n + \frac{s - n}{r} \cdot 3 \cdot k_1 \tilde{A}_W^* \quad (7.14)$$

for

$$\tilde{A}_W^* = \frac{1}{M} \sum_g \sqrt{A_g}$$

Proof.

$$\begin{aligned} \tilde{\gamma}_W &= \frac{1}{M} \sum_g \bar{\gamma}_g \\ &= \frac{1}{M} \sum_g \left[n + (s - n) \cdot 3 \cdot \frac{k_1 \sqrt{A_g}}{r} \right] \\ &= n + \frac{s - n}{r} 3k_1 \sum_g \frac{\sqrt{A_g}}{M} \end{aligned}$$

□

The $\tilde{\mathcal{A}}_W^*$ can be considered a spatial factor in the aggregation effect. Its value may change as the scale or zoning change. For example, it can be determined in the following hypothetical example of Figure (7.15).

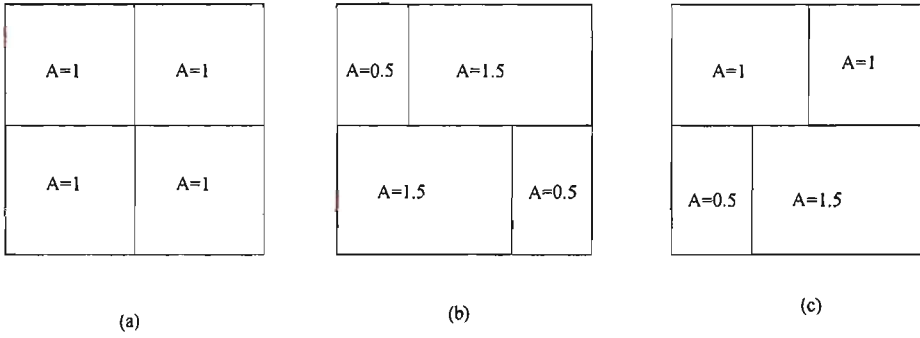


Figure 7.15. Some examples of the $\tilde{\mathcal{A}}_W^*$ values from different zoning at the region of area 4.0, (a) $\tilde{\mathcal{A}}_W^* = 1.0$, (b) $\tilde{\mathcal{A}}_W^* = 0.9659$, and (c) $\tilde{\mathcal{A}}_W^* = 0.9830$

Theorem 7.5.4. *The ratio of weighted group level variance over the individual level variance can be approximated as*

$$\frac{N\tilde{S}_{yy}}{S_{yy}} \approx \frac{M}{M-1} \left(1 + (\bar{N} - 1) \frac{s-n}{s} \left[1 - 3k_1 \frac{\tilde{\mathcal{A}}_W^*}{r} \right] \right) \quad (7.15)$$

for $k_1 = 0.5108$ (see equation 5.68).

Proof. Substituting (7.14) into (7.8) gives

$$\begin{aligned} \frac{N\tilde{S}_{yy}}{S_{yy}} &= \bar{N} \frac{M}{M-1} \left(1 - \frac{\bar{N}-1}{\bar{N}} \frac{1}{S_{yy}} \left[n + (s-n) \frac{3k_1 \tilde{\mathcal{A}}_W^*}{r} \right] \right) \\ &= \frac{M}{M-1} \left(\bar{N} - (\bar{N}-1) \frac{n}{S_{yy}} - (\bar{N}-1) \frac{s-n}{S_{yy}} \frac{3k_1 \tilde{\mathcal{A}}_W^*}{r} \right) \end{aligned}$$

Assume S_{yy} is equal to the sill s , then

$$\begin{aligned} \frac{N\tilde{S}_{yy}}{S_{yy}} &= \frac{M}{M-1} \left(\bar{N} - (\bar{N}-1) \frac{n}{s} - (\bar{N}-1) \frac{s-n}{s} \frac{3k_1 \tilde{\mathcal{A}}_W^*}{r} \right) \\ &= \frac{M}{M-1} \left((\bar{N}-1+1) - (\bar{N}-1) \frac{n}{s} - (\bar{N}-1) \frac{s-n}{s} \frac{3k_1 \tilde{\mathcal{A}}_W^*}{r} \right) \\ &= \frac{M}{M-1} \left((\bar{N}-1) \frac{s-n}{s} \left[1 - \frac{3k_1 \tilde{\mathcal{A}}_W^*}{r} \right] + 1 \right) \end{aligned}$$

□

Corollary 7.5.5. *Assume that $S_{yy} = s$, then the weighted group level variance can be approximated by*

$$N\tilde{S}_{yy} \approx \frac{M}{M-1} \left(s + (\bar{N}-1)(s-n) \left[1 - 3k_1 \frac{\tilde{\mathcal{A}}_W^*}{r} \right] \right) \quad (7.16)$$

The corollary (7.5.5) provides a connection between the non-spatial statistic ${}_N\bar{S}_{yy}$ and the spatial parameter population as described by n , s , and r . The important aspect also is shown by the \tilde{A}_W^* , which is the group area factor. Result (7.16) suggest a fairly straight forward method for estimating the parameter. We could calculate ${}_N\bar{S}_{yy}$ for different scales and zoning, under the assumption that the groups areas are within the range r . This will give a set of a values of ${}_N\bar{S}_{yy}$, \bar{N} , M , \tilde{A}_W^* , and then non-linear regression can be used to estimate the unknown parameters n , s , r . Equation (7.16) can be thought of as an another semivariogram model from a spatial structure with the \tilde{A}_W^* factor as an analog of the \bar{d} . The impact of the scale is clearly seen through the appearance of the factor $(\bar{N} - 1)$. Zoning affects \tilde{A}_W^* . Although this derivation was done for the exponential semivariogram model case, the idea can be applied for other semivariogram models easily.

7.5.2 The scale effect

Given some data for a particular grouping, the original groups can be formed into larger groups. This can be done several times, each resulting in an average group size :

$$\bar{N}_1, \dots, \bar{N}_K; \quad \text{for } k = 1, \dots, K$$

where $\bar{N}_k = N/M_k$. Each realization also gives a different ${}_N\bar{S}_{yy}$ and \tilde{A}_W^* , say

$${}_N\bar{S}_{yy1}, \dots, {}_N\bar{S}_{yyK}; \quad \text{and } {}_1\tilde{A}_W^*, \dots, {}_K\tilde{A}_W^*$$

This can be drawn as follows in figure (7.16).

Based on model (7.15) we have

$${}_N\bar{S}_{yyk} = \frac{M_k}{M_k - 1} \left(s + (\bar{N}_k - 1)(s - n) \left[1 - 3k_1 \frac{{}_k\tilde{A}_W^*}{r} \right] \right) \quad (7.17)$$

Given the values of ${}_N\bar{S}_{yyk}$, \bar{N}_k , k_1 , and ${}_k\tilde{A}_W^*$ then equation (7.17) has three unknown parameters n , s , and r . In relation to the spatial autocorrelation, let define $\rho(0) = (1 - \frac{n}{s})$, then we may have

$${}_N\bar{S}_{yyk} = \frac{M_k}{M_k - 1} \cdot s \left(1 + (\bar{N}_k - 1)\rho(0) \left[1 - 3k_1 \frac{{}_k\tilde{A}_W^*}{r} \right] \right)$$

where the $\rho(0)$ can be interpreted as the intra-household spatial correlation. If the observations are IID then the $n = s$ and $\rho(0) = 0$. This case implies ${}_N\bar{S}_{yyk}$ is proportional to the s by a factor $\frac{M_k}{M_k - 1}$. If

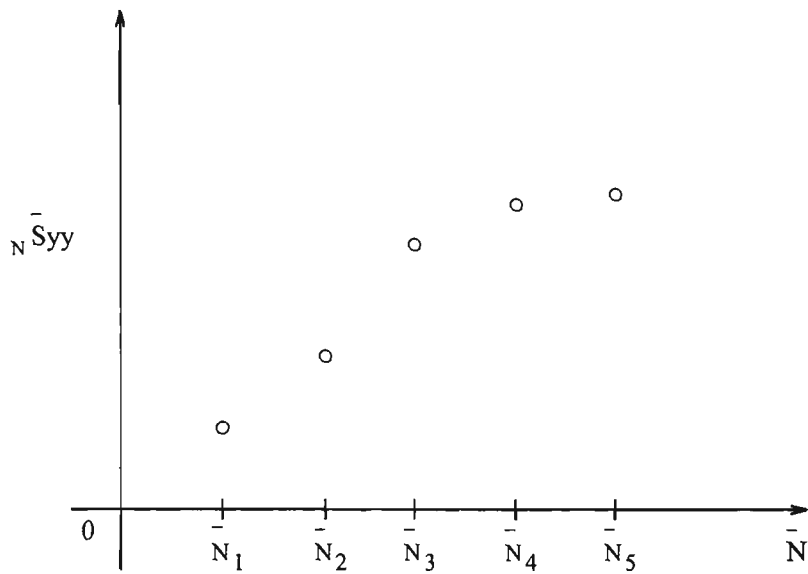


Figure 7.16. The weighted group level variance at different scale

the observations are not IID, then variation in $N\bar{S}_{yyk}$ depends on the magnitude of the $\rho(0)$ and also the value of r . Applying non-linear regression methods similar to those in section (5.4) we can estimate the parameter n , s , r , and therefore $\rho(0)$ as well.

7.5.3 The zoning effect

The zoning effect can be used by varying the arrangement of the groups at a particular \bar{N} . Suppose that we have $t = \{1, \dots, T\}$ realization of the zoning. This situation implies a variation in $N\bar{S}_{yy}$ and \tilde{A}_W^* , that is

$$N\bar{S}_{yy1}, \dots, N\bar{S}_{yyT}; \quad \text{and} \quad {}_1\tilde{A}_W^*, \dots, {}_T\tilde{A}_W^*$$

This zoning realization can be drawn as in figure (7.17).

Based on model (7.15)

$$N\bar{S}_{yyt} = \frac{M}{M-1} \left(s + (\bar{N} - 1)(s - n) \left[1 - 3k_1 \frac{{}_t\tilde{A}_W^*}{r} \right] \right) \tag{7.18}$$

Equation (7.18) shows that the value of $N\bar{S}_{yy}$ at the t th realization of the zoning is dependent on the \tilde{A}_W^* . The \tilde{A}_W^* is reflecting the area of the group, which may change on every realization of the zoning.

Given T realization of the zoning scheme, then we can compute the $N\bar{S}_{yy}$ and \tilde{A}_W^* for every realization. Given the values of k_1 then we can estimate the parameter n , s , and r by using regression method. The regression method relies on there being variation in \tilde{A}_W^* and so, for the same scale, it is desirable to use zoning with as many different values of \tilde{A}_W^* as possible.

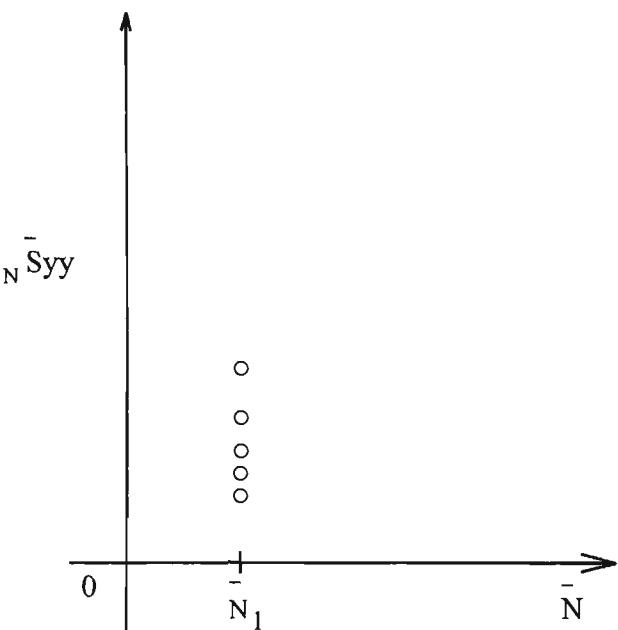


Figure 7.17. The weighted group level variance at different zoning scheme

7.6 Discussion

Equation (7.8) expresses the aggregation effect in terms of the ratio between weighted group level variance and the individual level variance. The case of the exponential semivariogram model leads to equation (7.15).

Equation (7.16) establishes a link between a non- spatial statistic ($\bar{N}\bar{S}_{yy}$) and spatial parameters such as nugget, sill, range, and also the group area components. The equation can also be considered as a starting point in using the MAUP to investigate the individual level spatial parameters. This equation can be sketched as displayed in figure (7.18). Using the \bar{N} as the x-axis, the figure shows that variation at a particular \bar{N} is due to the zoning effect and variation among \bar{N} indicates the scale effect. Using (7.16) we could develop a procedure to estimate n , s , r . This could then be applied to data such as presented in figure (7.18). The independent variable in (7.16) is \tilde{A}_w^* , and so different zoning and scales should be used that create appreciable variation in the variable. However, (7.16) was obtained using the assumption that $3\tilde{d}_g/r$ was small, and so the groupings used should be consistent with this assumption. Alternatively higher order terms in the Taylor series expansion of $\exp()$ could be used.

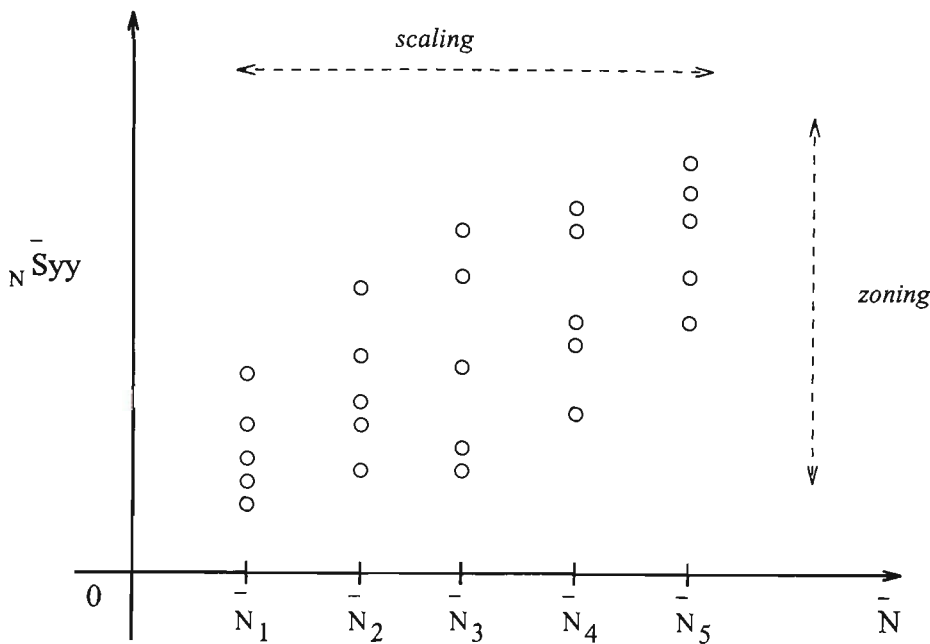


Figure 7.18. The scale and zoning effect

This chapter discuss the theoretical aspects of using the MAUP as a tool to estimate the individual level variogram parameters. It is feasible theoretically, but further empirical investigation and clarification is required.

7.7 Summary

The MAUP can be used as a tool in variogram analysis, and can be used to estimate the individual level variogram parameters from the group level data. Different zoning and scaling realizations are needed to estimate the individual level variogram parameters. The estimated parameters can be used to adjust the estimated group level variogram parameters, which is useful for further analysis, such interpolation procedure by kriging method (Carrat & Valleron, 1992).

Equation (7.1) shows the expectation of the aggregation effect in terms of difference between weighted group level variance and the individual level variance. The equation formulated the scale and zone effect in term of $\bar{C}_{N\bar{\Delta}}$ and $\bar{C}_{N\bar{\Delta}}$. Consider the zoning effect in the case with \bar{N} constant, which implies the $\bar{C}_{N\bar{\Delta}}$ and $\bar{C}_{N\bar{\Delta}}$ are zero. Hence the expectation of the aggregation effect is

$$E(N\bar{S}_{yy} - S_{yy}) = -\frac{M(\bar{N} - 1)}{M - 1} (\bar{\Delta} - \tilde{\Delta}_w) \quad (7.19)$$

Chapter 8

Empirical Analysis

In this section the implications and implementation of the theory and methods described in the previous chapter, are illustrated through an analysis of the Australian Population and Housing Census 1991 data. The census data are available as group level data, and the Adelaide region is chosen as the study area. Appendix (G) contains a description of the characteristics being used in the Adelaide region.

8.1 Geographical aspects of Adelaide region

Adelaide is the capital city of South Australia, which had a population of approximately one million people in 1991. Geographically the region is located around 138.47 to 138.74 degrees of longitude and -35.22 to -34.66 degrees of latitude. The region covers an area of 670.57 km^2 . A map of Adelaide region is displayed in Figure (8.1a).

In the 1991 Australian Census of Population and Housing the Adelaide region was partitioned into 1713 collection districts (CD). The CD's boundaries may be viewed in Figure (8.1a). The CDs have an average area of 0.39 km squares, with a range from a minimum of 0.04 km^2 up to maximum 12.64 km^2 . There were 767,030 people aged 15 or more counted at the census night, and the average number of people 15 years and more per CD is 447.77

In the Australian census, the collection district is the smallest geographical unit used for data dissemination (Castles, 1991). In the 1991 Australian Census of Population and Housing, a spatial characteristics was attached to each CD, that is a CD's centroid. The CD can be considered approximately as a polygon,

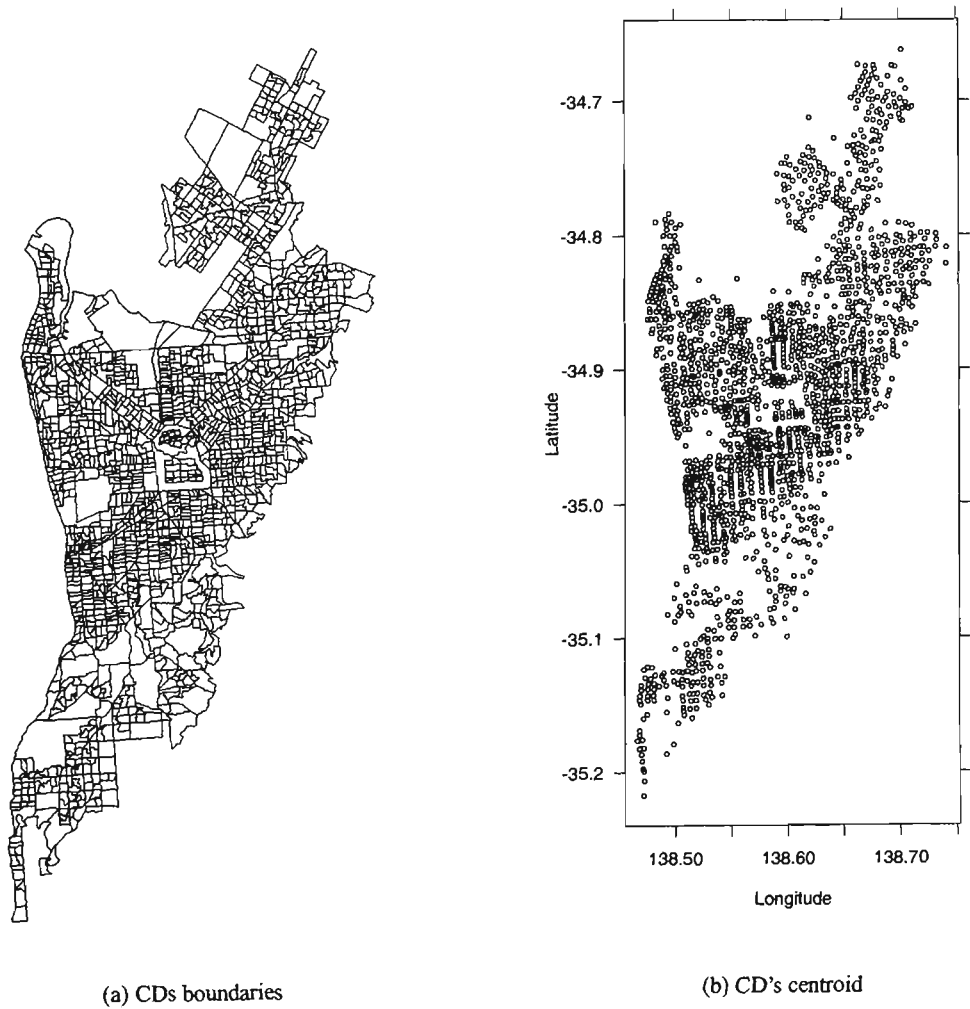


Figure 8.1. Adelaide region, (a) CD boundaries – based on Australian Census 1991, (b) CD's centroid. The region is approximately 33 km wide and 55 km long.

with its centre determined according to the boundaries of the polygon. Griffith and Amrhein (1991) discussed a procedure to determine a centroid for any shape of the polygon. In the 1991 census output, the CD centroids are recorded in terms of longitude and latitude.

The fundamental unit of observation in the census is a person, which are clustered within the households. The census did not record the geographic location of the households, but did record the CD in which the households were located. Hence the geographical locations of all households within any particular CD are represented by the geographical location of the CD's centroid. Figure (8.1b) shows the locations of the CD's centroid in the Adelaide region.

The distance between CDs can be represented as the distance between the centroids of the CDs. For the Adelaide region, the smallest distance between two CDs is 0.11 km and the largest distance between two CDs is 66.51 km.

Four sets of variables will be considered in this chapter.

1. The first set concerns the labor force, which includes three characteristics, the employment rate, unemployment rate, and labor force participation rate itself.
2. The second set concerns income characteristics, which is categorized into three different categories, income less than 20000, income between 20000 and 40000, and income over 40000. The income was stated as annual income in Australian dollars.
3. The third set concerns the nature of the employment, which includes three different characteristics, wage or salary earner, self employed persons, and employer.
4. The fourth set covers the aspect of qualification achievements, which are the formal and informal qualification rate.

8.2 Description of the characteristics

Data concerning the characteristics were recorded as the frequency or number of people at collection district level. These frequencies can be converted to a rate value. The rate is defined by a ratio between the frequency and the group size, which is number of people within the CD. The rate is defined by

$$p_g = \frac{f_g}{N_g} \tag{8.1}$$

here, f_g is the frequency or number of people belonging to a specific characteristics within the g th group, and N_g is number of people whose age is 15 years and over within the g th group. Table (8.1) summarizes the distribution of the rates of each studied characteristic.

Table 8.1. Description of the characteristics (rate %)

| Characteristic | Min. | Q1 | Median | Mean | Q3 | Max. |
|----------------------|------|-------|--------|-------|-------|--------|
| Employment | 0.00 | 45.50 | 53.07 | 53.01 | 61.24 | 98.69 |
| Unemployment | 0.00 | 4.77 | 6.55 | 7.27 | 9.06 | 29.65 |
| Labor part. | 0.00 | 53.40 | 60.28 | 60.28 | 68.13 | 98.69 |
| Income < 20000 | 6.67 | 55.21 | 61.24 | 61.32 | 67.27 | 100.00 |
| Income 20000 - 40000 | 0.00 | 22.31 | 26.25 | 25.96 | 29.71 | 80.00 |
| Income 40000 - over | 0.00 | 1.71 | 3.64 | 5.24 | 7.28 | 71.31 |
| Wage earner | 0.00 | 39.14 | 45.32 | 45.63 | 52.35 | 100.00 |
| self employed | 0.00 | 3.11 | 4.25 | 4.25 | 5.33 | 10.29 |
| employer | 0.00 | 1.31 | 2.34 | 2.87 | 3.83 | 19.24 |
| Formal qualif. | 0.00 | 6.37 | 10.73 | 12.80 | 17.54 | 42.86 |
| Informal qualif. | 0.00 | 10.60 | 13.32 | 13.35 | 15.98 | 26.96 |

The mean rate in table (8.1) was calculated as an ordinary mean of the group level data. This is called as unweighted group mean, and is defined by

$$\tilde{p} = \frac{1}{M} \sum_{g=1}^M p_g \tag{8.2}$$

where M is number of CDs within the study region. The unweighted group level variance is

$${}_1\tilde{S}_{yy} = \frac{1}{M-1} \sum_{g=1}^M (p_g - \tilde{p})^2 \tag{8.3}$$

The weighted mean of the rate can be determined by

$$P = \frac{\sum_{g=1}^M N_g \cdot p_g}{\sum_{g=1}^M N_g} \tag{8.4}$$

the weighted group level variance is defined by

$${}_N\bar{S}_{yy} = \frac{1}{M-1} \sum_{g=1}^M N_g \cdot (p_g - P)^2$$

(8.5)

Since the data involved are actually dichotomous at the individual level data, we can calculate the individual level variance from the overall proportion. For the unweighted group level mean rate, we have the

$$\tilde{S}_{yy} = \tilde{p} \cdot (1 - \tilde{p})$$

(8.6)

and from the weighted group mean rate

$$S_{yy} = P \cdot (1 - P)$$

(8.7)

Note that S_{yy} is the individual level variance.

Table 8.2. The unweighted and weighted mean rate and variance

| Variables | Mean | | Indiv. level var. | | Group level var. | | |
|----------------------|-------------|--------|-------------------|----------|--------------------|--------------------|---------------------------|
| | \tilde{p} | P | \tilde{S}_{yy} | S_{yy} | ${}_1\bar{S}_{yy}$ | ${}_N\bar{S}_{yy}$ | ${}_N\bar{S}_{yy}/S_{yy}$ |
| Employment | 0.5301 | 0.5363 | 0.2491 | 0.2487 | 0.0134 | 5.8739 | 23.6184 |
| Unemployment | 0.0727 | 0.0718 | 0.0674 | 0.0666 | 0.0013 | 0.5321 | 7.9895 |
| Labor part. | 0.6028 | 0.6081 | 0.2394 | 0.2383 | 0.0112 | 4.9085 | 20.5980 |
| Income < 20000 | 0.6132 | 0.6105 | 0.2372 | 0.2378 | 0.0082 | 3.4513 | 14.5135 |
| Income 20000 - 40000 | 0.2596 | 0.2617 | 0.1922 | 0.1932 | 0.0044 | 1.8344 | 9.4948 |
| Income 40000 - over | 0.0524 | 0.0583 | 0.0497 | 0.0549 | 0.0026 | 1.4366 | 26.1676 |
| Wage earner | 0.4563 | 0.4620 | 0.2481 | 0.2486 | 0.0100 | 4.3664 | 17.5640 |
| self employed | 0.0425 | 0.0430 | 0.0407 | 0.0411 | 0.0003 | 0.1198 | 2.9148 |
| employer | 0.0287 | 0.0285 | 0.0279 | 0.0277 | 0.0006 | 0.2252 | 8.1300 |
| Formal qual. | 0.1280 | 0.1250 | 0.1116 | 0.1094 | 0.0073 | 3.1227 | 28.5439 |
| Informal qual. | 0.1335 | 0.1355 | 0.1157 | 0.1171 | 0.0015 | 0.6590 | 5.6277 |

Table (8.2) shows that the difference between the unweighted and weighted mean of the rates are very small. There is a large difference between the unweighted and weighted group level variance, where the unweighted is smaller than the weighted one. Meanwhile the individual level variances calculated from the unweighted and weighted means are very close. Steel and Holt (1996b) show that ${}_N\bar{S}_{yy}$ will be unbiased

for S_{yy} when the groups are randomly formed. The difference between the weighted group level variance and the individual level variance shown in table (8.2) demonstrate that there are large aggregation effects in the Adelaide census data.

The statistics in table (8.1) and (8.2) are non-spatial statistics, which do not give any information on spatial variability. This variability can be investigated by looking at the changes in the observation values relative to their locations. Two approaches will be used, the first involves looking the spatial distribution graphically and the second involves applying semivariogram and cross-semivariogram analysis.

8.3 Spatial graphical description of the characteristics

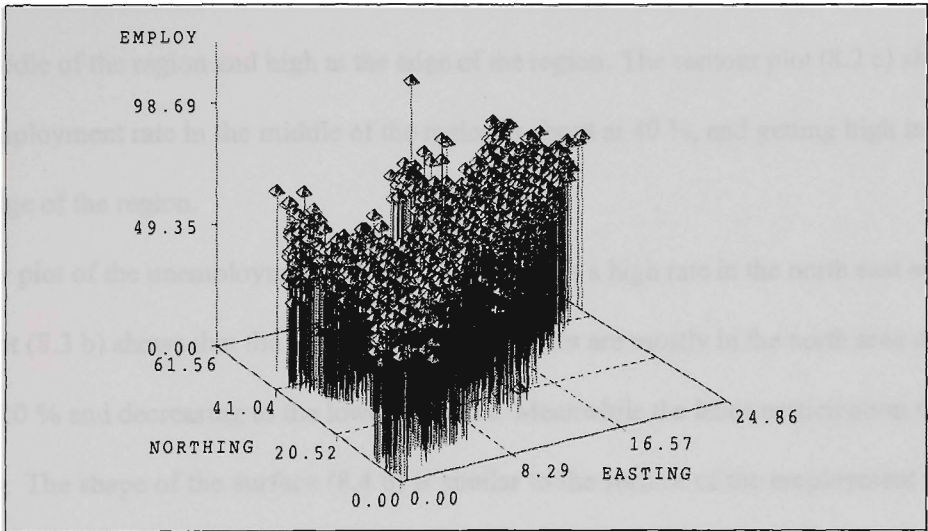
We will use three types of graphical representation to analyse the spatial variability in the data. These are 3-d scatter plot, surface plot, and contour plot. The 3-d scatter plot is a plot of the observations points according to their location in a cartesian coordinate system and their characteristic's value as the height of the points. The high and low of the characteristic values are shown across the region and represent a rough surface. The surface plot is a plot of surface generated according to the local regression smoother method (Venables & Ripley, 1994). The local regression smoother method calculates smoothed characteristic values according to their location in easting and northing, which is formulated by the following model

$$Y_i = \beta_e e_i + \beta_n n_i + \beta (e_i \times n_i)$$

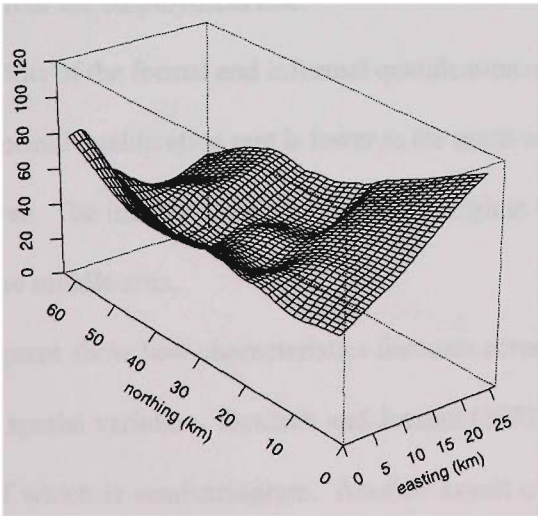
where (e_i, n_i) is the geographical location of the observation Y_i in easting and northing coordinate, and the β_e, β_n, β are parameters of the coefficient $e, n, (e \times n)$, respectively. In S-Plus (Kaluzny, Vega, Cardoso, & Shelly, 1998), this method is defined by the function

```
loess(yi~ei*ni, degree=1)
```

The contour plot is another representation of the generated surface in a two dimensional axis system. The plot contains dots and lines, which show the locations of the observations point and the contour line, respectively. The contour lines are also generated by the local regression smoother method. The results are shown in figures (8.2), (8.3), (8.4) for the labor force characteristics, and figure (A.1) through (A.8) of appendix (A) for other variables.

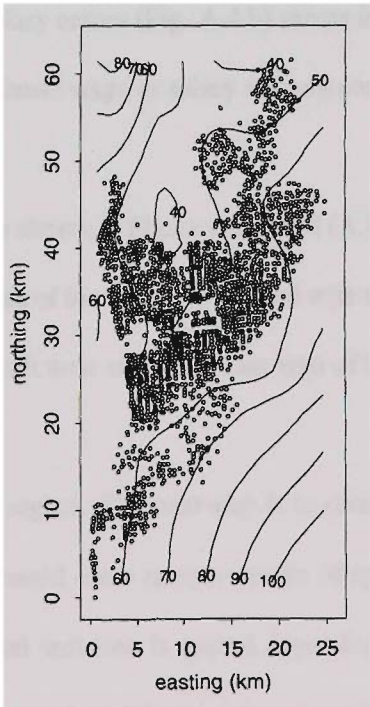


(a) 3-d scatter plot



employment rate

(b) Surface plot



(c) Contour plot

Figure 8.2. Scatter plot, contour and surface plot of the employment rate

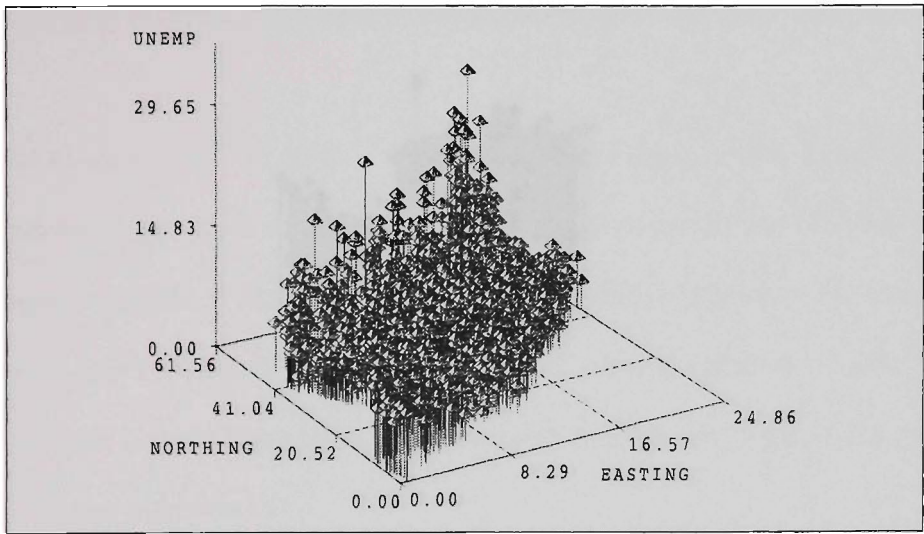
The scatter plot of the employment rate (8.2 a) shows a high employment rate in the east of the region. But this perception can be deceptive, since some points can be hidden by the tall points. A better view can be obtained by looking at the surface plot on figure (8.2 b). It shows that the minimum employment rate around the middle of the region and high at the edge of the region. The contour plot (8.2 c) shows that the level of the employment rate in the middle of the region is about at 40 %, and getting high into 60 % and 70 % at the edge of the region.

The scatter plot of the unemployment rate (fig. 8.3 a) shows a high rate in the north east area, and also the surface plot (8.3 b) shows that the high unemployment rates are mostly in the north area of the region, starting from 10 % and decreasing to the lower east area. Meanwhile the labor participation rate is shown in figure (8.4). The shape of the surface (8.4 b) is similar to the surface of the employment rate. This is because employment is the main component of the labor force participation rate.

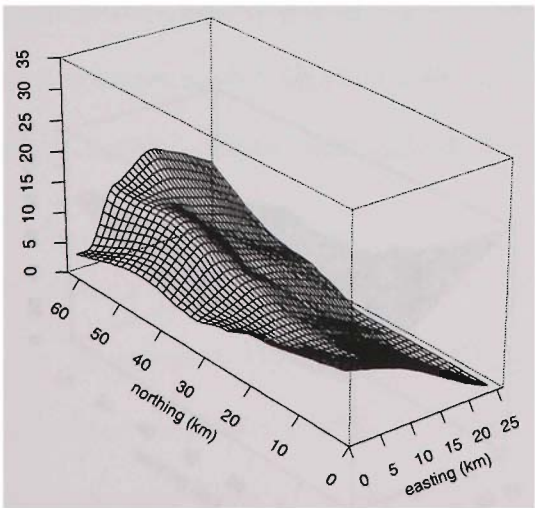
The employment characteristics are broken down into three aspects, in terms of wage or salary earner, self employed, and employer. The surface plot of the wage or salary earner (Fig. A.4 b) seems identical to the surface plot of the employment rate (Fig. 8.2 b). This is because wage or salary earners comprise the major portion of the employment rate.

Scatter plots of the formal and informal qualification rate are shown in Figure (A.7) and (A.8), respectively. The formal qualification rate is lower in the north west area of the region compared with the middle to the east area. The informal qualification rate is high in the north west and south east area of the region, and low in the middle area.

These figures show how characteristics fluctuate across the region. The next step is to consider measurement of spatial variation. Deutsch and Journel (1992) discussed some measurements of spatial variation, one of which is semivariogram. Another aspect of spatial variation is spatial dependency, which indicates how one observation is affected by another observation when it is related to the geographical distance between them. This aspect is called as spatial autocorrelation (Anselin, 1988). The relationship between semivariogram and spatial autocorrelation was discussed in section (5.1.5).

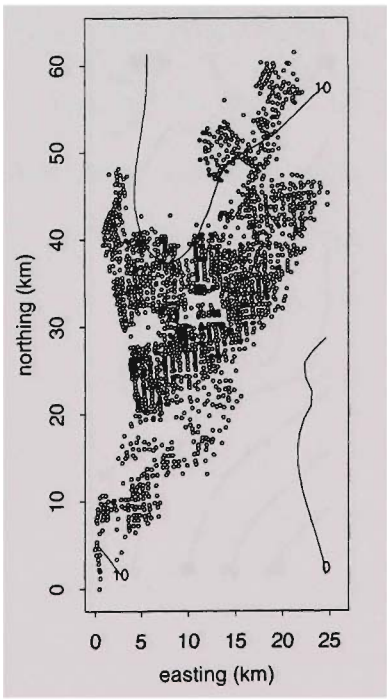


(a) 3-d scatter plot



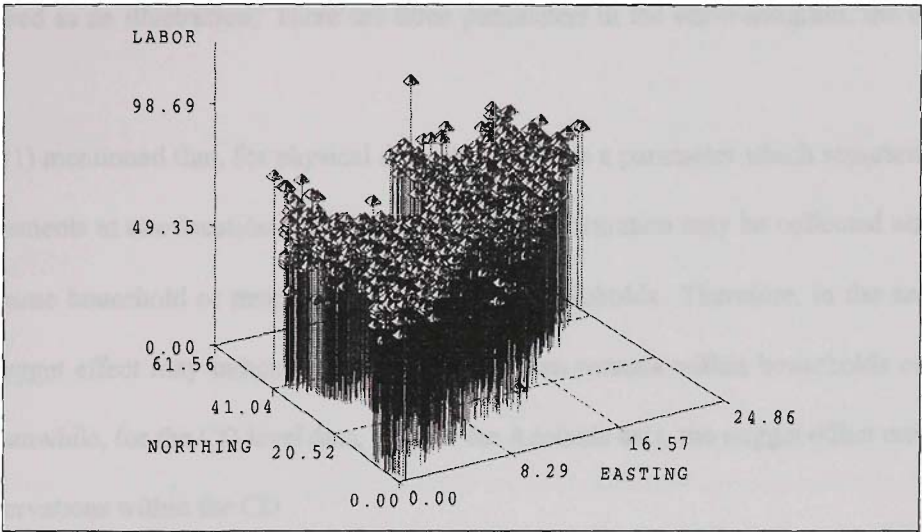
unemployment rate

(b) Surface plot

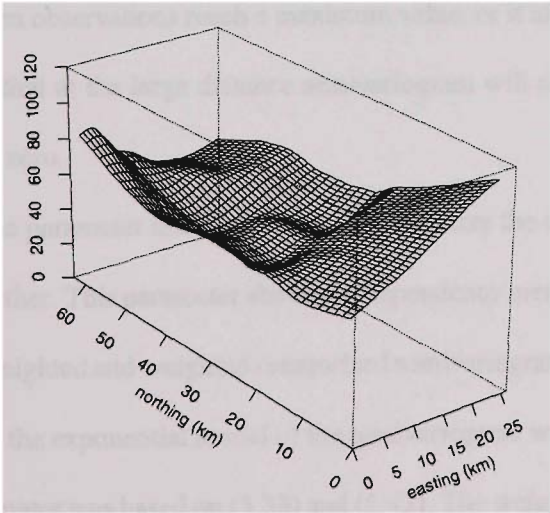


(c) Contour plot

Figure 8.3. Scatter plot, contour and surface plot of the unemployment rate

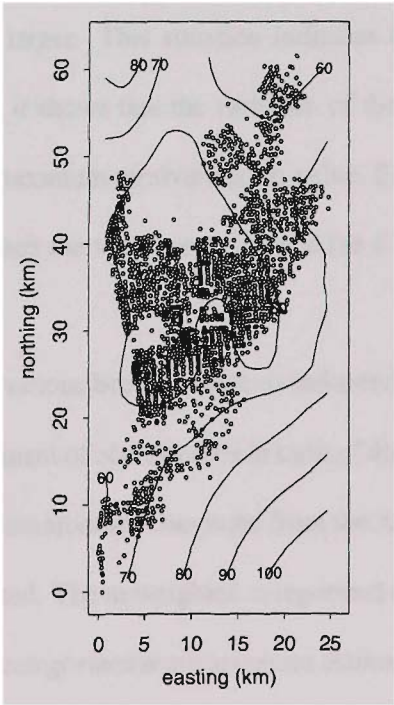


(a) 3-d scatter plot



labor part. rate

(b) Surface plot



(c) Contour plot

Figure 8.4. Scatter plot, contour and surface plot of the labor participation rate

8.4 Semivariogram analysis of the Adelaide data

The discussion here is focused on developing a semivariogram model. The exponential semivariogram model will be used as an illustration. There are three parameters in the semivariogram, the nugget, sill, and range.

Cressie (1991) mentioned that, for physical data, the nugget is a parameter which represents the variation of measurements at one location point. In social data, information may be collected about several peoples in the same household or peoples at the groups of households. Therefore, in the unit level social data, the nugget effect may indicate the variation between persons within households or groups of households. Meanwhile, for the CD level data, such as the Adelaide data, the nugget effect may represent variation of observations within the CD.

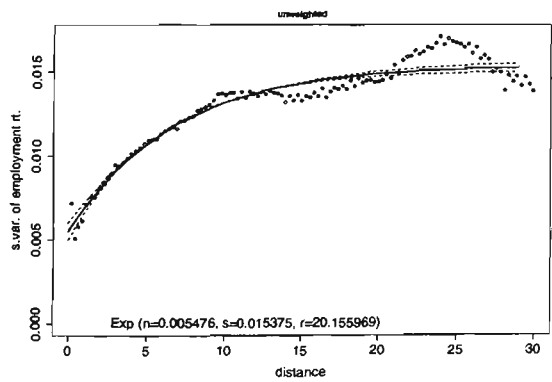
The second parameter of the semivariogram model is the sill. Figure (5.1) shows that the sill is the asymptotic line of the semivariogram plot when the distance is large. The observations may tend to be independent of each other as the distances between them gets larger. This situation indicates that the covariogram has reached a minimum value. On the other hand, it shows that the variation of the differences between observations reach a maximum value, or it has a maximum semivariogram value. Equation (5.8) shows that at the large distance semivariogram will approach the variance, σ^2 , since the $C(d_{ij})$ is approaching zero.

The range parameter indicates the distances where the observations become close to independent between each other. This parameter shows the dependency measurement of observations in term of distances.

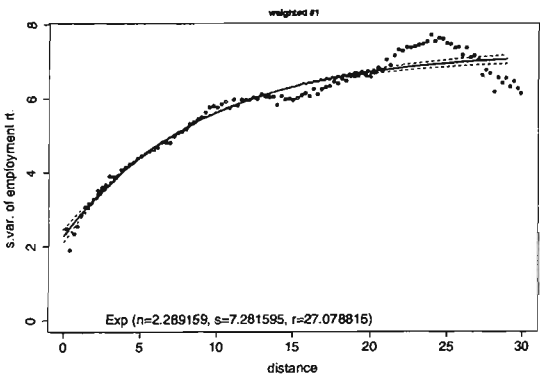
The unweighted and weighted categorized semivariograms estimators are computed from the Adelaide CD data and the exponential model of the semivariogram was fitted. The unweighted categorized semivariogram estimator was based on (5.33) and (5.42). The weighted categorized semivariogram estimator was based on (5.114), (5.122), and (5.42). The weighted least squares method (Cressie, 1985) is applied to fit the empirical semivariogram into the model (see section 5.1.3). The SAS procedure is used to perform this task (see appendix F). The results are shown partly in table (8.3), (8.4), (8.5), and figure (8.5), (8.6), (8.7). Other variables are shown in appendix (A).

Table 8.3. Estimated parameters of the exponential semivariogram model of the employment rate

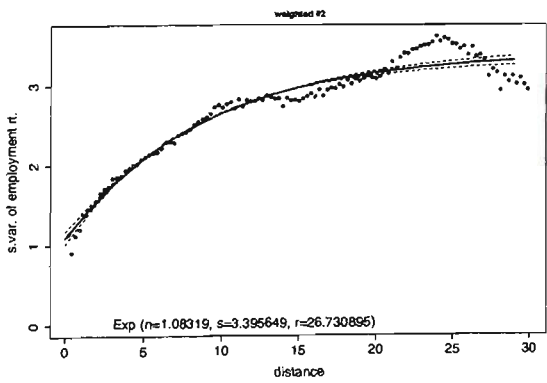
| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.005476 | 0.000249 | 0.004983 | 0.005969 |
| | Sill | 0.015375 | 0.000157 | 0.015064 | 0.015685 |
| | Range | 20.155969 | 1.177222 | 17.823439 | 22.488500 |
| weighted #1 | Nugget | 2.289159 | 0.085139 | 2.120466 | 2.457852 |
| | Sill | 7.281595 | 0.094774 | 7.093811 | 7.469379 |
| | Range | 27.078815 | 1.474798 | 24.156670 | 30.000960 |
| weighted #2 | Nugget | 1.083190 | 0.040079 | 1.003777 | 1.162603 |
| | Sill | 3.395649 | 0.043329 | 3.309797 | 3.481500 |
| | Range | 26.730895 | 1.453698 | 23.850558 | 29.611231 |



(a) Unweighted



(b) Weighting factor 1



(c) Weighting factor 2

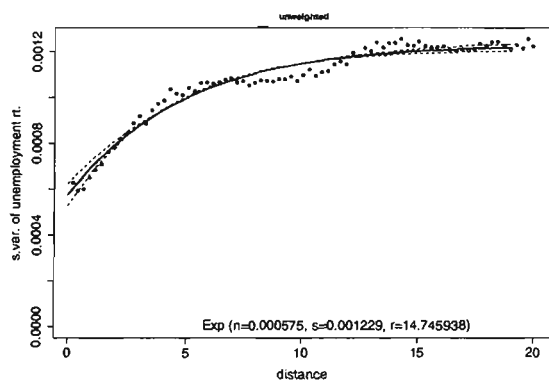
Figure 8.5. Semivariogram model fitting for the employment rate

Figure (8.5a) shows the unweighted semivariogram and the weighted semivariograms of the employment rate. The circle points represent the empirical or calculated semivariogram, and the solid line represent the fitted exponential semivariogram model. The sill of the unweighted semivariogram is 0.015375, which is close with the unweighted group level variance (0.0134, see table 8.2). The sill of the weighted #1 (7.28) differs from the weighted group level variance (5.87, table 8.2). Figure (8.5-b) shows that it may be caused by variation at the distance larger than 15 km. Below the distance 15 km the sill reach the value 6. The nugget is not zero for all cases of unweighted, weighted #1, or weighted #2, which indicates variations within the CD. The range differs significantly between unweighted and weighted, since the estimate of the unweighted range (20.16) is outside the 95% confidence interval of either the weighted #1 (24.16;30.00) or the weighted #2 (23.85;29.61). On the other hand, the estimated range does not differ significantly between weighted #1 (27.08) and weighted #2 (26.73).

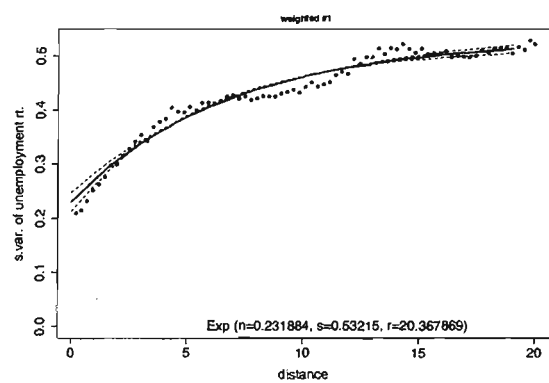
Table 8.4. Estimated parameters of the exponential semivariogram model of the unemployment rate

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000575 | 0.000024 | 0.000527 | 0.000622 |
| | Sill | 0.001229 | 0.000011 | 0.001207 | 0.001251 |
| | Range | 14.745938 | 1.067368 | 12.619154 | 16.872721 |
| weighted #1 | Nugget | 0.231884 | 0.008767 | 0.214414 | 0.249353 |
| | Sill | 0.532150 | 0.007715 | 0.516778 | 0.547523 |
| | Range | 20.367869 | 1.652878 | 17.074428 | 23.661310 |
| weighted #2 | Nugget | 0.111510 | 0.004190 | 0.103161 | 0.119859 |
| | Sill | 0.260971 | 0.003905 | 0.253189 | 0.268753 |
| | Range | 20.892583 | 1.682942 | 17.539239 | 24.245927 |

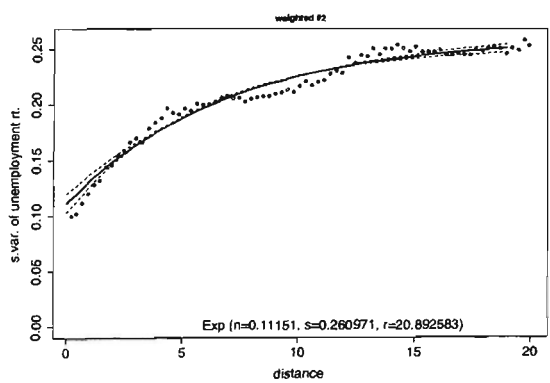
Table (8.4) and figure (8.6) show semivariogram analysis of the unemployment rate. The exponential semivariogram model of the unemployment rate has estimated nugget, sill, and range 0.000575, 0.001229, and 14.745938, respectively. Figure (8.3) shows a trend of spatial location from south-east (SE) to north-west (NW). This trend is shown on the semivariogram graph (figure 8.6), which has two levels. The first one is start from zero to around 10 kms, and the second is start from 10 to 20 kms. If the cut-off is 10 kms, then the estimated range might be reduced to around 5 kms. The anisotropic condition cannot be indicated



(a) Unweighted



(b) Weighting factor 1



(c) Weighting factor 2

Figure 8.6. Exponential semivariogram model fitting for the unemployment rate

only by this omnidirectional semivariogram, but it may be shown by comparison of the range from four or eight types of directional semivariogram (see Carr, 1995, page 170).

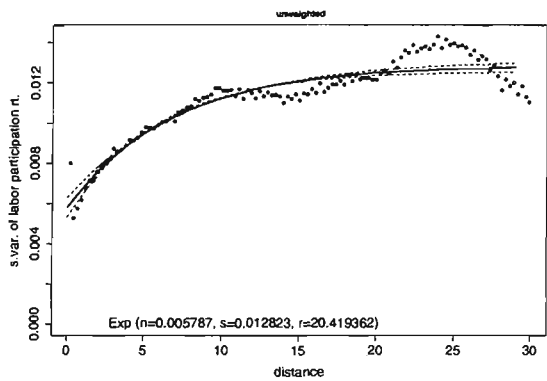
The unweighted sill is close with the unweighted variance (0.0013, see table 8.2). The weighted sill (0.53) is also similar with the weighted group level variance (0.53). The estimated range differs significantly between the unweighted (14.75) and weighted (20.37 and 20.89), but does not differ much between the weighted #1 (20.37) and weighted #2 (20.89).

Table 8.5. Estimated parameters of the exponential semivariogram model of the labor participation rate

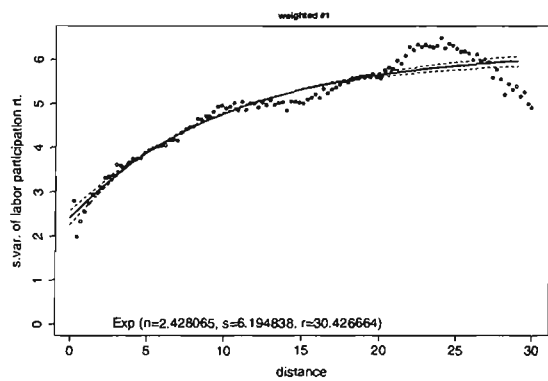
| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.005787 | 0.000244 | 0.005304 | 0.006271 |
| | Sill | 0.012823 | 0.000142 | 0.012543 | 0.013104 |
| | Range | 20.419362 | 1.558532 | 17.331308 | 23.507415 |
| weighted #1 | Nugget | 2.428065 | 0.082243 | 2.265109 | 2.591021 |
| | Sill | 6.194838 | 0.103469 | 5.989826 | 6.399850 |
| | Range | 30.426664 | 2.264362 | 25.940089 | 34.913239 |
| weighted #2 | Nugget | 1.138438 | 0.037570 | 1.063997 | 1.212880 |
| | Sill | 2.907754 | 0.049328 | 2.810017 | 3.005492 |
| | Range | 31.128410 | 2.312467 | 26.546520 | 35.710300 |

Table (8.5) and figure (8.7) show semivariogram analysis of the labor force participation rate. The unweighted semivariogram of the labor force participation rate is similar to that of the employment rate, since the employment rate is the largest portion. The estimated nugget, sill, and range of the unweighted semivariogram are 0.00578749, 0.01282333, and 20.419361, respectively. The estimate of weighted sill (6.19) differs with the weighted group level variance in table (8.2) (4.90). Again, the shifted weighted sill from the weighted group level variance can be caused by the variation at the distance 15 km and more (see figure 8.7-b).

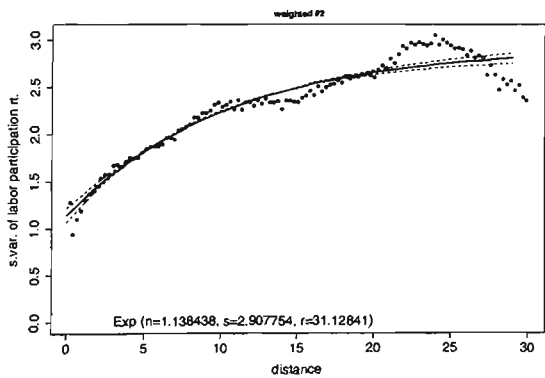
For the other variables, the result of semivariogram analysis is presented in section (A.2) of appendix (A). The interpretation are quite similar with the previous variables, hence we will summarize the results in general.



(a) Unweighted



(b) Weighting factor 1



(c) Weighting factor 2

Figure 8.7. Exponential semivariogram model fitting for the labor participation rate

8.4.1 Summary

To summarize the result, some statistics are tabulated in Tables (8.6), (8.7), (8.8), and (8.9). Additional statistics are computed, they are spatial correlation at distance 0 and spatial correlation at distance equal to the range (r). For example we can calculate the unweighted nugget effect within the group, for employment rate, in term of $\rho(0)$ by applying theorem (5.1.5), that is

$$\rho(0) = 1 - \frac{0.005476}{0.015375} = 0.643837$$

This value is equal to the spatial correlation at distance 0. The distance here has the meaning of the distance between groups. Meanwhile by using equation (5.21), we can also evaluate the spatial correlation at distance equal to the range (r) for the employment rate, that is

$$\rho(r) = \left(1 - \frac{0.005476}{0.015375}\right) \exp(-3) = 0.0321$$

The weighted version can also be computed in the same way.

Tables (8.6) shows a summary of the semivariogram analysis of the labor force characteristics. The unweighted estimate of the sill is similar with the unweighted group level variance for the employment rate, unemployment rate, and labor force participation rate. The difference occurs in the weighted #1 sill in the employment rate and labor force participation rate with the $N\bar{S}_{yy}$. Figure (8.5) and (8.7) show a possible cause of this difference, which is the variation at distance 15 km and more. The estimated range of the unweighted semivariogram of the employment rate is 20.16 kms (see table 8.3) indicates the distance (km) where the observations become almost independent each other. Table (8.6) shows empirically that spatial correlation is decaying from 0.6438 to 0.0321 at the distance between groups from 0 to 20.16 km.

Table (8.7) shows a summary of semivariogram analysis of the income characteristics. The unweighted sill of the variables are similar to the unweighted group level variance, and also for the weighted #1 sill are similar to the $N\bar{S}_{yy}$. The rate of income below 20,000 and income over 40,000 have a smaller range than the range of rate of income 20,000-40,000. The unweighted or weighted nugget are not zero for all variables. The rate of income over 40,000 has small nugget compared with its sill, which implies high correlation at distance 0.

Table 8.6. Summary of labor force characteristics

| Characteristic | | Unweighted | weighted | |
|--------------------------|-------------------|------------|-----------|-----------|
| | | | #1 | #2 |
| Employment rate | \bar{S}_{yy} | 0.0134 | 5.8739 | |
| | \hat{n} | 0.005476 | 2.289159 | 1.08319 |
| | \hat{s} | 0.015375 | 7.281595 | 3.395649 |
| | \hat{r} | 20.155969 | 27.078815 | 26.730895 |
| | \hat{n}/\hat{s} | 0.3562 | 0.3144 | 0.3190 |
| | $\rho(0)$ | 0.6438 | 0.6856 | 0.6810 |
| | $\rho(r)$ | 0.0321 | 0.0341 | 0.0339 |
| Unemployment rate | \bar{S}_{yy} | 0.0013 | 0.5321 | |
| | \hat{n} | 0.000575 | 0.231884 | 0.11151 |
| | \hat{s} | 0.001229 | 0.53245 | 0.260971 |
| | \hat{r} | 14.745938 | 20.367869 | 20.892583 |
| | \hat{n}/\hat{s} | 0.4679 | 0.4355 | 0.4273 |
| | $\rho(0)$ | 0.5321 | 0.5645 | 0.5727 |
| | $\rho(r)$ | 0.0265 | 0.0281 | 0.0285 |
| Labor participation rate | \bar{S}_{yy} | 0.0112 | 4.9085 | |
| | \hat{n} | 0.005787 | 2.428065 | 1.138438 |
| | \hat{s} | 0.012823 | 6.194838 | 2.907754 |
| | \hat{r} | 20.419362 | 30.426664 | 31.12841 |
| | \hat{n}/\hat{s} | 0.4513 | 0.3919 | 0.3915 |
| | $\rho(0)$ | 0.5487 | 0.6081 | 0.6085 |
| | $\rho(r)$ | 0.0273 | 0.0303 | 0.0303 |

Table 8.7. Summary of the income characteristics

| Characteristic | | Unweighted | weighted | |
|------------------------------|-------------------|------------|-----------|-----------|
| | | | #1 | #2 |
| Rate of income below 20,000 | \bar{S}_{yy} | 0.0082 | 3.4513 | |
| | \hat{n} | 0.001792 | 0.687636 | 0.334117 |
| | \hat{s} | 0.008559 | 3.646971 | 1.738171 |
| | \hat{r} | 6.924523 | 8.558424 | 8.941471 |
| | \hat{n}/\hat{s} | 0.2094 | 0.1885 | 0.1922 |
| | $\rho(0)$ | 0.7906 | 0.8115 | 0.8078 |
| | $\rho(r)$ | 0.0394 | 0.0404 | 0.0402 |
| Rate of income 20,000-40,000 | \bar{S}_{yy} | 0.0044 | 1.8344 | |
| | \hat{n} | 0.001914 | 0.772853 | 0.369552 |
| | \hat{s} | 0.004547 | 2.224875 | 1.010783 |
| | \hat{r} | 15.155897 | 30.270201 | 26.196292 |
| | \hat{n}/\hat{s} | 0.4209 | 0.3474 | 0.3656 |
| | $\rho(0)$ | 0.5791 | 0.6526 | 0.6344 |
| | $\rho(r)$ | 0.0288 | 0.0325 | 0.0316 |
| Rate of income over 40,000 | \bar{S}_{yy} | 0.0026 | 1.4366 | |
| | \hat{n} | 0.00017 | 0.033202 | 0.01762 |
| | \hat{s} | 0.002801 | 1.53447 | 0.652238 |
| | \hat{r} | 5.294999 | 4.565517 | 6.03821 |
| | \hat{n}/\hat{s} | 0.0607 | 0.0216 | 0.0270 |
| | $\rho(0)$ | 0.9393 | 0.9784 | 0.9730 |
| | $\rho(r)$ | 0.0468 | 0.0487 | 0.0484 |

Table 8.8. Summary of the nature of the employment characteristics

| Characteristic | | Unweighted | weighted | |
|-------------------------------|-------------------|------------|-----------|-----------|
| | | | #1 | #2 |
| Rate of wage or salary earner | \bar{S}_{yy} | 0.0100 | 4.3664 | |
| | \hat{n} | 0.004797 | 1.804646 | 0.862768 |
| | \hat{s} | 0.013138 | 6.428906 | 2.97899 |
| | \hat{r} | 36.054854 | 45.175892 | 44.41955 |
| | \hat{n}/\hat{s} | 0.3651 | 0.2807 | 0.2896 |
| | $\rho(0)$ | 0.6349 | 0.7193 | 0.7104 |
| | $\rho(r)$ | 0.0316 | 0.0358 | 0.0354 |
| Rate of self employed | \bar{S}_{yy} | 0.0003 | 0.1198 | |
| | \hat{n} | 0.000186 | 0.075139 | 0.036784 |
| | \hat{s} | 0.0003 | 0.132648 | 0.06432 |
| | \hat{r} | 16.687985 | 25.843623 | 27.724176 |
| | \hat{n}/\hat{s} | 0.6200 | 0.5665 | 0.5719 |
| | $\rho(0)$ | 0.3800 | 0.4335 | 0.4281 |
| | $\rho(r)$ | 0.0189 | 0.0216 | 0.0213 |
| Rate of employer | \bar{S}_{yy} | 0.0006 | 0.2252 | |
| | \hat{n} | 0.000117 | 0.052682 | 0.028307 |
| | \hat{s} | 0.00059 | 0.241481 | 0.118443 |
| | \hat{r} | 6.178353 | 7.480126 | 8.213822 |
| | \hat{n}/\hat{s} | 0.1983 | 0.2182 | 0.2390 |
| | $\rho(0)$ | 0.8017 | 0.7818 | 0.7610 |
| | $\rho(r)$ | 0.0399 | 0.0389 | 0.0379 |

Table 8.9. Summary of the qualification achievement characteristics

| Characteristic | | Unweighted | weighted | |
|-----------------------------|-------------------|------------|-----------|-----------|
| | | | #1 | #2 |
| Formal qualification rate | \bar{S}_{yy} | 0.0073 | 3.1227 | |
| | \hat{n} | 0.000248 | 0.106191 | 0.043951 |
| | \hat{s} | 0.008548 | 3.756959 | 1.825394 |
| | \hat{r} | 13.364826 | 15.23778 | 16.121832 |
| | \hat{n}/\hat{s} | 0.0290 | 0.0283 | 0.0241 |
| | $\rho(0)$ | 0.9710 | 0.9717 | 0.9759 |
| | $\rho(r)$ | 0.0483 | 0.0484 | 0.0486 |
| Informal qualification rate | \bar{S}_{yy} | 0.0015 | 0.6590 | |
| | \hat{n} | 0.000549 | 0.221561 | 0.103544 |
| | \hat{s} | 0.002054 | 0.985302 | 0.456666 |
| | \hat{r} | 34.968376 | 42.302545 | 40.463875 |
| | \hat{n}/\hat{s} | 0.2673 | 0.2249 | 0.2267 |
| | $\rho(0)$ | 0.7327 | 0.7751 | 0.7733 |
| | $\rho(r)$ | 0.0365 | 0.0386 | 0.0385 |

The spatial correlation, either $\rho(0)$ or $\rho(r)$, shows a similar value between the unweighted or weighted version. It indicates that the spatial correlation is not affected much by the weighting factors.

The estimated parameters seems to be affected by the weighting factors, but they do not differ significantly between weighted #1 and weighted #2. The estimated range is much appreciated from the unweighted to the weighted, but does not differ between weighted #1 and #2. Further results for the remaining variables are given in Tables (8.8) and (8.9). The estimated unweighted sill is similar with the ${}_1\bar{S}_{yy}$ for all variables, but some of the estimated of weighted #1 sill are similar with the ${}_N\bar{S}_{yy}$. The difference occurred in the employment rate, labor force participation rate, rate of wage or salary earner, and informal qualification rate. It may be caused by the variation at distance of 15 km or more (see figure 8.5, 8.7, A.12, and A.16). Tables (8.6) through (8.9) indicate that high nugget compared with sill imply a lower spatial correlation at distance zero or r .

8.5 The micro sample of the Adelaide data

The previous section considered group level data, which is an aggregation of the census data from the individual level to the CD level. In this section, a micro sample of census data will be introduced. It can play a role in the adjustment of the group level analysis or it can be used as a benchmark for comparison of the proposed methods of analysis.

The micro sample contains a sample of approximately one percent of the Adelaide population. No geographical location information is recorded, therefore we can not analyse the individual level data with spatial analysis methods. However, individual level data can play a role in spatial analysis when used in combination with spatial group level data.

The tabulation of the characteristics from the micro sample of Adelaide is given in table (A.9). The total number of individuals in the sample is 10,210 with 8,196 aged 15 years and more. Table (A.9) shows a frequency and percentage of the characteristics based on the total number of individuals in the sample. Table (8.10) shows the rate of the characteristics which are defined to be a percentage of the frequency over the total number of individuals aged 15 and more in the sample.

Table 8.10. The rate and variance of the characteristics from the micro sample of Adelaide

| Characteristic | Micro sample | | Census | |
|------------------------|------------------------|------------------------------|-------------|----------------------|
| | Rate \hat{p}_{ms} | Variance \hat{s}_{ms}^2 | Rate P | Variance S_{yy} |
| Employment | 0.5456 | 0.2479 | 0.5363 | 0.2487 |
| Unemployment | 0.0722 | 0.0670 | 0.0718 | 0.0666 |
| Labor participation | 0.6179 | 0.2361 | 0.6081 | 0.2383 |
| Income < 20000 | 0.6057 | 0.2388 | 0.6105 | 0.2378 |
| Income 20000-40000 | 0.2611 | 0.1929 | 0.2617 | 0.1932 |
| Income 40000 over | 0.0569 | 0.0536 | 0.0583 | 0.0549 |
| wage or salary earner | 0.4705 | 0.2491 | 0.4620 | 0.2486 |
| self employed person | 0.0458 | 0.0410 | 0.0430 | 0.0411 |
| employer | 0.0288 | 0.0280 | 0.0285 | 0.0277 |
| Formal qualification | 0.1214 | 0.1067 | 0.1250 | 0.1094 |
| Informal qualification | 0.1342 | 0.1162 | 0.1355 | 0.1171 |

The estimated variance, \hat{s}_{ms}^2 , from the micro sample can be used as the initial value of the sill, in estimating the parameter n , s , and r as described in section (5.4). We consider its use in this estimation process in the next section.

8.6 Estimation of individual level semivariogram parameters by non-linear model

This section applies the methods discussed in section (5.4) to estimate individual level semivariogram parameters of the employment rate. Three different groupings are considered, the CD level, SSC level, and DPC level (see section 7.3.2 for grouping factor detail). There is 1713 groups at CD level, 313 group at SSC level, and 102 groups at DPC level.

The non-linear regression model is applied for these three different grouping factors by calculating the Γ_{gh}^f , where $f = \{1, 2, 3\}$, Γ_{gh}^1 for the CD level, Γ_{gh}^2 for the SSC level, and Γ_{gh}^3 for the DPC level. The working model is defined by

$$\begin{aligned} \Gamma_{gh}^f = & s \left(\frac{1}{2N_g^f} - \frac{1}{2N_h^f} \right) - (s - n) \cdot \exp \left[\frac{-3d_{gh}^f}{r} \right] \\ & + \frac{N_g^f - 1}{2N_g^f} (s - n) \cdot \exp \left[\frac{-3k_1 \sqrt{A_g^f}}{r} \right] + \frac{N_h^f - 1}{2N_h^f} (s - n) \cdot \exp \left[\frac{-3k_1 \sqrt{A_h^f}}{r} \right] \end{aligned} \tag{8.8}$$

Model (8.8) suggests that parameter n , s , and r can be estimated no matter what the grouping factor level is, i.e. estimating parameters using the CD level may give the same results as estimating parameters using the SSC level.

Estimation of the parameters can be considered in two situations, those are when the individual data are available, and individual data are not available. The first situation implies that individual level variance can be computed, which provides the initial value of the sill. In this situation, two methods are applied, those are method (1a) when all parameters are estimated, and method (1b) where the sill is fixed at individual level variance, and n and r are estimated. For the both methods, the initial value of the nugget is zero (see section 5.4.3), and the initial value of the range is determined by (5.110). The second situation implies that only the group level variance (unweighted and weighted) can be calculated. This resulting method (2) defines the initial value of the nugget, sill, and range as equal to the estimated group level parameters from semivariogram analysis of group level data (see section 5.4.1). The estimation result is shown in table (8.11).

Table (8.3) shows the estimated nugget, sill, and range for the unweighted group level at the CD level, which are 0.005476, 0.015375, and 20.155969, respectively. The corresponding estimated parameter in Table (8.11) can be compared, which show that the non-linear regression method has adjusted the estimated of unweighted group level parameters at the CD level, except for the estimated range. This is also a case when the estimation was done at the SSC level and DPC level. Except for the SSC level, the estimated range becomes larger than the other level.

In each level of data, the approach (1a) and (2) (see section 5.4.3) give almost identical estimates of nugget and sill, but it is quite different with approach (1b). However the estimates for three approaches differ depending on the level of data used in the analysis. In particular the use of DPC level data gives quite different results from using CD or SSC level data. The results obtained using approach (1b), where the sill is equated with the individual level variance obtained from the micro-sample, are more consistent across the different levels, although the range estimate varies considerably. They also differ considerably from the estimate obtained from approach (1a) and (2).

Table 8.11. Estimated n , s , and r of the employment rate semivariogram model by non-linear model at different grouping factor levels, when individual and group level variance is known

| Grouping factors | approach | Estimated parameters | | |
|------------------|----------|----------------------|--------------------|---------------------|
| | | nugget (\hat{n}) | sill (\hat{s}) | range (\hat{r}) |
| CD level | 1a | 1.6021 | 1.6139 | 20.9832 |
| | (se) | (0.0054) | (0.0054) | (0.2075) |
| | 1b | 0.2335 | 0.2487 | 13.7282 |
| | (se) | (0.00002) | – | (0.10736) |
| | 2 | 1.6021 | 1.6139 | 20.9835 |
| | (se) | (0.0054) | (0.0054) | (0.2075) |
| SSC level | 1a | 1.2034 | 1.2164 | 35.0045 |
| | (se) | (0.0801) | (0.0800) | (1.3556) |
| | 1b | 0.2353 | 0.2487 | 30.0001 |
| | (se) | (0.0001) | – | (0.9551) |
| | 2 | 1.2034 | 1.2165 | 35.0056 |
| | (se) | (0.0801) | (0.0801) | (1.3554) |
| DPC level | 1a | 9.8619 | 9.8699 | 20.0833 |
| | (se) | (0.5285) | (0.5283) | (3.0162) |
| | 1b | 0.2370 | 0.2487 | 18.8854 |
| | (se) | (0.0002) | – | (1.9934) |
| | 2 | 9.8585 | 9.8665 | 19.9528 |
| | (se) | (0.5285) | (0.5283) | (2.9978) |

Note : se=standard error of the estimated

The CD and SSC level give almost the same estimated of the nugget and sill, but they are different with the estimated at the DPC level. This situation suggested that the scale, such as $\frac{N}{M}$, still have an effect into the estimation of the nugget and sill. But the estimation of the range may not be much affected by the scale.

8.7 Cross-semivariogram analysis

The bivariate cases will be considered examining the association between formal qualification rate and the rate of income below 20,000 and income over 40,000 at the CD level data. The main objective of this exercise is to illustrate the effectiveness of cross-semivariogram analysis in exploring the relationship structure between characteristics as displayed in diagram (6.1). Firstly we may look at the correlation coefficient from non-spatial perspective as in table (8.12), which gives the correlation estimated from CD level data.

Table 8.12. Non-spatial correlation coefficient of employment rate versus its component at the CD level data

| | income below 20,000 | income over 40,000 |
|---------------------------|---------------------|--------------------|
| Formal qualification rate | -0.5962 | 0.7223 |

The exponential cross-semivariogram was estimated using CD level data and the unweighted approach and the two weighting described in section (6.7). Results are given in table (8.13) and (8.14). Figures (8.8) and (8.9) show the empirical and fitted cross-semivariogram. These show some interesting behaviour for distances exceeding 10-15 kms. The figures show a wave pattern beyond the distance 10-15 kms along through the distance 30 km. This situation suggest that the constant mean assumptions may be violated. This may be resolved by applying an adjustment to remove a trend (detrended) within the data (Cressie, 1991).

dependency among points are reduced. It can be understood that

We can compute spatial correlation ($\rho_{ab}(d_{ij})$) and coefficient of codispersion ($\rho_{ab}^{\gamma}(d_{ij})$) at distance 0 and r_{ab} by applying equation (6.53) and (6.55). It is noted that σ_{ab} is equal to the estimated sill of the

Table 8.13. Estimated parameters of the exponential cross-semivariogram model of the formal qualification rate and rate of income below 20,000 at CD level

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000134 | 0.000356 | -0.00057374 | 0.000841 |
| | Sill | -0.004883 | 0.000083 | -0.005048 | -0.004719 |
| | Range | 6.066235 | 0.832444 | 4.403389 | 7.719080 |
| weighted #1 | Nugget | 0.056572 | 0.123439 | -0.188519 | 0.301664 |
| | Sill | -2.028022 | 0.031722 | -2.091008 | -1.965035 |
| | Range | 6.409151 | 0.761712 | 4.896745 | 7.921556 |
| weighted #2 | Nugget | 0.022364 | 0.053280 | -0.083425 | 0.128152 |
| | Sill | -1.002786 | 0.014814 | -1.032220 | -0.973371 |
| | Range | 6.816045 | 0.726416 | 5.373722 | 8.258370 |

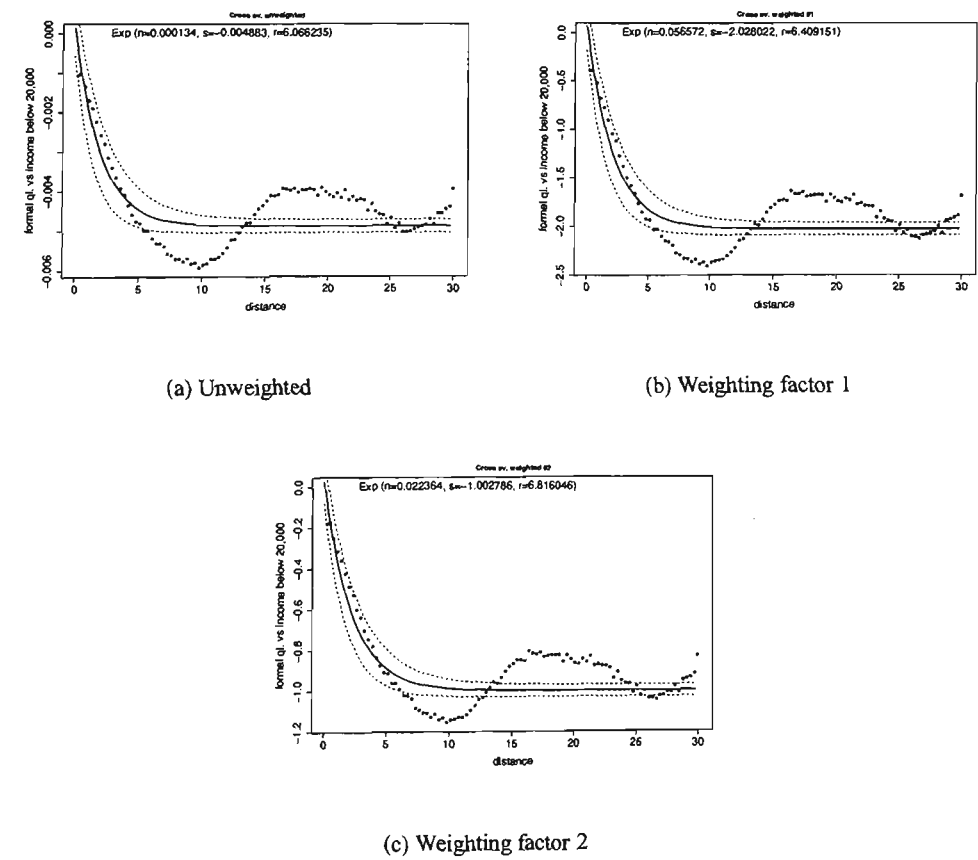
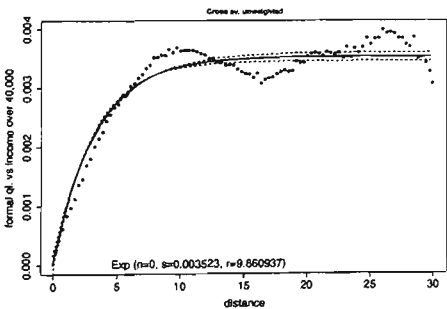


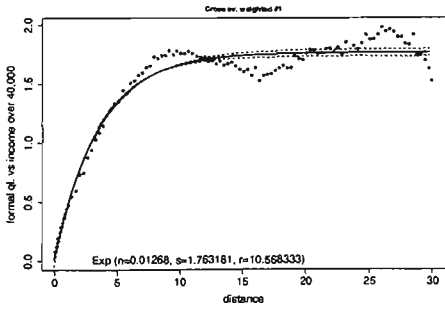
Figure 8.8. Cross-semivariogram model fitting for the formal qualification rate with the rate of income below 20,000

Table 8.14. Estimated parameters of the exponential cross-semivariogram model of the formal qualification rate and rate of income over 40,000

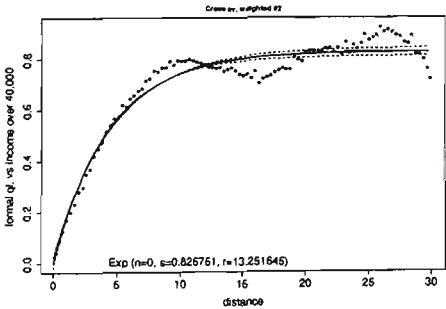
| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.0 | 0.000078 | -0.000155 | 0.000155 |
| | Sill | 0.003523 | 0.000036 | 0.003451 | 0.003595 |
| | Range | 9.860937 | 0.532701 | 8.803242 | 10.918635 |
| weighted #1 | Nugget | 0.012680 | 0.030474 | -0.047828 | 0.073187 |
| | Sill | 1.763181 | 0.015283 | 1.732835 | 1.793527 |
| | Range | 10.568333 | 0.457421 | 9.660108 | 11.476558 |
| weighted #2 | Nugget | 0.0 | 0.010557 | -0.020961 | 0.020961 |
| | Sill | 0.826751 | 0.008216 | 0.810439 | 0.843063 |
| | Range | 13.251645 | 0.521647 | 12.215898 | 14.287392 |



(a) Unweighted



(b) Weighting factor 1



(c) Weighting factor 2

Figure 8.9. Cross-semivariogram model fitting for the formal qualification rate with the rate of income over 40,000

cross-semivariogram, and σ_a^2 or σ_b^2 are the estimated sill of the respective semivariograms. Meanwhile we can define $C_{ab}(r_{ab})$ by using relationship in (6.50) and applying equation (5.21). The results are shown in table (8.15). The results in table (8.15) shows that the values of $\rho_{ab}(0)$ are very close to the correlation coefficient as shown in table (8.12), and also the coefficient of codispersion at the distance equal to the estimated range. The weighting factor on the cross-semivariogram gave a small effect on either the spatial correlation or coefficient of codispersion.

Table 8.15. Spatial correlation and codispersion coefficient at distance 0 and r_{ab} .

| Formal qualification rate vs | $\rho_{ab}(0)$ | $\rho_{ab}(r_{ab})$ | $\rho_{ab}^\gamma(0)$ | $\rho_{ab}^\gamma(r_{ab})$ |
|------------------------------|----------------|---------------------|-----------------------|----------------------------|
| income below 20,000 | | | | |
| unweighted | -0.5709 | -0.6694 | 0.2004 | -0.5665 |
| weighted #1 | -0.5479 | -0.6342 | 0.2094 | -0.5440 |
| weighted #2 | -0.5630 | -0.6482 | 0.1845 | -0.5591 |
| income over 40,000 | | | | |
| unweighted | 0.7200 | 0.7539 | 0 | 0.7184 |
| weighted #1 | 0.7343 | 0.7477 | 0.2135 | 0.7337 |
| weighted #2 | 0.7577 | 0.7776 | 0 | 0.7567 |

8.7.1 Estimation of individual level cross-semivariogram parameter using non-linear regression methods

Again consider bivariate cases of formal qualification rate with the rate of income below 20,000 and rate of income over 40,000. The group level data are defined by three different grouping factors, CD level, SSC level, and DPC level. The non-spatial correlation for each level are shown in table (8.16). The absolute value of the correlations increase from CD level to SSC level. When moving from the SSC to DPC level there is a very small increase for income over 40,000 and a decrease in absolute for income below 20,000.

The estimation method using non-linear regression was discussed in section (6.8), and is based on the non-categorized version of the group level cross-semivariogram. Consider the exponential cross-semivariogram where equation (6.92) is applied. Consider two situations, (1) when the micro sample data are also available, and (2) when only group level data are available. In the first situation, the variances can be calculated for each of the variables involved. This information from the micro sample is useful as the initial value of the sill. The initial value of the nugget is zero, and the initial value of the range is r_{ab}

Table 8.16. Non-spatial correlation coefficient of employment rate versus its component

| | grouping factor | income < 20,000 | income > 40,000 |
|------------------|-----------------|-----------------|-----------------|
| Formal qlf. rate | CD | -0.5962 | 0.7223 |
| | SSC | -0.7020 | 0.8604 |
| | DPC | -0.6317 | 0.8677 |

that of the weighted #2 cross-semivariogram (from table 8.13 and 8.14). In the second situation, the group level cross-semivariogram parameters, \hat{n}_{ab} , \hat{s}_{ab} , and \hat{r}_{ab} , will be estimated and used as the initial value of the n_{ab} , s_{ab} , and r_{ab} .

Equation (6.92) shows that the parameters (n_{ab} , s_{ab} , and r_{ab}) can be estimated using a particular group level data. It implicitly says that those three grouping factors will give similar parameter estimates. Fortran programs and SAS procedures, as presented in appendix (E), were used and the results are presented in table (8.17).

Table 8.17. Estimation of n , s , and r of the employment rate versus income below 20,000 and employment rate versus income over 40,000

| grouping | situation #1 | | | situation #2 | | |
|---|------------------|------------------|------------------|------------------|------------------|------------------|
| | \hat{n}_{ab}^1 | \hat{s}_{ab}^1 | \hat{r}_{ab}^1 | \hat{n}_{ab}^2 | \hat{s}_{ab}^2 | \hat{r}_{ab}^2 |
| Employment rate vs rate income < 20,000 | | | | | | |
| CD | 0.0 | -0.0054 | 4.9242 | 0.0 | -0.0054 | 5.2553 |
| (se) | (0.0056) | (0.0056) | (0.0741) | (0.0056) | (0.0056) | (0.0769) |
| SSC | 0.0 | -0.0104 | 2.3394 | 0.0 | -0.0104 | 2.3394 |
| (se) | (0.0899) | (0.0899) | (0.0809) | (0.0901) | (0.0899) | (0.0810) |
| DPC | 0.0 | -0.0058 | 7.0943 | 0.0 | -0.0058 | 7.0943 |
| (se) | (0.4504) | (0.4502) | (0.9724) | (0.4504) | (0.4502) | (0.9725) |
| Employment rate vs rate income > 40,000 | | | | | | |
| CD | 0.0 | 0.0039 | 9.8688 | 0.0 | 0.0039 | 10.1121 |
| (se) | (0.0041) | (0.0041) | (0.1073) | (0.0041) | (0.0041) | (0.1105) |
| SSC | 0.8844 | 0.8883 | 9.6043 | 0.8844 | 0.8882 | 9.6043 |
| (se) | (0.0581) | (0.0580) | (0.4702) | (0.0581) | (0.0581) | (0.4702) |
| DPC | 0.5994 | 0.6033 | 10.5130 | 0.5994 | 0.6033 | 10.5130 |
| (se) | (0.3271) | (0.3270) | (1.2801) | (0.3271) | (0.3270) | (1.2801) |

Note : se = standard error of the estimated

Table (8.17) shows that the situation #1 and #2 give a similar estimated parameter. The standard error of the estimated parameter at the CD level gives the smallest values compared with the SSC level and DPC

level due to the number of groups in the analysis. For the cross-semivariogram involving income below 20,000, the same nugget estimate of zero is obtained using data for all three levels. The estimate sill values are similar for the CD and DPC level, but larger for the SSC level. The estimated range values differ for the three levels. For the cross-semivariogram involving income over 40,000 the estimated range parameters are similar for all three levels. The nugget and range estimates are similar for the SSC and DPC levels, but much smaller when CD level data are used.

8.8 Summary

Investigation of the distribution of the data was done by spatial approaches, such as by 3-d scatter plot, contour plot, and surface plot. The variation over the region was also investigated and measured by looking at semivariogram analysis and spatial autocorrelation.

The group level semivariograms were produced for each characteristics. The theoretical derivation has shown that the individual level semivariogram is somewhere above the unweighted group level semivariogram (see section 5.5). This was also confirmed by the simulation result. The weighting factors were applied in calculation of the group level semivariogram. The weighting factors make significant adjustment to the unweighted group level semivariogram. Adjustment on the semivariogram value by weighting affect the estimated parameters. The nugget, sill, and range are greatly increased in going from the unweighted to the weighted version. The nugget and sill also differ between using two weights, but the range does not show a big difference.

Data from a micro sample can play an important role in the analysis of the group level data. The micro samples directly can be used as a benchmark for estimation of the individual level parameters based on the group level data. Also, the micro sample improves the estimate of parameters of the individual level semivariogram from the group level data (see section 5.4). In real data, the use of micro sample was discussed particularly for the Adelaide region.

Estimation of the individual level semivariogram parameters from the group level data was done for the employment rate of the Adelaide data, by using non-linear regression method. The results in table (8.11) show that information of the individual level statistics, such as S_{yy} , improved the estimated nugget and

range (method 1b) and also reduced the standard error of the estimate. Table (8.18) shows a comparison of the estimated variance from the micro sample, \hat{s}_{ms}^2 , the individual level variance S_{yy} , and the estimated sill by the non-linear method for three different grouping levels. Methods 1a and method 2 gave a similar result for the estimated sill, but they differ with the \hat{s}_{ms}^2 and S_{yy} , except for the method 1b which directly uses the individual level variance. Hence, while method (1a) and (2) appeared to work well with simulated data, their performance with real data suggest further development may be necessary.

Table 8.18. Comparison of \hat{s}_{ms}^2 , S_{yy} , and estimated sill from the non-linear estimation methods of the employment rate of Adelaide data

| method | Estimated sill | | | sample | census |
|--------|----------------|-----------------|-----------------|------------------|----------|
| | \hat{s}^{cd} | \hat{s}^{ssc} | \hat{s}^{dpc} | \hat{s}_{ms}^2 | S_{yy} |
| 1a | 1.6139 | 1.2164 | 9.8699 | 0.2479 | 0.2487 |
| 1b | 0.2479 | 0.2479 | 0.2479 | | |
| 2 | 1.6139 | 1.2165 | 9.8665 | | |

Note : 1a and 1b = individual sample data are available,
2=group level data are available

It is interesting to look at the relationship between the sill of the group level semivariogram parameter estimates with the result of the micro sample. Comparison of the result may be found in Table (8.19). Table (8.19) shows that the unweighted sill is close to the unweighted group level variance. The estimated sill from the weighted #1 semivariogram is close to the weighted group level variance, except for the employment rate, labor force participation rate, and the rate of salary earner. These weighted sill of these variables are shifted from the weighted group level variance, due to the variation at the distance 15 km and more, see figure (8.5-b), (8.7-b), and (A.12-b).

Table 8.19. The comparison of the sill of the group level semivariogram parameter estimates and the variance of the micro sample

| Characteristic | Sill Γ_{gh} | Variance ${}_l\bar{S}_{yy}$ | Sill $\hat{\Gamma}_{gh}^{wl}$ | Variance ${}_N\bar{S}_{yy}$ |
|----------------------|-----------------------|--------------------------------|----------------------------------|--------------------------------|
| Employment | 0.0154 | 0.0134 | 7.2816 | 5.8739 |
| Unemployment | 0.0012 | 0.0013 | 0.5322 | 0.5321 |
| Labor participation | 0.0128 | 0.0112 | 6.1948 | 4.9085 |
| Income below 20,000 | 0.0086 | 0.0082 | 3.6470 | 3.4513 |
| Income 20,000-40,000 | 0.0045 | 0.0044 | 2.2249 | 1.8344 |
| Income over 40,000 | 0.0028 | 0.0026 | 1.5345 | 1.4366 |
| Wage/salary earner | 0.0131 | 0.0100 | 6.4289 | 4.3664 |
| Self employed | 0.0003 | 0.0003 | 0.1326 | 0.1198 |
| Employer | 0.0006 | 0.0006 | 0.2415 | 0.2252 |
| Formal qualification | 0.0085 | 0.0073 | 3.7570 | 3.1227 |
| Informal qlf. | 0.0021 | 0.0015 | 0.9853 | 0.6590 |

Chapter 9

Summary and Discussion

This thesis has considered some of the issues associated with the analysis of social data, in particular census data, taking into account the spatial aspect of the data. Such data are often available in group level aggregate form and so we have focused on a method of analysing such data. This led to the consideration of aggregation effects and the well known MAUP. Theory was developed that clearly showed the role of the individual level spatial relationship, as reflected in the individual level semivariogram.

Aggregate social data may contain information on the geographic location of the groups and spatial methods of analysing these data were considered. Attention was focused on estimating individual level semivariograms and cross-semivariograms using spatial group level data. The spatial group level data are assumed to be a transformation of data from a particular random process Y_i . A random process Y_i may be formulated in general into

$$Y_i(\ell) = \mu_i(\ell) + S(\ell) + \epsilon_i \quad (9.1)$$

where $\mu_i(\ell)$ is a mean component of $Y_i(\ell)$, which may present in the analysis as trend or differences in mean levels (drift), $S(\ell)$ is a random component which is defined as irregular spatially correlated, and ϵ_i is a random component which is due to measurement error or short range spatial variation (Burrough & McDonnell, 1998), throughout the thesis, it is assumed a constant mean, μ , which implies the slope or drift are omitted in the analysis. The violation of this assumption is assumed to be relatively small compared with the effect of $S(\ell)$ at a short distance, but the effect of the drift is probably shown at a long

distance. This violation may be observed in terms of a periodical wave as shown in the semivariogram or cross-semivariogram graphs, e.g. figure (A.14), or figure (8.9).

Empirical work evaluating these methods was done using simulated data and data from the 1991 Australian census. This chapter discusses the key points from those developments.

9.1 Basic aggregation effect

Aggregation is a common process used in summarizing data. For the Australian census data, the aggregation was done at CD level, since the CD is the smallest geographical division for collecting data. Analysing aggregated data will often give a different result from the same analysis based on the corresponding unit level data. This phenomena has been defined as the ecological fallacy. The difference between the results of the analysis may be defined as the aggregation effect. In this thesis, the aggregation effect was examined by looking at the difference and the ratio of the variation at two levels. In case of the group level and individual level data, then the aggregation effect for a variable can be evaluated using the difference and the ratio between group level variance and individual level variance.

In chapter (4) we defined two types of group level variances, the unweighted and weighted group level variance, which were defined in Equation (4.11) and (4.13), respectively. There is a simple relationship between individual level variance and weighted group level variance as shown in Equation (4.21). Equation (4.21) gives a break down of the individual level variance into the weighted group level variance and the sum of within group individual level variance. This relationship facilitates a derivation of the aggregation effect.

The key result for aggregation effect in terms of the differences of variances is given by

$$E(N\bar{S}_{yy} - S_{yy}) = -\tilde{\Sigma}\bar{C}_{N\bar{\Sigma}} - \frac{M(\bar{N} - 1)}{M - 1}\bar{\Delta} + \tilde{\Delta}_W \left(\frac{M(\bar{N} - 1)}{M - 1} + \bar{N}\bar{C}_{N\bar{\Delta}} \right) + N\bar{S}_{\mu\mu} - S_{\mu\mu} \quad (9.2)$$

where $\tilde{\Sigma}$ is the unweighted average group level variance $\bar{\Sigma}_g$, $\bar{C}_{N\bar{\Sigma}}$ is the relative covariance between N_g and $\bar{\Sigma}_g$, $\bar{\Delta}$ is the population covariance, $\tilde{\Delta}_W$ is the unweighted average within group covariance, and $\bar{C}_{N\bar{\Delta}}$ is the relative covariance between N_g and $\bar{\Delta}_g$.

This equation was formulated as the difference of weighted group level variance and individual level variance. In some cases we can eliminate the factor $\bar{C}_{N\bar{\Sigma}}$ and $\bar{C}_{N\bar{\Delta}}$, for example when the N_g or $\bar{\Delta}_g$ are

constant. The main factors are $\bar{\Delta}$ and $\tilde{\Delta}_W$. The factor $\bar{\Delta}$ is independent of the scale or zoning effect, but the factor $\tilde{\Delta}_W$ is affected by the scale and zoning. Hence understanding how the factor $\tilde{\Delta}_W$ behaves is the main issue. The expectation of the aggregation effect in terms of difference mainly comes from the difference between $\bar{\Delta}$ and $\tilde{\Delta}_W$ (see equation 4.2.12).

When the number of groups is large, the term $\frac{M}{M-1}$ may approach to unity, and the aggregation effect is proportional to the average group size times the difference of the average covariance of units within groups and the average covariance between units in the population (see equation 4.55).

9.2 Semivariogram approach to aggregation effect

So far the only spatial aspect taken into account is the group to which individuals belong, through the factor $\tilde{\Delta}_W$. A key idea in the spatial perspective is that entities close to each other are more likely to be similar than entities far away from each other. This leads to the idea that the similarity of the characteristics of units may be a function of the distance between them. Matérn (1986) defined the concept of the characteristic's values changing over space as spatial variability. Spatial variability can be measured in several ways, see Deutsch and Journel (1992, page 40), and one of them is using semivariogram analysis. Anselin (1988) discussed spatial correlation and spatial autocorrelation, and Cliff and Ord (1981) determined the relationship between semivariogram and spatial autocorrelation.

In analysing aggregate social data, one may set targets of inference which do not explicitly involve any spatial aspect, however spatial analysis should be done to tackle the dependency of the observations. Some studies has been mentioned spatial variability and aggregation effects (section 3). But mostly, these issues are examined separately. In this thesis the aims was to investigate these matters in a comprehensive way, that is exploring aggregation effect by taking into account the spatial variability. This research focuses on applying semivariogram methods in analysing social data and developing more appropriate methods of analysing spatially aggregated social data.

9.2.1 Empirical perspective on the semivariogram and aggregation effect

The key point in section (4.3) is the relationship between group level semivariogram and the components of the individual level semivariogram, which was formulated in theorem (4.3.6). This theorem shows that the group level semivariogram can be broken down into three main components, $\bar{\hat{\gamma}}_{gh}$, $\bar{\hat{\gamma}}_g$, and $\bar{\hat{\gamma}}_h$. The first component comes from the average of the individual level semivariogram between the groups, and the last two components are the average of the individual level semivariogram within the groups. This theorem provide a relationship between the group level and individual level semivariogram values without assuming any model.

In theorem (4.3.8) the empirical weighted group level variance was related to the individual level semivariogram value. The empirical weighted group level variance contains three factors, $\frac{N-1}{M-1} \bar{\hat{\gamma}}$, $\bar{N} \bar{\hat{\gamma}}_w$, and $(\bar{C}_{N\bar{\hat{\gamma}}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}})$. After allowing for change in N and M , the first factor is not affected by the scale or zoning, but the second and third factors are. This result shows clearly how the aggregation effect can be related to the spatial statistic of the population, as measured by the $\hat{\gamma}_{ij}$ values.

9.2.2 Theoretical perspective on the semivariogram and aggregation effect

The semivariogram approach measures the similarity between observations by looking at the variance of the difference in pairs of observations. It is estimated by the square of the difference of the pairs observations (see corollary 5.1.3). Results (4.76) shows that the variance of the difference is related to the difference between the average of the variance of the units and their covariance. Model assumptions are needed to express the semivariogram in terms of the distance between pairs of observations. The assumptions were that the process was intrinsically stationary (see section 5.1) and second order stationary (5.4). Under the isotropic condition, then we may express equation (4.76) as a function of distance, where distance is defined as a cartesian distance (5.5) between pairs of observations.

Applying these assumptions provides a way of relating the semivariogram to the distance, and a semivariogram model can be developed. Some commonly used models are the exponential, spherical, and Gaussian models. Some other models are also used, such as power, rational quadratic, and wave model. The common parameters of a semivariogram model are nugget, sill, and range. Several methods may be

applied to fit semivariogram models, such as weighted least squares, maximum likelihood (ML), restricted maximum likelihood (REML), minimum norm quadratic, and generalized linear model. The weighted least square method is applied in the estimation procedure through out this thesis.

The nugget parameter of the semivariogram model indicates variation of observations at the same location. In physical sciences, ideally, one object at the same location can be measured identically. A different result will indicate a measurement error. But in social science, the situation may be different. For example, several respondents might be taken from one household which in practice is a single location. In another case it may be taken from a groups of household, e.g. an apartment, which may be recorded as one geographic location.

The sill parameter may refer to the asymptotic value of the semivariogram model, such as in exponential, spherical, or Gaussian model. This point indicates that the semivariogram has reached its constant value, which is the variance of the process.

The nugget and sill parameters determine the level of semivariogram, but the last parameter – range – will reflect the distance at which spatial relationships are weak. That is a distance where the semivariogram line is approaching a flat line (Figure 5.1). Moreover, it correspond to distance after which the observations are close to independent of each other. In other words, we can say that there is a spatial inter-relationship between observations at a distance less than the range distance.

In the scope of CD level, then semivariogram analysis may give an interesting overview of the population. The three parameters of the semivariogram model, nugget, sill, and range, can be used to give a description of variation at the household level within the CD, variation at the CD level, and variation of spatial level, respectively. In the same way, it can be thought that when the individual level semivariogram model was estimated successfully, then the parameters may identify variation at the unit level (e.g. person within household), variation among households, and variation at spatial level.

There is a basic relationship between semivariogram and covariogram, such as defined in (5.8), that is,

$$\gamma(d) = \sigma^2 - C(d) \quad (9.3)$$

where d is a distance and $\sigma^2 \approx S_{yy}$. In the context of individual semivariogram, d represents the distance between two individual observations. But in the context of aggregated semivariogram, d may represent

the distance between two different groups. This equation is analogous to equation (4.78). The first term of equation (9.3) refers to the population variance. This is a case when $\Sigma(l_i) = \Sigma(l_j)$ for all $i \neq j$. And the second term refers to the covariance function. The covariance function may be represented as a function of the distance as well. We would expect that the covariance function will take a larger value at a close distance than at a far distance, and its value will approach zero as the distance gets larger. This relation between variogram and covariogram can be used to set up a variance-covariance matrix corresponding to specific variogram model. Then a random process $\mathbf{Y}|\mathbf{L}$ can be generated based on this variance covariance matrix.

Equation (4.96) and (4.98) formulated the aggregation effect in terms of difference and ratio of weighted group level variance and individual level variance, respectively. Assume the constant mean over the population, then they can be expressed as

$$E(N\bar{S}_{yy} - S_{yy}) = \left(\frac{N-M}{M-1}\right) \bar{\gamma} - \bar{N}\tilde{\gamma}_W \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}}\right) \quad (9.4)$$

and

$$\frac{E(N\bar{S}_{yy})}{E(S_{yy})} = \frac{N-1}{M-1} - \bar{N}\frac{\tilde{\gamma}_W}{\bar{\gamma}} \left(\bar{C}_{N\bar{\gamma}} + \frac{M}{M-1} \cdot \frac{\bar{N}-1}{\bar{N}}\right) \quad (9.5)$$

Both expression of the aggregation effect contain the factor $\bar{C}_{N\bar{\gamma}}$, which is the coefficient of co-variation between N_g and $\bar{\gamma}_g$. If the N_g or $\bar{\gamma}_g$ is constant, $\bar{C}_{N\bar{\gamma}}$ is zero and, in general, we expect that the factor $\bar{C}_{N\bar{\gamma}}$ is very small. Therefore the main factor in addition to \bar{N} affecting the scale or zoning is $\tilde{\gamma}_W$.

In case of constant group size the factor $\bar{C}_{N\bar{\gamma}}$ is zero, therefore the expectation of the aggregation effect mainly contains $\left(\frac{M(\bar{N}-1)}{M-1}\right)$ and $(\bar{\gamma} - \tilde{\gamma}_W)$. The \bar{N} and M are the average of groups size and number of group within the population, respectively, and are usually known. The $\tilde{\gamma}_W$ term indicates the average within group semivariogram. Aggregation effect will equal to zero when $\bar{\gamma} = \tilde{\gamma}_W$. This is the case when dispersion of the observations in the population is identical with dispersion of the observations within all the groups.

9.3 Relationship of group level semivariogram and individual level semivariogram

The basic relationship between the components of the individual level and group level semivariogram was given in theorem (5.2.2), which is the expectation of the theorem (4.3.8). The later theorem was developed as an algebraic expression, and the the former theorem was set up by considering some assumptions for the semivariogram. This key relationship was defined in equation (5.34), which involved the mean of the between group semivariogram function, the mean of within group semivariogram function, and the groups sizes.

Taylor series expansion was applied to (5.34) to derive a relationship between the group level semivariogram and the parameters of the individual level semivariogram. Several shapes were considered, although for later development the shape of the groups was assumed to be as a circle, since the distance moment in the circle shape was the simplest compared with other shapes (see section 5.3.2). The result (5.51) shows the group level semivariogram as a function of the individual level semivariogram evaluated at d_{gh} , \bar{d}_g and \bar{d}_h . The others components were \bar{d}_{gh} , $S_{d_{gh}}^2$, $S_{d_g}^2$, and $S_{d_h}^2$. The first two are the mean and variance of the distances of pairs between two groups, and the last two is the mean of the distances of pairs within the group. These quantities have been evaluated in section (5.3.2) for some common shapes of the group.

The simple case, the exponential semivariogram model for the individual level semivariogram was considered. Further assuming that the first and second derivative of the individual level semivariogram are small, then the approximation was simplified into (5.93) and then (5.96). This approximation was applied in section (5.4) to the estimation of parameters of the individual semivariogram model, using a non-linear regression method. Simulation results showed that the existence of non-spatial individual sample (such as in micro sample data) can be used to improve the accuracy of the estimation. But without any information of the individual sample, the estimation procedure can still work well, though some outliers still have to be tackled carefully.

9.4 Group level semivariogram adjustment

Adjustments were also done by putting a weighting factor into the group level semivariogram calculation. Appropriate weighting factors were developed in case of constant group size (\bar{N}). The two weighting factors can be found in the table (9.1). The expectation was defined by assuming a constant spatial autocorrelation within the group $\bar{\rho}$. The expectation result showed that spatial autocorrelation within the group ($\bar{\rho}$) and its group size (\bar{N}) make an important contribution to the group level semivariogram.

Table 9.1. The weighting factors and the expectation of $(\bar{Y}_g - \bar{Y}_h)^2$

| <i>f</i> | Weighting factor (<i>w_f</i>) | <i>E</i> (·) when <i>N_g</i> = <i>N_h</i> = \bar{N} , $\rho_g = \rho_h = \bar{\rho}$ |
|----------|---|--|
| 1 | ${}_N\hat{\Gamma}_{gh}$ | $E({}_N\hat{\Gamma}) = \sigma^2 \left(1 + \frac{\bar{N}-1}{\bar{M}-1} M \bar{\rho} \right)$ |
| 2 | $\left(\frac{1}{N_g} + \frac{1}{N_h} \right)^{-1}$ | $E \left(w_2 \cdot (\bar{Y}_g - \bar{Y}_h)^2 \right) = \sigma^2 (1 + (\bar{N} - 1) \cdot \bar{\rho})$ |

Chapter 10

Conclusion

The thesis has shown the role of semivariogram analysis (univariate case) and cross-semivariogram analysis (bivariate case) in understanding the aggregation effect of the social data. The aggregation effect, which includes the two main aspects of the scaling effect and zoning effect, is mainly determined by the presence of spatial dependency within the data. Another factor is the relationship between group size and within group variation. Investigating the first cause of dependency has led to the spatial analysis of aggregate social data by applying semivariogram and cross-semivariogram analysis.

10.1 Contributions, recommendations, and limitations

10.1.1 Some contributions

This research has led to several achievements in investigating the role of the semivariogram and cross-semivariogram in the analysis of social data, in particular for aggregation effects and the MAUP.

First

Semivariogram analysis can be used as an exploratory tool to reveal the existence of the dependency within social data from a spatial perspective. The nugget of the semivariogram may indicate some degree of intra-level correlation. If the individual level semivariogram can be estimated, then the nugget will indicate the intra-household correlation. If the group level semivariogram is estimated, then the nugget will show intra-group correlation. The cross-semivariogram can be applied for the bivariate case. The sill reflects the

variance of the observation in the univariate case, and the covariance of the observations in the bivariate case. The range shows the level of dependency within the data in term of distance of pair of observations. Therefore within the range the observations tends to have higher relationship than the observations outside the range.

Individual level data containing information on the location of individual is rarely available, however, aggregate group level data with geographic location is sometimes available and so a method to appropriately use such data will be very useful.

Second

Theory is developed for the aggregation effect which shows that the within group semivariogram ($\tilde{\gamma}_w$) plays an important role in determining the aggregation effect. The aggregation effect in terms of the difference or ratio of group level and individual level variances indicates the importance of the average within group semivariogram factor. Using semivariogram analysis this factor can be related to a semivariogram model. The factor $\bar{C}_{N\tilde{\gamma}}$ in the aggregation effect will often be small and make little contribution. The factor $\tilde{\gamma}$ is not affected by either scale or zoning.

Third

The theory related aggregation effect to the empirical semivariogram values. Further assumptions, such as intrinsically stationarity and second order stationarity can be applied. The study shows that the group level semivariogram, which was derived from the aggregated data, will give biased estimates of the individual level semivariogram. This bias is a major cause of the aggregation effect. Using Taylor series method, the decomposition of the group level semivariogram was obtained

$$\begin{aligned} \Gamma_{gh} \approx & \gamma(d_{gh}) \\ & + \gamma'(d_{gh})(\bar{d}_{gh} - d_{gh}) + \frac{\gamma''(d_{gh})}{2} S_{d_{gh}}^2 + \frac{\gamma''(d_{gh})}{2} (\bar{d}_{gh} - d_{gh})^2 \\ & - \frac{N_g - 1}{2N_g} \left[\gamma(\bar{d}_g) + \frac{\gamma''(\bar{d}_g)}{2} S_{d_g}^2 \right] - \frac{N_h - 1}{2N_h} \left[\gamma(\bar{d}_h) + \frac{\gamma''(\bar{d}_h)}{2} S_{d_h}^2 \right] \end{aligned} \quad (10.1)$$

This equation shows that the group level semivariogram is represented by a function of individual level semivariogram, which are affected by several factors, \bar{d}_{gh} , $S_{d_{gh}}^2$, \bar{d}_g , \bar{d}_h , $S_{d_g}^2$, and $S_{d_h}^2$. These factors can be

represented as a function of area and shape of the group. Hence it shows how the role of area and shape of the group affected the group level semivariogram. Consider an example for the exponential model of the semivariogram, which can be formulated as

$$\begin{aligned}
 \Gamma_{gh} \approx & s \cdot \left\{ \frac{1}{2N_g} + \frac{1}{2N_h} \right\} - (s - n) \cdot \exp \left[\frac{-3d_{gh}}{r} \right] \\
 & + \frac{3}{r} (s - n) \cdot \exp \left[\frac{-3d_{gh}}{r} \right] \left\{ (\bar{d}_{gh} - d_{gh}) - \frac{3}{2r} S_{d_{gh}}^2 - \frac{3}{r} (\bar{d}_{gh} - d_{gh})^2 \right\} \\
 & + \frac{N_g - 1}{2N_g} (s - n) \cdot \exp \left[\frac{-3\bar{d}_g}{r} \right] \left\{ 1 + \frac{9}{2r^2} \cdot S_{d_g}^2 \right\} \\
 & + \frac{N_h - 1}{2N_h} (s - n) \cdot \exp \left[\frac{-3\bar{d}_h}{r} \right] \left\{ 1 + \frac{9}{2r^2} \cdot S_{d_h}^2 \right\}
 \end{aligned} \tag{10.2}$$

Fourth

Applying a semivariogram model to describe the factor of average within group semivariogram suggests an approach to estimate the parameters of the individual level semivariogram from the group level semivariogram. The theorems on semivariogram can be extended into bivariate case using the idea of a cross-semivariogram.

Equation (10.2) explicitly showed that the individual level parameters of a semivariogram model can be estimated by making assumptions the shapes of groups. This assumption leads to an approximation for the \bar{d}_g , \bar{d}_h , $S_{d_g}^2$, and $S_{d_h}^2$. Given the $\hat{\Gamma}_{gh}$, \mathcal{A}_g , \mathcal{A}_h , N_g , N_h , and d_{gh} , then the parameters n , s , and r can be estimated through non-linear regression methods. The method is easily extended into bivariate case in terms of cross semivariogram.

Introducing the micro sample gives some advantages, besides being a bench-marking tool of the estimated parameter sill, it also can be used to improve on the estimation procedure of the parameters of individual level semivariogram.

Fifth

The emphasis in the literature and this thesis has mainly been on explaining and accounting for the MAUP. The role of spatial relationship between individuals, as reflected by the semivariogram has been identified. The basic viewpoint can be reversed, and it is shown how the MAUP may be used as a method of spatial analysis. A method is developed for the estimation of parameters of the individual level semivariogram

from different realization of the group level semivariogram. These different realizations of the group level data will involve the scale effect or the zoning effect. These effects can be formulated into semivariogram analysis by using the decomposition of the aggregation effect into components determined by the individual level semivariogram.

Sixth

The consideration of the weighting factors for the group level semivariogram can give a range for the estimated parameters of the individual level semivariogram. These values can be useful in determining the initial value in the formal estimation procedure of the parameters of individual level semivariogram from group level data.

10.1.2 Recommendations

Semivariogram and cross-semivariogram analysis are effective means of exploring the existence of spatial dependency within social data. The estimation of the parameters of individual level semivariogram and cross-semivariogram can be done using group level data in two situations. The first situation is when the micro sample data are available, and the variance of the specified characteristics are accessible, which can be used to determine the initial values of the estimation procedure. The second situation is when micro sample data are not available and the estimated weighted group level semivariogram can be used to give the initial value of the estimation procedure.

If the group sizes or within group means of the semivariogram are constant the analysis of aggregation effect using semivariogram analysis is simplified. If these factors cannot be controlled then the aggregation effect on the semivariogram analysis includes the factor $\tilde{C}_{N\bar{y}}$. However this factor was shown to have little effect in simulation status.

The estimation of parameters of the individual level semivariogram can be done by applying the MAUP approach in group level semivariogram.

Other semivariogram models can be used following the proposed method here. The exponential semivariogram model was used in this study only for example, but the basic theorems were derived in general.

10.1.3 Limitations

The basic approach for the estimation of parameters of the individual level semivariogram uses an approximation of the group level semivariogram into components involving the distance function and some moments of the *pdf* of distance. The approximation involved the first and second derivative of the distance function. This study has considered the simplest approximation when the second order terms were ignored. This condition implies elimination of the factors from $S_{d_{gh}}^2$ and \bar{d}_{gh} , which are the variance and mean of pairs distance of the *g*th and *h*th group. The approximation discussed uses the d_{gh} which is a distance between centroids. The centroid is defined by the boundary definition, which may rise the boundary problem issue (see section 3.5.3), but it does not intact in this research.

The simulation of the grouping process was done for the group's shape of square, rectangle, or "L shape", to determine the mean and variance of distance within the group. Therefore the approximation of mean and variance of distance within the group of all shape was used. These quantities were used extensively in the estimation procedure.

The use of the MAUP as a tool in the semivariogram analysis was done mainly by simulation. Implementation for real data sets has not been carried out.

10.2 Further research

This research has developed a general approach for the analysis of aggregate, spatial social data. The approach resolves the MAUP by showing how it is determined by the spatial structure at the individual level as measured by the empirical semivariogram values. It develops a method for estimating individual level semivariogram models using aggregate data. This approach warrants further development and evaluation. Some of the issues than can be investigated further are briefly discussed here.

The results concerning how the observed aggregation effect and the group level semivariogram values are related to the individual level semivariogram values (i.e. Corollary 4.3.9, Theorem 4.3.6) do not make any assumptions concerning the nature of the characteristics involves. However, the theoretical semivariogram models used may not be entirely appropriate for binary and categorized data. The general approach

can be extended to include models appropriate for such variables. For example for the binary data, the half squared of difference can only take one of two values, 0 or $\frac{1}{2}$.

The implication of the approach for bivariate analysis can be developed further to consider the estimation of regression and correlation between different variables at distance d . Also the implications of the general approach for multiple regression and multivariate analysis techniques can be developed.

The theory in Chapter (5) concerning the distribution of the distance between pairs of points can be used to further investigate the role of the shape of the geographical groups in the aggregation effects and the MAUP. This effort may include of varying of boundary definition, which imply on variation of centroid over the groups.

The usefulness of including higher order terms in the Taylor Series expansion in the non-linear regression method could be considered. The performance of this method without higher order terms in the simulation results suggests that in many cases including additional terms may not be necessary. However, it could be beneficial to determine the conditions in which the additional terms would be useful.

Further evaluation of the methods of estimating the individual level semivariogram model from group level data could be undertaken. Simulation involving a greater variety of parameter values, semivariogram models and size and shape of groups would be useful to help assess the performance in a wide range of conditions. Application to further real data sets would also be useful. The use of the MAUP as analysis tool in its own right discussed in chapter (7) could also do with more extensive empirical evaluation.

Monte-Carlo type computer intensive methods could be developed as an alternative to the Taylor Series expansion approach use section (5.4). This could be built around equation (5.28). At each iteration of the estimation method the terms $\bar{\gamma}_{gh}$, $\bar{\gamma}_g$, $\bar{\gamma}_h$ could be evaluated by simulation using the current estimated parameter values. A key issue is whether the increase in computer intensity is worthwhile.

More work could be done on diagnostics and standard error. The reliability of the standard errors that are produced by the non-linear regression method needs to be considered. We also need to develop diagnostics to help identify when the use of group level data can produce reliable estimates of the individual level parameters.

The issue of within household level correlation could be taken further. The semivariogram models used imply a very simple model for these correlations. More realistic and complex models could be developed and the estimation method modified accordingly. Also, in some cases, for example in the Australian Census, it is possible to obtain household sample data which would enable direct estimation of these correlations. These estimates could then be combined with the analysis of the aggregate group level data.

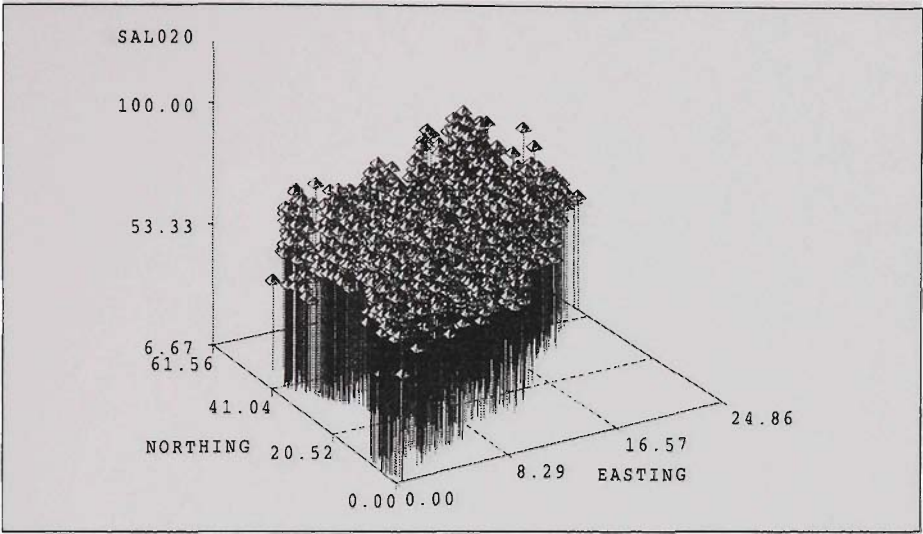
The semivariogram models used in this research have been applied globally. Moreover, the usual assumption of intrinsic stationarity implies that the expectation of the difference between values has no spatial trend. If there is such a trend then there are methods to remove them before undertaking the semivariogram analysis. An alternative would be to apply the theory and methods in this thesis locally and link this research with the ideas of Local Indicators of Spatial Association, LISA (Anselin, 1995). All the key results will apply within any defined neighborhood.

Appendix A

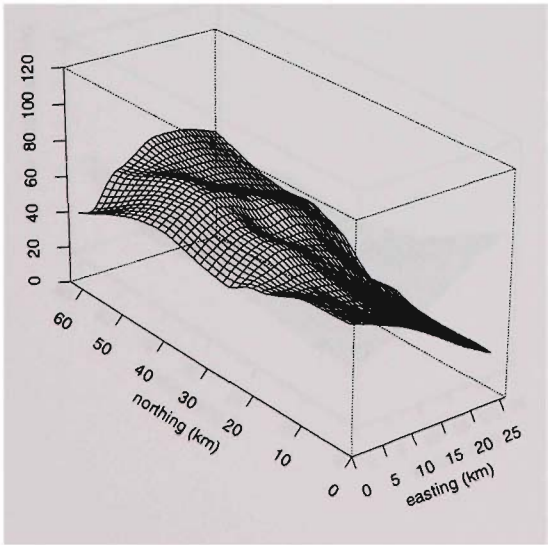
Results of analysing Adelaide CD data

This appendix contains some results of graphical view of spatial perspective of the studied characteristics of the Adelaide CD data. The results also include semivariogram analysis of Adelaide data. This results are presented in form of tables and graphics. The description of this result is found in section (8.3) and (8.4) for the graphical view and cross-semivariogram, respectively. The last section is a tabulation of the micro sample data for Adelaide.

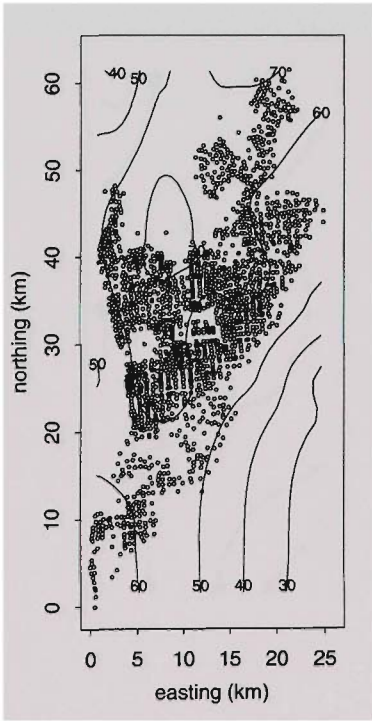
A.1 Graphical view of spatial perspective of the characteristics



(a) 3-d scatter plot

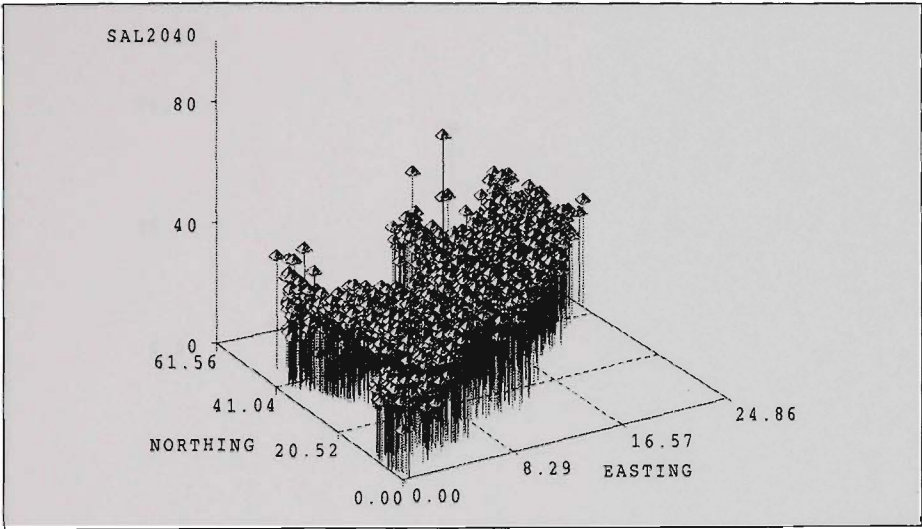


salary 0-20 thousand
(b) Surface plot

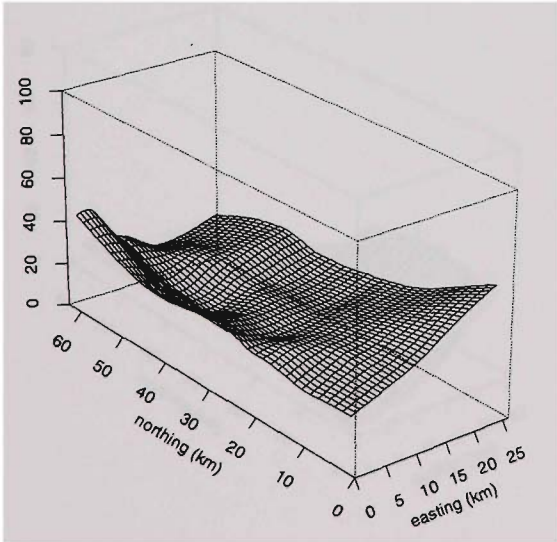


(c) Contour plot

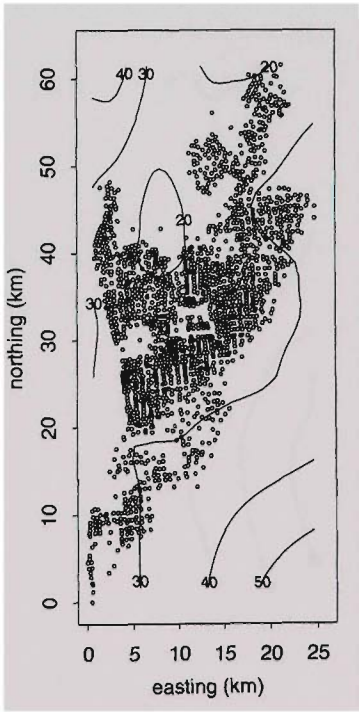
Figure A.1. Scatter plot, contour and surface plot of rate of the income less than 20000



(a) 3-d scatter plot

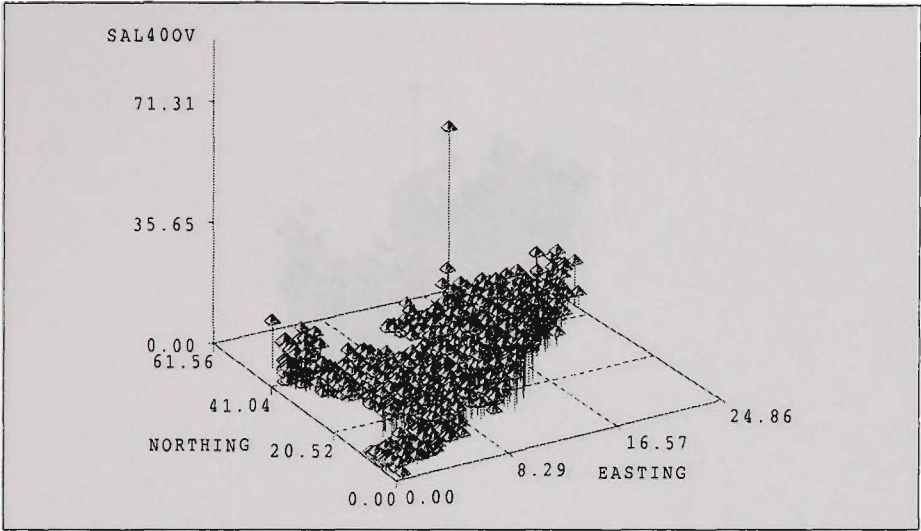


(b) Surface plot

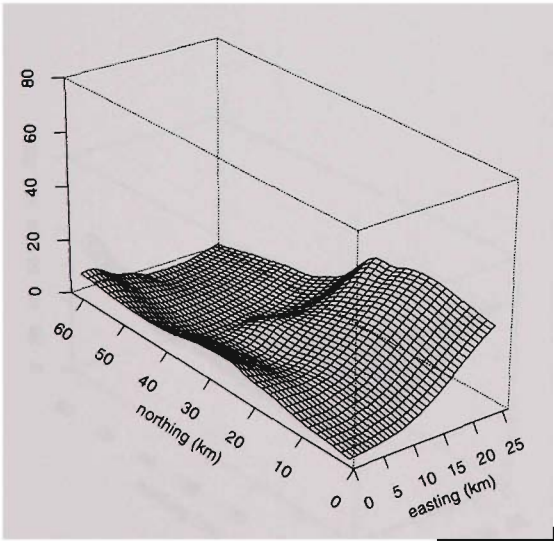


(c) Contour plot

Figure A.2. Scatter plot, contour and surface plot of rate of the income between 20000 to 40000

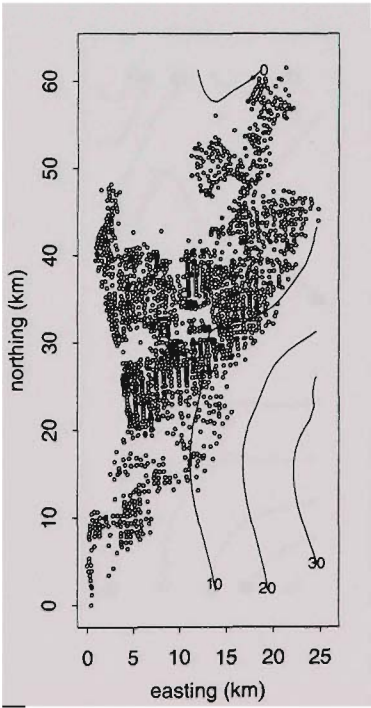


(a) 3-d scatter plot



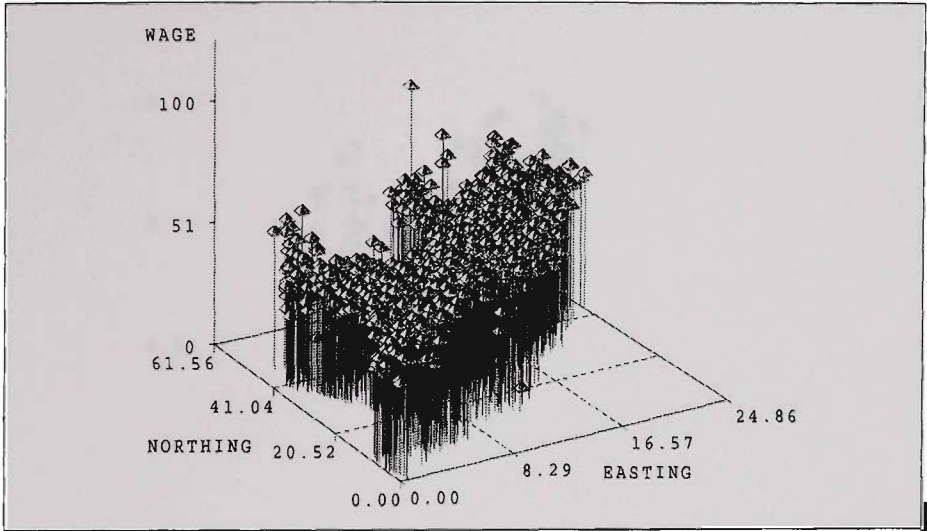
salary over 40 thsnd

(b) Surface plot

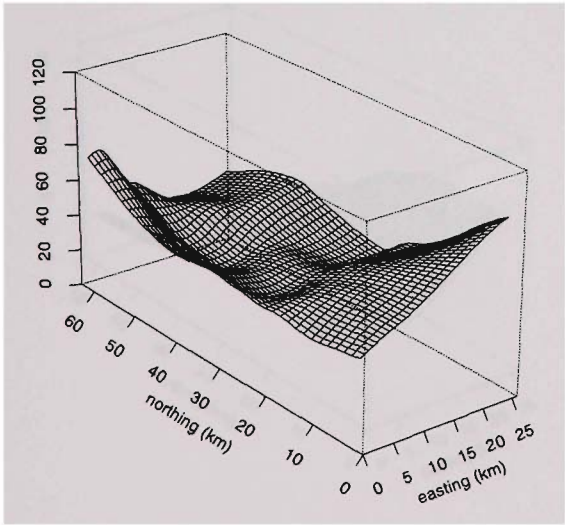


(c) Contour plot

Figure A.3. Scatter plot, contour and surface plot of rate of the income greater than 40000

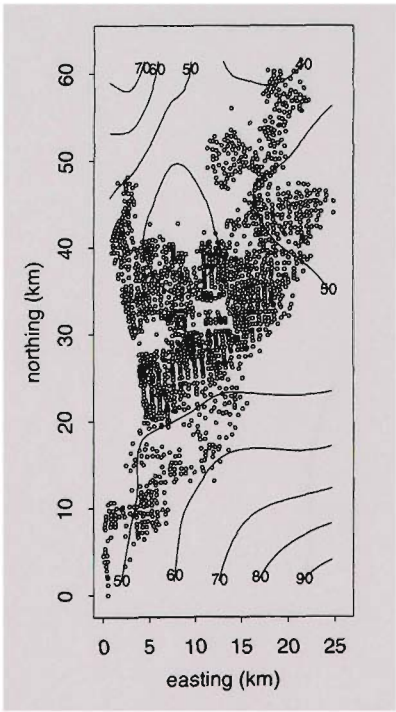


(a) 3-d scatter plot



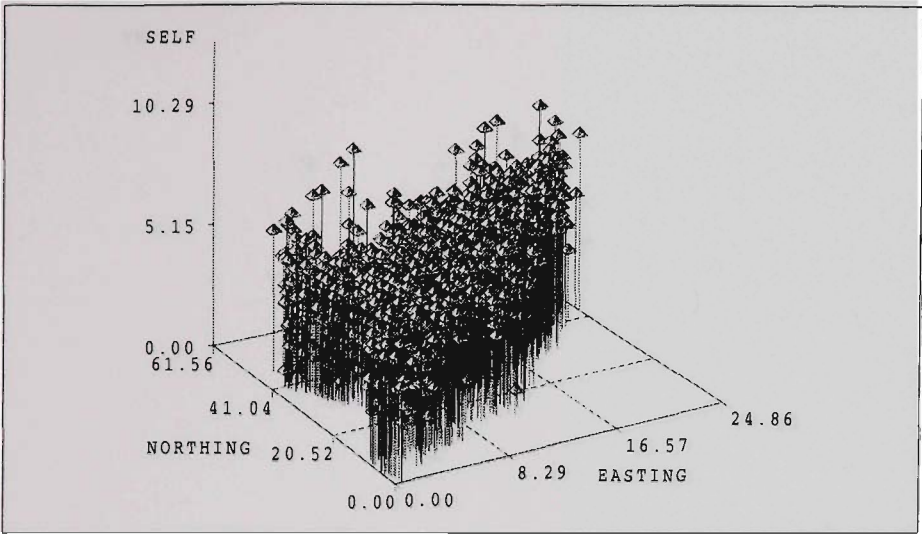
wage or salary earner

(b) Surface plot

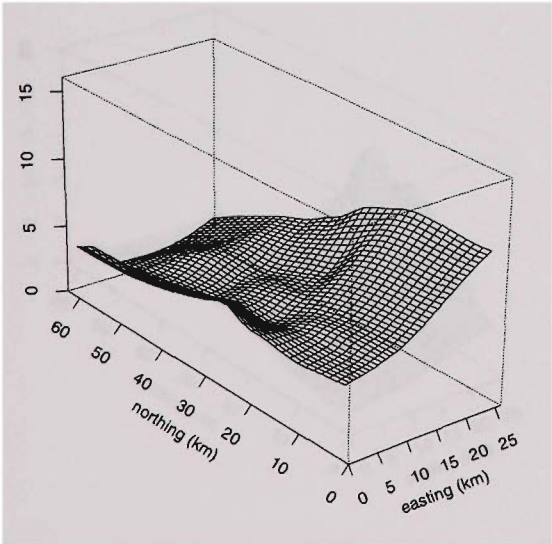


(c) Contour plot

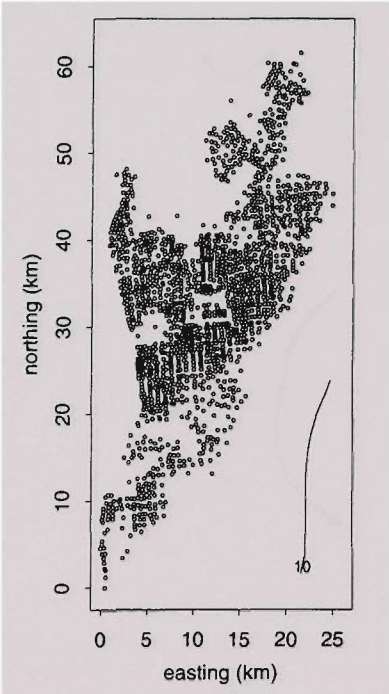
Figure A.4. Scatter plot, contour and surface plot of rate of the wage or salary earner



(a) 3-d scatter plot

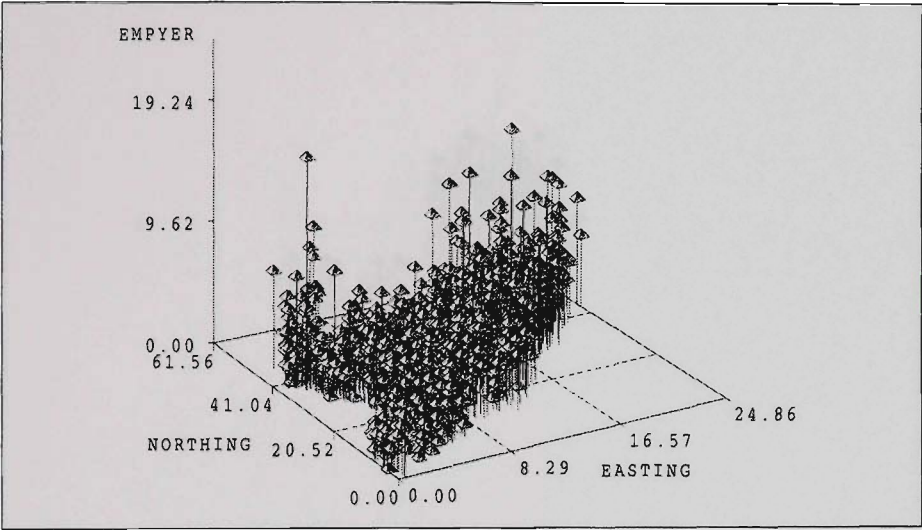


(b) Surface plot

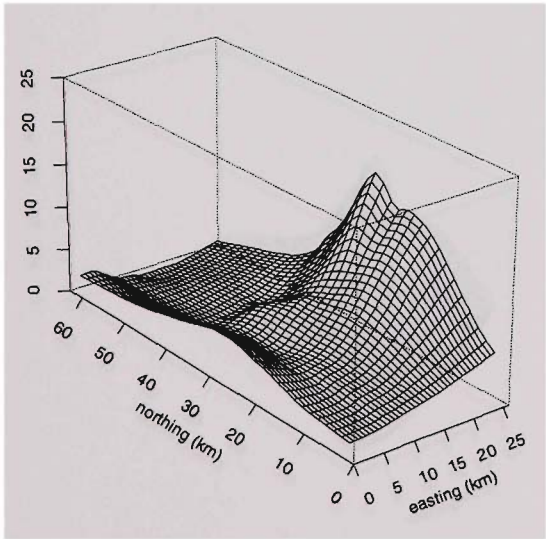


(c) Contour plot

Figure A.5. Scatter plot, contour and surface plot of rate of the self employed persons

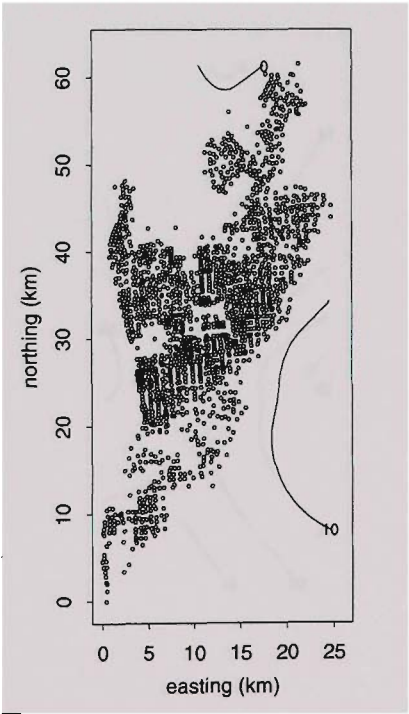


(a) 3-d scatter plot



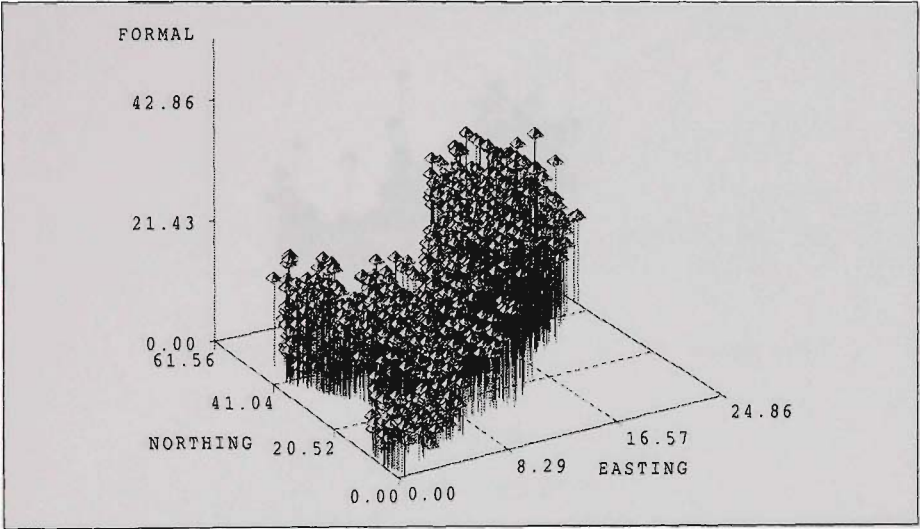
employer

(b) Surface plot

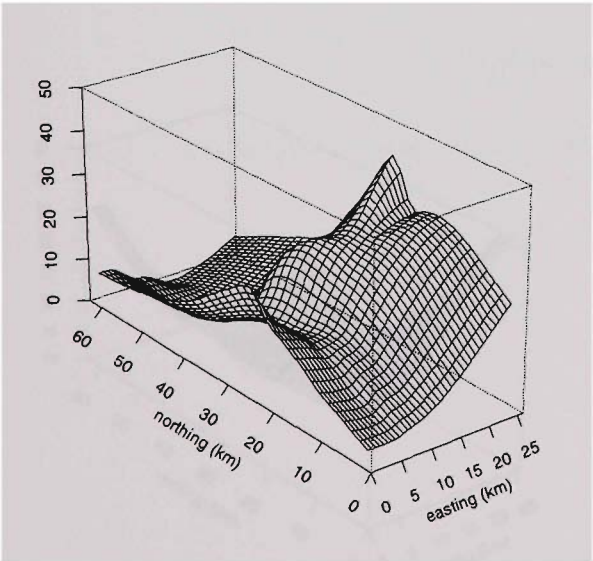


(c) Contour plot

Figure A.6. Scatter plot, contour and surface plot of rate of the employer

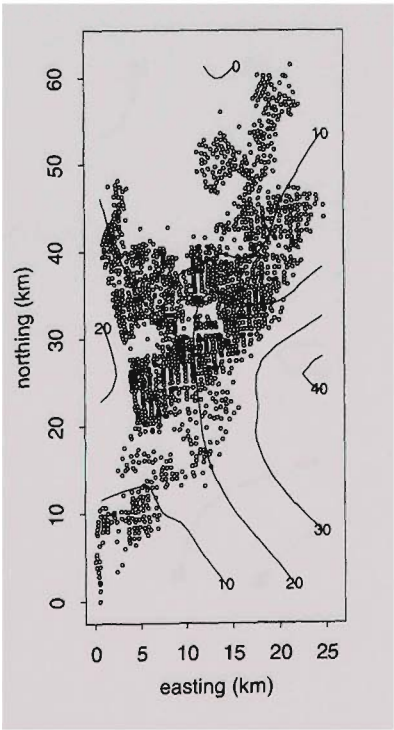


(a) 3-d scatter plot



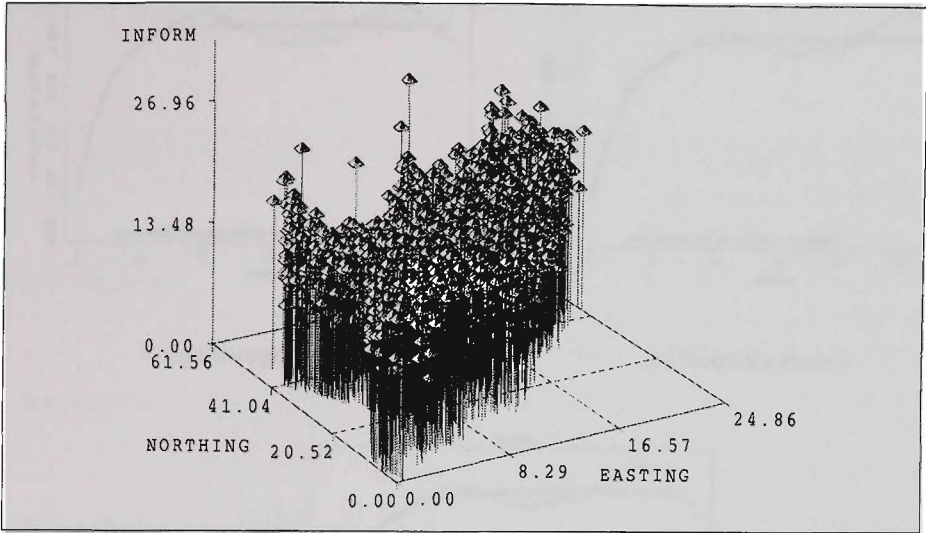
formal qual. rate

(b) Surface plot

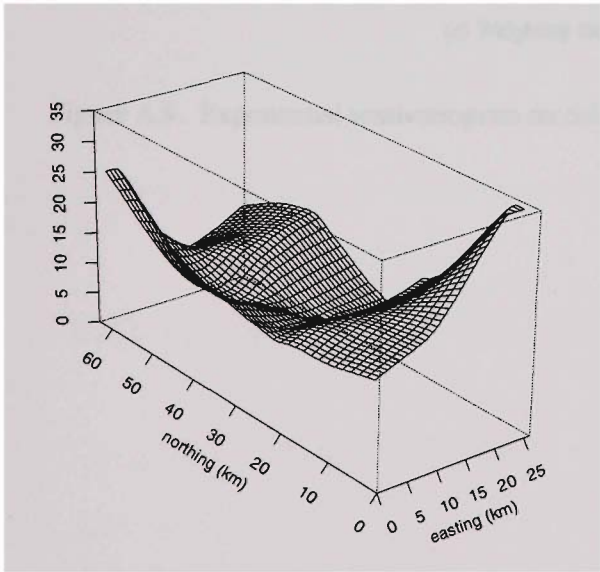


(c) Contour plot

Figure A.7. Scatter plot, contour and surface plot of rate of the formal qualification

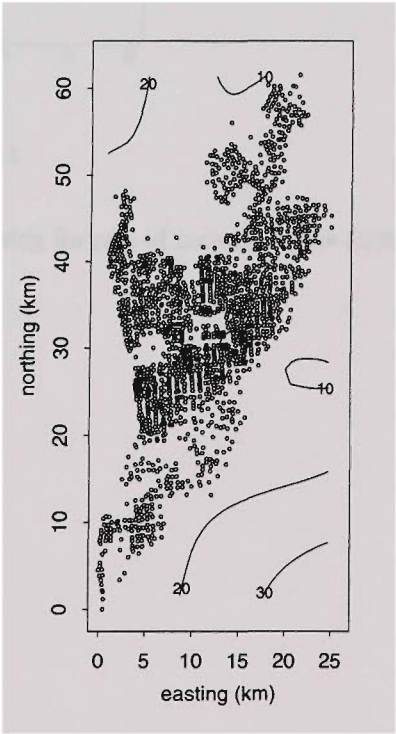


(a) 3-d scatter plot



informal qual. rate

(b) Surface plot



(c) Contour plot

Figure A.8. Scatter plot, contour and surface plot of rate of the informal qualification

A.2 Semivariogram results

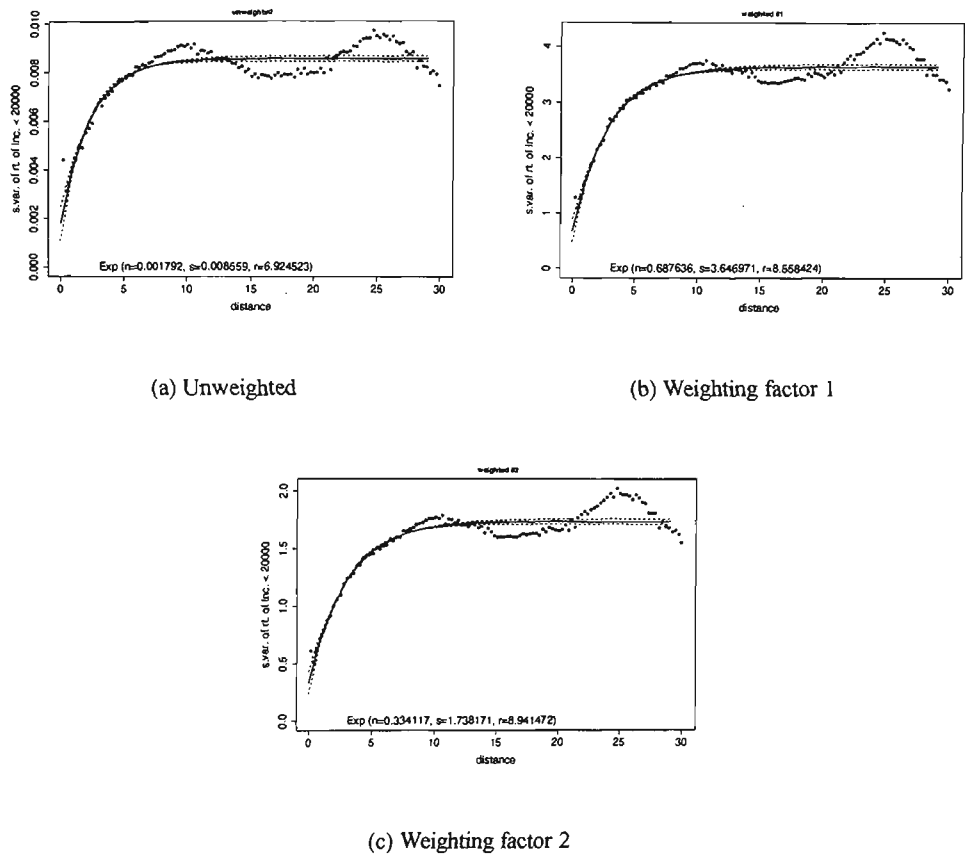


Figure A.9. Exponential semivariogram model fitting for rate of income below 20000

Table A.1. Estimated parameters of the exponential semivariogram model of rate of the income below 20000

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|----------|----------------|-------------------------------------|----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.001792 | 0.000356 | 0.001087 | 0.002497 |
| | sill | 0.008559 | 0.000056 | 0.008448 | 0.008670 |
| | range | 6.924523 | 0.518899 | 5.896385 | 7.952662 |
| weighted #1 | Nugget | 0.687636 | 0.105825 | 0.477957 | 0.897316 |
| | sill | 3.646971 | 0.023123 | 3.601155 | 3.692786 |
| | range | 8.558424 | 0.484751 | 7.597945 | 9.518903 |
| weighted #2 | Nugget | 0.334117 | 0.049171 | 0.236690 | 0.431544 |
| | sill | 1.738171 | 0.011354 | 1.715674 | 1.760667 |
| | range | 8.941471 | 0.503810 | 7.943230 | 9.939713 |

Table A.2. Estimated parameters of the exponential semivariogram model of rate of the income between 20000 and 40000

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.001914 | 0.000077 | 0.001760 | 0.002067 |
| | sill | 0.004547 | 0.000039 | 0.004469 | 0.004625 |
| | range | 15.155897 | 0.925685 | 13.311425 | 17.000369 |
| weighted #1 | Nugget | 0.772853 | 0.022599 | 0.727823 | 0.817883 |
| | sill | 2.224875 | 0.047474 | 2.130281 | 2.319469 |
| | range | 30.270201 | 2.247739 | 25.791471 | 34.748930 |
| weighted #2 | Nugget | 0.369552 | 0.010779 | 0.348074 | 0.391030 |
| | sill | 1.010783 | 0.016669 | 0.977569 | 1.043997 |
| | range | 26.196292 | 1.722009 | 22.765104 | 29.627479 |

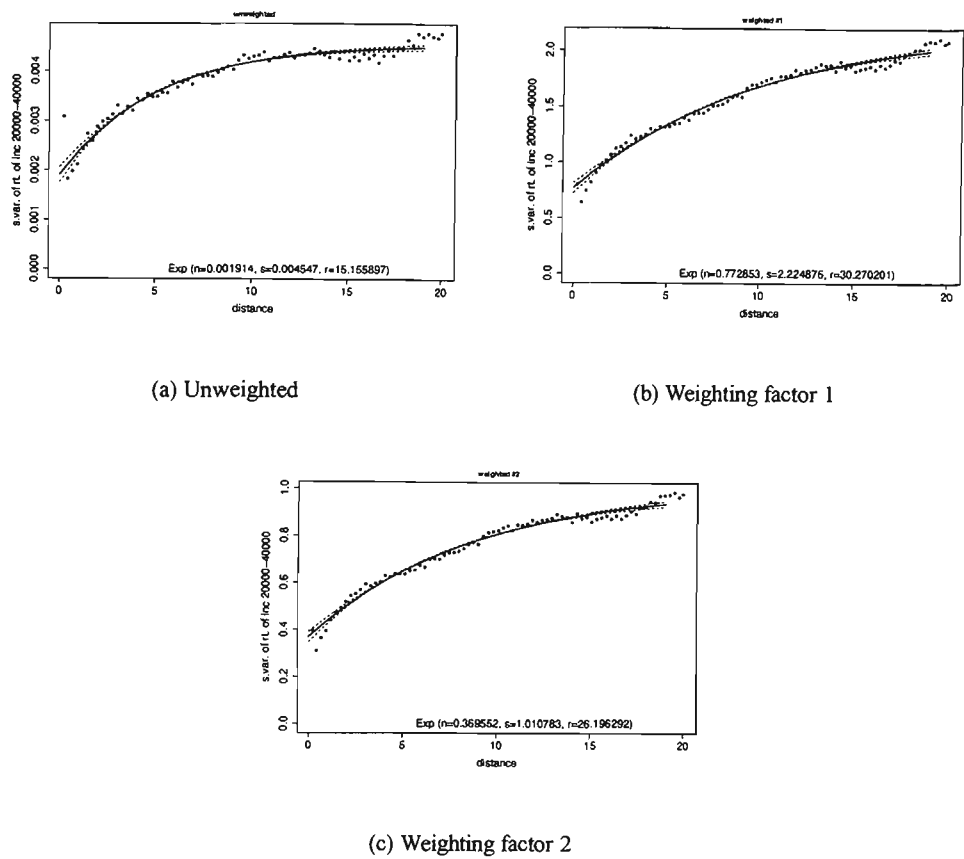


Figure A.10. Exponential semivariogram model fitting for rate of income 20000-40000

Table A.3. Estimated parameters of the exponential semivariogram model of rate of the income over 40000

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|----------|----------------|-------------------------------------|----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000170 | 0.000152 | -0.000132 | 0.000471 |
| | sill | 0.002801 | 0.000022 | 0.002758 | 0.002844 |
| | range | 5.294999 | 0.464484 | 4.374677 | 6.215322 |
| weighted #1 | Nugget | 0.033202 | 0.105597 | -0.176025 | 0.242430 |
| | sill | 1.534477 | 0.013226 | 1.508271 | 1.560682 |
| | range | 4.565517 | 0.483617 | 3.607285 | 5.523749 |
| weighted #2 | Nugget | 0.017620 | 0.022156 | -0.026280 | 0.061520 |
| | sill | 0.652238 | 0.004225 | 0.643867 | 0.660609 |
| | range | 6.038210 | 0.358914 | 5.327063 | 6.749357 |

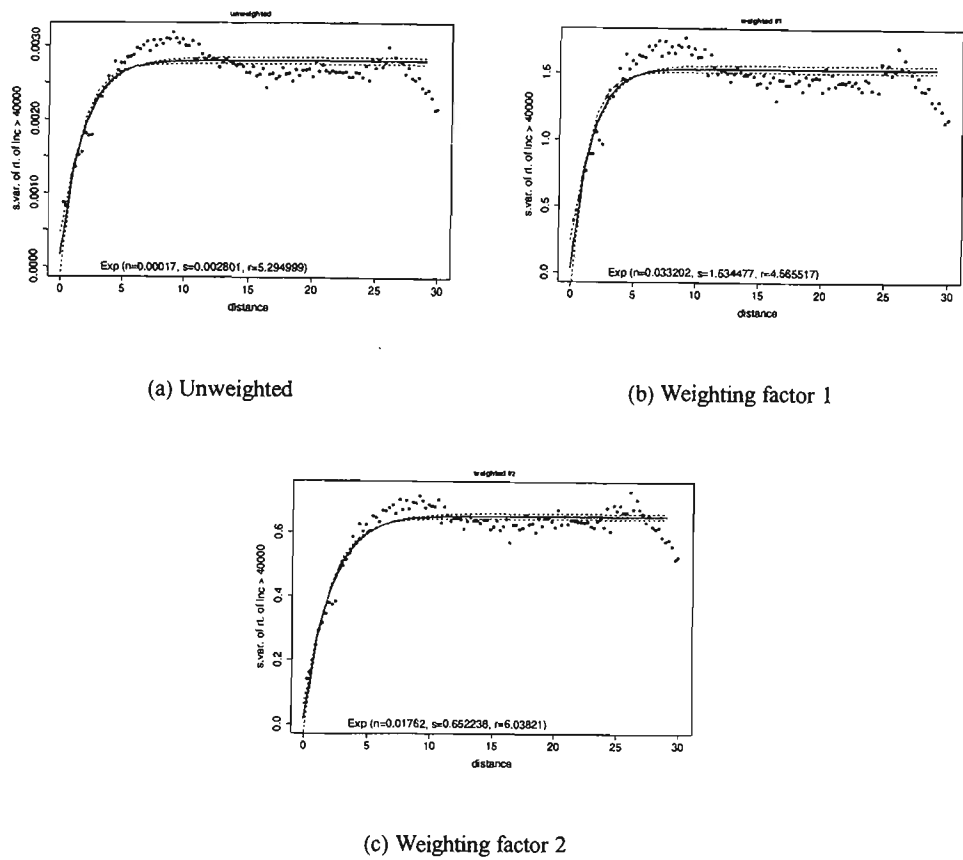


Figure A.11. Exponential semivariogram model fitting for rate of income over 40000

Table A.4. Estimated parameters of the exponential semivariogram model of rate of the wage or salary earner

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.004797 | 0.000163 | 0.004474 | 0.005121 |
| | sill | 0.013138 | 0.000290 | 0.012564 | 0.013713 |
| | range | 36.054854 | 2.991282 | 30.127969 | 41.981739 |
| weighted #1 | Nugget | 1.804646 | 0.059134 | 1.687479 | 1.921813 |
| | sill | 6.428906 | 0.177830 | 6.076555 | 6.781257 |
| | range | 45.175892 | 3.517550 | 38.206267 | 52.145516 |
| weighted #2 | Nugget | 0.862768 | 0.027338 | 0.808601 | 0.916936 |
| | sill | 2.978990 | 0.078712 | 2.823032 | 3.134948 |
| | range | 44.419550 | 3.386918 | 37.708757 | 51.130342 |

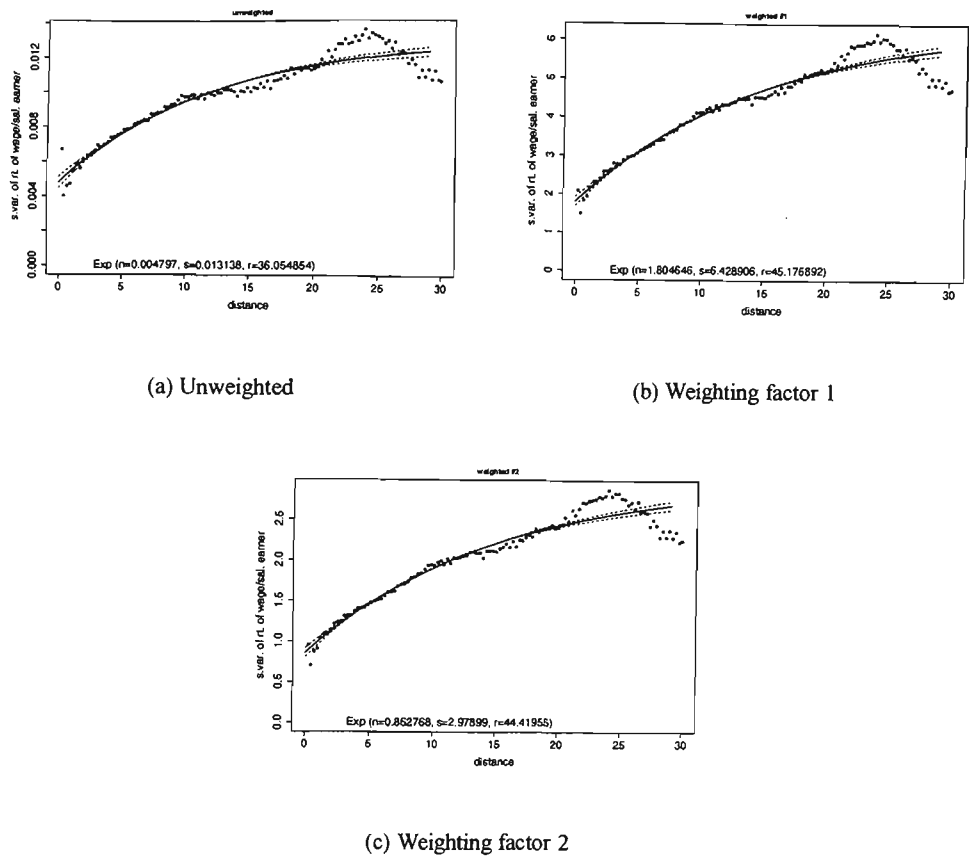


Figure A.12. Exponential semivariogram model fitting for rate of wage or salary earner

Table A.5. Estimated parameters of the exponential semivariogram model of rate of the self employed person

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000186 | 0.0000038 | 0.000178 | 0.000193 |
| | sill | 0.000300 | 0.0000014 | 0.000297 | 0.000302 |
| | range | 16.687985 | 0.9564958 | 14.792798 | 18.583173 |
| weighted #1 | Nugget | 0.075139 | 0.0010778 | 0.073003 | 0.077275 |
| | sill | 0.132648 | 0.0008590 | 0.130946 | 0.134350 |
| | range | 25.843623 | 1.2494484 | 23.367984 | 28.319263 |
| weighted #2 | Nugget | 0.036784 | 0.000509 | 0.0357761 | 0.037793 |
| | sill | 0.064320 | 0.000460 | 0.0634097 | 0.065231 |
| | range | 27.724176 | 1.418303 | 24.9139700 | 30.534382 |

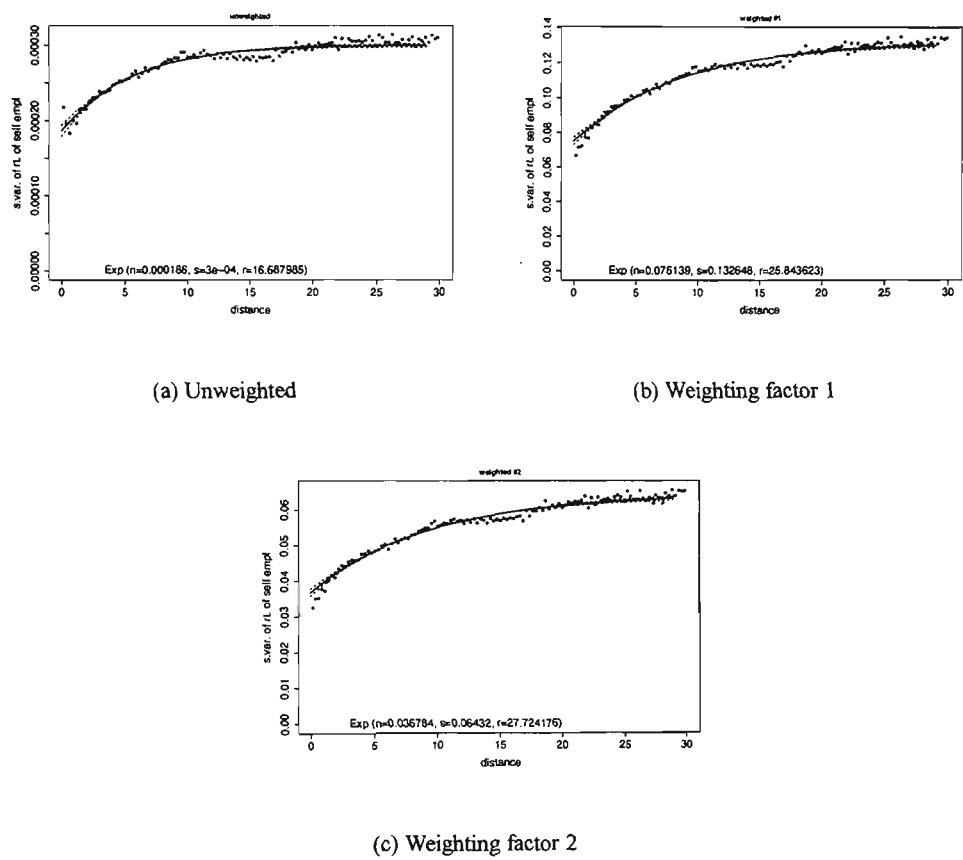


Figure A.13. Exponential semivariogram model fitting for rate of self employed

Table A.6. Estimated parameters of the exponential semivariogram model of rate of the employer

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|----------|----------------|-------------------------------------|----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000117 | 0.000019 | 0.000079 | 0.000154 |
| | sill | 0.000590 | 0.000003 | 0.000584 | 0.000595 |
| | range | 6.178353 | 0.339634 | 5.505406 | 6.851299 |
| weighted #1 | Nugget | 0.052682 | 0.004961 | 0.042852 | 0.062513 |
| | sill | 0.241481 | 0.000857 | 0.239784 | 0.243178 |
| | range | 7.480126 | 0.286174 | 6.913104 | 8.047148 |
| weighted #2 | Nugget | 0.028307 | 0.002219 | 0.023910 | 0.032704 |
| | sill | 0.118443 | 0.000422 | 0.117607 | 0.119279 |
| | range | 8.213822 | 0.299967 | 7.619471 | 8.808172 |

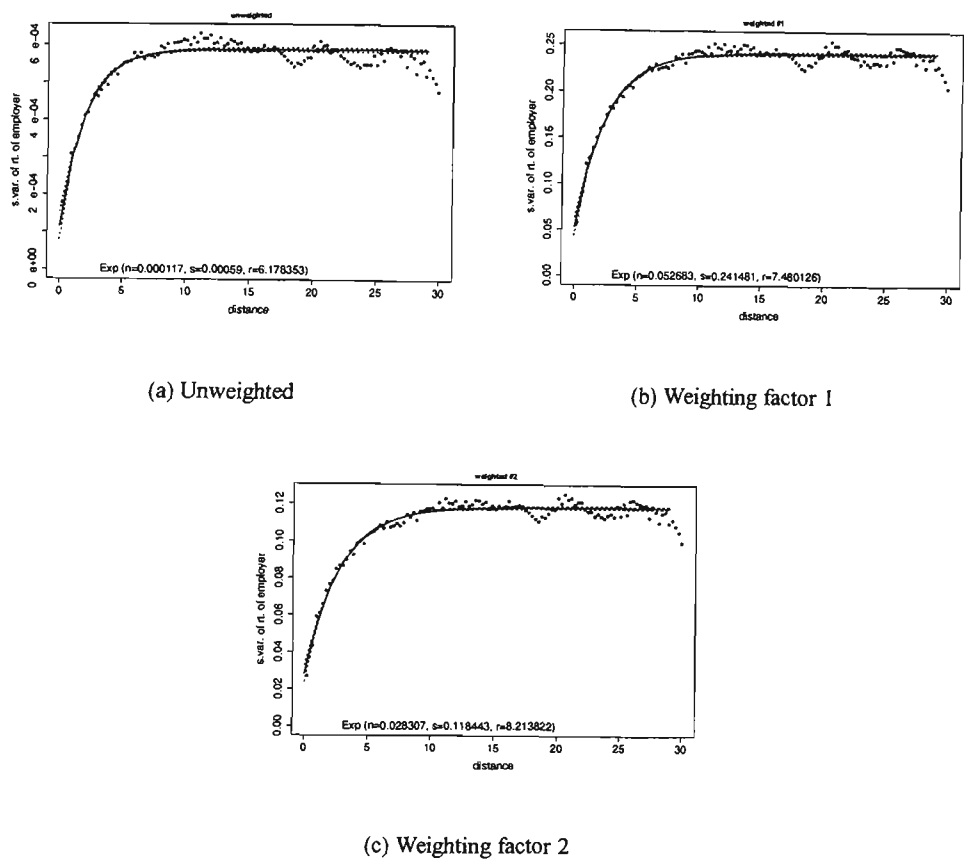


Figure A.14. Exponential semivariogram model fitting for rate of employer

Table A.7. Estimated parameters of the exponential semivariogram model of rate of the formal qualification

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000248 | 0.000118 | 0.000015 | 0.000482 |
| | sill | 0.008548 | 0.000079 | 0.008392 | 0.008705 |
| | range | 13.364826 | 0.522447 | 12.329657 | 14.399994 |
| weighted #1 | Nugget | 0.106191 | 0.040159 | 0.026621 | 0.185761 |
| | sill | 3.756959 | 0.033716 | 3.690155 | 3.823764 |
| | range | 15.237780 | 0.515335 | 14.216704 | 16.258857 |
| weighted #2 | Nugget | 0.043951 | 0.017944 | 0.008396 | 0.079506 |
| | sill | 1.825394 | 0.017006 | 1.791698 | 1.859090 |
| | range | 16.121832 | 0.533867 | 15.064035 | 17.179628 |

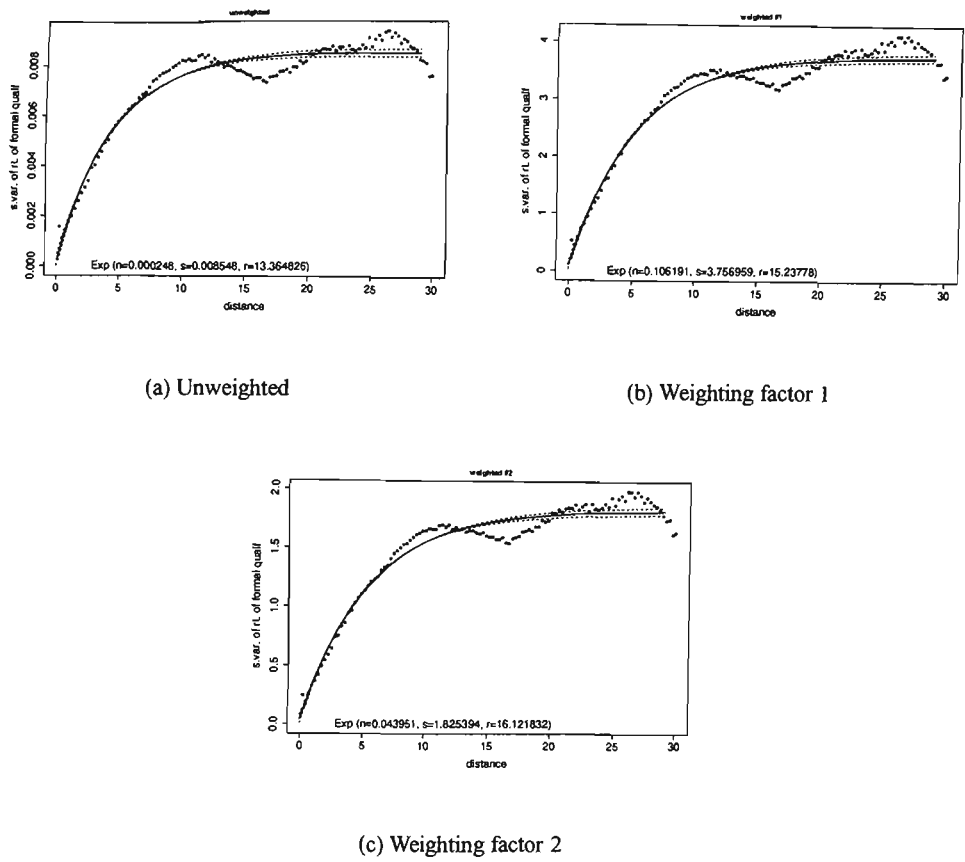


Figure A.15. Exponential semivariogram model fitting for formal qualification rate

Table A.8. Estimated parameters of the exponential semivariogram model of rate of the informal qualification

| Weighting | Parameter | Estimate | Standard Error | Asymptotic 95 % Confidence Interval | |
|-------------|-----------|-----------|----------------|-------------------------------------|-----------|
| | | | | Lower | Upper |
| Unweighted | Nugget | 0.000549 | 0.000021 | 0.000507 | 0.000591 |
| | sill | 0.002054 | 0.000040 | 0.001974 | 0.002134 |
| | range | 34.968376 | 2.209768 | 30.589973 | 39.346780 |
| weighted #1 | Nugget | 0.221561 | 0.010085 | 0.201578 | 0.241544 |
| | sill | 0.985302 | 0.029108 | 0.927629 | 1.042977 |
| | range | 42.302545 | 3.322150 | 35.720084 | 48.885006 |
| weighted #2 | Nugget | 0.103544 | 0.004588 | 0.094453 | 0.112636 |
| | sill | 0.456666 | 0.012151 | 0.432590 | 0.480743 |
| | range | 40.463875 | 2.946279 | 34.626160 | 46.301590 |

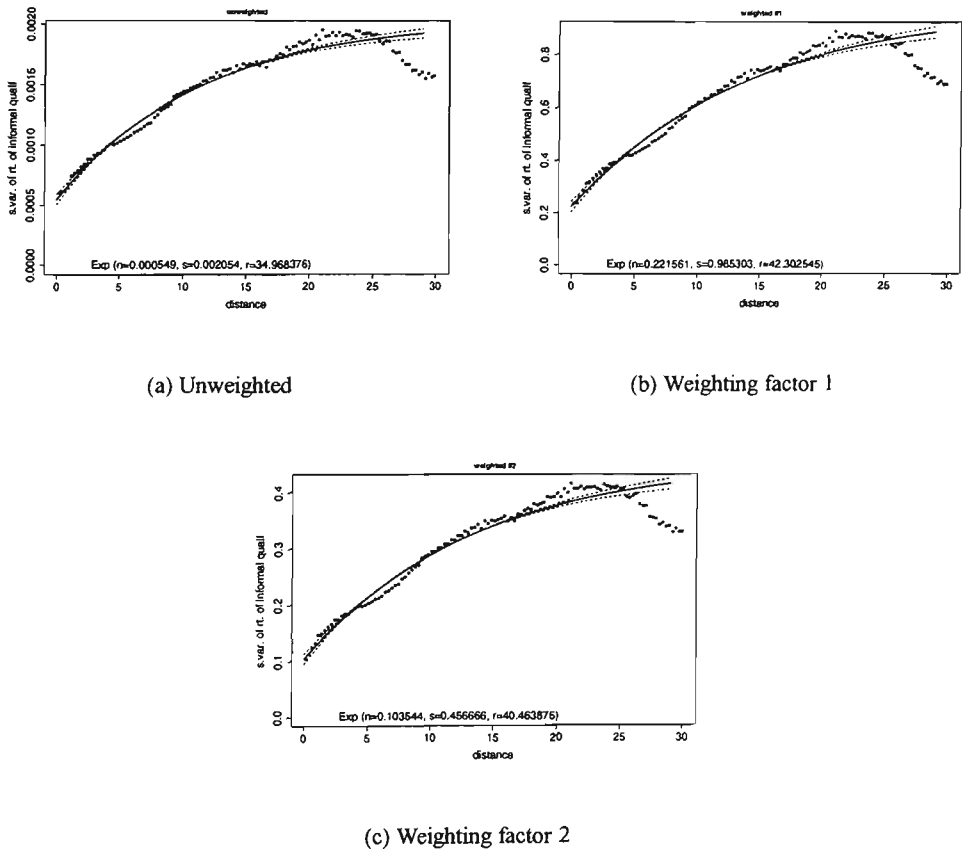


Figure A.16. Exponential semivariogram model fitting for the informal qualification rate

A.3 Tabulation of micro-sample data of the Adelaide region

Table A.9. The tabulation of the labor force status, income, nature of income, and qualification level

| LABOUR FORCE STATUS (LFSP) | Frequency | Pct | Cum. Freq | Cum. Pct. |
|---|-----------|---------|-----------|-----------|
| + wage or salary earner | 3856 | 37.8 | 3856 | 37.8 |
| + self employed | 351 | 3.4 | 4207 | 41.2 |
| + employer | 236 | 2.3 | 4443 | 43.5 |
| + unpaid helper | 29 | 0.3 | 4472 | 43.8 |
| EMPLOYMENT :..... | (4472) | (43.80) | | |
| + unemployed-looking for full time work | 484 | 4.7 | 4956 | 48.5 |
| + unemployed-looking for part time work | 108 | 1.1 | 5064 | 49.6 |
| UNEMPLOYMENT:..... | (592) | (5.80) | | |
| not in labour force aged 15+ | 3027 | 29.6 | 8091 | 79.2 |
| not stated + not applicable | 2129 | 20.7 | 10210 | 100.0 |
| INDIVIDUAL INCOME (INCP) | Frequency | Pct | Cum. Freq | Cum. Pct. |
| INCOME BELOW 20000 :..... | 4964 | 48.62 | 4964 | 48.62 |
| INCOME 20000-40000 :..... | 2140 | 20.96 | 7104 | 69.58 |
| INCOME OVER 40000 :..... | 466 | 4.56 | 7570 | 74.14 |
| not stated + not applicable | 2640 | 25.86 | 10210 | 100.0 |
| QUALIFICATION (HIGHEST) LEVEL (QLLP) | Frequency | Pct | Cum. Freq | Cum. Pct. |
| FORMAL QUALIFICATION :..... | 995 | 9.75 | 995 | 9.75 |
| INFORMAL QUALIFICATION :..... | 1100 | 10.77 | 2095 | 20.52 |
| inadequately desc.+not stated+not applic. | 8115 | 79.48 | 10210 | 100.0 |

Appendix B

Longitude & Latitude Conversion

Introduction

This section intend to give a brief description of transformation of the coordinate system from the latitude longitude into the Cartesian system. This appendix related with the empirical work in chapter (8).

The Australian Census 1991 data provided the longitude and latitude of each CD's centroid. The geographical characteristics may be incorporated into the analysis of the data.

Most of the computation in spatial analysis is done using a Cartesian system, such as Cartesian distance. Therefore we need to convert the longitude latitude coordinate system into Cartesian coordinate system. Further discussion of the coordinate system may be found in Maling (1992).

The earth has an irregular shape. It can be approached by a theoretical shape, such as spherical, biaxial ellipsoid, triaxial ellipsoid, and ovaloid (Rajagopalan, 1992). Conversion of the geographical coordinate from one system to another should be done carefully and defined specifically according to their geographic location. Some detail of the procedures and convention of the conversion may be found in Bonham-Carter (1994). For the Australian latitude longitude data, the conversion will follow the Australian 1965 convention. The Australian 1965 convention put some geometric parameter of the earth, those are major axes $(a) = 6378.160$ km and minor axes $(b) = 6356.7747$ km.

The Cartesian coordinate system may be computed from the longitude latitude data by calculating a distance between a fixed point and the current point. The fixed point is defined arbitrarily. Maling (1992) and Bonham-Carter (1994) defined the distance between two points as,

$$s = R \cdot \cos^{-1}(\sin(\varphi_a) \cdot \sin(\varphi_b) + \cos(\varphi_a) \cdot \cos(\varphi_b) \cdot \cos(\Delta_\lambda)) \quad (\text{B.1})$$

where, R is the earth's radius, φ_a and φ_b are the latitude coordinates of the two points, and Δ_λ is the longitude difference between the two points longitude, $(\lambda_a - \lambda_b)$. The conversion gives an approximation that one second of latitude is about 30 meters, or one degree is approximately 110 km (Bonham-Carter, 1994).

Maling (1992) noted that for the practical application of the conversion, it is sufficient to take a radius of the earth to be 6371.2 km. But to be more precise, the R may be computed by taking a reference to its latitude points, that is

$$R = \frac{\rho + \nu}{2} \quad (\text{B.2})$$

where,

$$\rho = \frac{a \cdot (1 - e^2)}{(1 - e^2 \cdot \sin^2(\varphi))^{3/2}} \quad (\text{B.3})$$

$$\nu = \frac{a}{(1 - e^2 \cdot \sin^2(\varphi))^{1/2}} \quad (\text{B.4})$$

and, $e^2 = \frac{a^2 - b^2}{a^2}$

The implementation of the above procedure has been written in S-Plus. The S-Plus procedures may be found in the following,

```
1 ## read a raw data file :
2 ##
3 adel<-read.table('adelx.prn',header=T)
4 ## INPUT :
5 ## header contains :
6 #      cdid lga
7 #      dpc  km2 lon lat
8 ## notes :
9 ##      dpc      : post code
10 ##      km2      : area in km square
11 ##      lon lat  : longitude and latitude
12 #####
13 ## OUTPUT :
14 ##      easting  : x distance (km) from arbitrary point.
15 ##      northing : y distance (km) from arbitrary point.
16 #####
17 ## Some constants to calculate the earth's radius,
```



```

18 ## base on Australian 1965 convention, (Bonham-Carter,1994).
19 ## Those constants are a and b ( in metres )
20 acoef <- 6378160
21 bcoef <- 6356774.7
22 ecoef2 <- (acoef^2 - bcoef^2) / (acoef^2)
23 #####
24 basep<-adel$z0105+adel$z0106
25 numcd <-length(adel$lon)
26 lon.a <- 138.46730
27 lat.a <- -35.21738
28 earthc <- 6371.2
29 #####
30 ## Note : earthc is a radius the earth in km (Maling, 1992).
31 ## Ref : Maling, D.H. (1992) Coordinate System and Map Projections
32 ## Pergamon press. England. pg 5
33 #####
34 lon1 <- rep(lon.a,numcd)
35 lat1 <- rep(lat.a,numcd)
36 lon2 <- adel$lon
37 lat2 <- rep(lat.a,numcd)
38 lon3 <- rep(lon.a,numcd)
39 lat3 <- adel$lat
40 easting <- rep(0,numcd)
41 northing <- rep(0,numcd)
42 earthu <- 0
43 for (i in 1:numcd) {
44   # calculate the earth's radius :
45   p <- (acoef*(1-ecoef2))/(1-ecoef2*(sin(lat3[i]))^2)^(1.5)
46   v <- (acoef) / (1-ecoef2*(sin(lat3[i]))^2)^(0.5)
47   R <- (p+v)/2
48   earthu<-(earthu+R)
49   # converting the longitude and latitude :
50   ab <- ((90 - lat1[i]) * pi) / 180
51   ac <- ((90 - lat2[i]) * pi) / 180
52   bc <- abs((lon1[i]-lon2[i])*pi)/180
53   easting[i] <- acos ( cos(ab)*cos(ac) +
54                       sin(ab)*sin(ac)*cos(bc) ) * (R/1000)
55   ac <- ((90-lat3[i])*pi)/180
56   bc <- abs((lon1[i]-lon3[i])*pi) /180
57   northing[i] <- acos ( cos(ab)*cos(ac) +
58                       sin(ab)*sin(ac)*cos(bc) ) * (R/1000)
59 }
60 earthu <- earthu/numcd
61 cat("Average R : ",earthu,"\n")

```

Appendix C

Evaluation of the Probability density function of the random distance within the region

The main discussion of this matter can be found in section (5.3.2). These procedures are intended to compute analytically the mean and variance of distance within the group of several shapes.

- The rectangle with length of the sides are defined as a and b ($a \geq b$). The w is a ratio of a/b .

```
1 ## w is a ration of a/b, A is area
2 ## Square : A=1 w=1
3 ## Rectangle : A=1 w=2 --> b=1/2 a
4 #####
5 a:=sqrt(A*w);
6 b:=sqrt(A/w);
7 fr:=4*R/(a^2*b^2);
8 fr1:=(0.5*Pi*a*b-a*R-b*R+0.5*R^2)*fr;
9 l2:=b;
10 fr2:=(a*b*arcsin(b/R)+a*sqrt(R^2-b^2)-a*R-0.5*b^2)*fr;
11 l3:=a;
12 fr3:=( a*b*(arcsin(b/R)-arccos(a/R))
13      +a*sqrt(R^2-b^2)+b*sqrt(R^2-a^2)
14      -0.5*(R^2+a^2+b^2) )*fr;
15 l4:=sqrt(a^2+b^2);
16 pdf:=int(fr1,R=0..l2)+int(fr2,R=l2..l3)+int(fr3,R=l3..l4);
17 Ed:=int(R*fr1,R=0..l2)+int(R*fr2,R=l2..l3)+int(R*fr3,R=l3..l4);
18 Ed2:=int(R^2*fr1,R=0..l2)+int(R^2*fr2,R=l2..l3)+int(R^2*fr3,R=l3..l4);
19 V:=Ed2-Ed^2;
20 P:=2*a+2*b;
```

- The circle with radius R .

```
1 fc := 4*d/A*(arccos(Pi*d/P)-1/2*Pi*d/P*(4-d^2*Pi/A)^(1/2));
2 P := 2*sqrt(Pi*A);
3 pdf:=int(fc,d=0..P/Pi);
4 Ed:=int(d*fc,d=0..P/Pi);
5 Ed2:=int(d^2*fc,d=0..P/Pi);
6 V:=Ed2-Ed^2;
```

- Regular hexagon with side length s .

```
1 sq3:=sqrt(3);
2 hx := 4*w/(9*sq3);
3 hx1 := hx*( 3*Pi-4*w*sq3+w^2*(sq3-Pi/3) );
4 hx2 := hx*( Pi*(5+w^2)-3*sqrt(12*w^2-9)-(4*w^2+6)*arcsin(sq3/(2*w)) );
5 hx3 := hx*( (2*w^2+24)*(arcsin(sq3/w)-Pi/3)-sq3*(w^2+6)+10*sqrt(3*w^2-9) );
6 Af:=sqrt(3/2 * sq3/A);
7 w:= v* Af;
8 s:=A/3 *sqrt(2*sqrt(3));
9 #pdf:=int(hx1*Af,w=0..1)+int(hx2,w=1..sq3)+int(hx3,w=sq3..2);
10 pdf:=int(hx1*Af,v=0..1/Af)+int(hx2*Af,v=1/Af..sq3/Af)+int(hx3*Af,v=sq3/Af..2/Af);
```

```
11 Ed:=int(v*hx1*Af,v=0..1/Af)+int(v*hx2*Af,v=1/Af..sq3/Af)+int(v*hx3*Af,v=sq3/Af..2/Af);
12 Ed2:=int(v^2*hx1*Af,v=0..1/Af)+int(v^2*hx2*Af,v=1/Af..sq3/Af)+int(v^2*hx3*Af,v=sq3/Af..2/Af);
13 V:=Ed2-Ed^2;
14 P:=6*s;
```

• The equilateral triangle with side length s .

```
1 sq3 := sqrt(3);
2 ft := 8/sq3 * w;
3 ft1 := ft * (Pi-4*w*sq3+w^2*(sq3+2*Pi/3));
4 ft2 := ft * (3*sqrt(12*w^2-9)-(4*w^2+6)*arccos(sq3/(2 * w) ));
5 w:=v/s;
6 s:=2*sqrt(A/sq3);
7 #A:=1;
8 pdf:=int(ft1*1/s,v=0..s)+int(ft2*1/s,v=sq3*s/2..s);
9 #pdf1:=int(ft1,v=0..s)+int(ft2,v=sq3*s/2..s);
10 Ed:=int(v*ft1*1/s,v=0..s)+int(v*ft2*1/s,v=sq3*s/2..s);
11 Ed2:=int(v^2*ft1*1/s,v=0..s)+int(v^2*ft2*1/s,v=sq3*s/2..s);
12 V:=Ed2-Ed^2;
13 P:=3*s;
14 #pdf:=int(ft1*s,w=0..1)+int(ft2*s,w=sq3/2..1);
```

Appendix D

SAS codes for semivariogram analysis of Illawarra data

The main discussion of this SAS procedure is in section (5.1.4). This procedure is intended for estimation of the semivariogram or cross-semivariogram model parameters.

```
1  data woll;
2  infile 'woll.dat';
3  input easting northing laborr1; run;
4  option ls=65 ps=75;
5  proc variogram data=woll outv=outv;
6  compute lagd=0.6 maxlag=50;
7  coordinates xc=easting yc=northing;
8  var laborr1; run;
9  data outv2; set outv;
10 vari=variog; type='regular'; output; run;
11 symbol1 i=join l=1 v=star c=black;
12 axis1 minor=none label=(c=black 'distance') offset=(3,3);
13 axis2 minor=(number=1) label=(c=black 'semivariogram') offset=(3,3);
14 filename grafout 'illsv1.eps'; goptions dev=PSLEPSF
15 gaccess=sasgaedt gsfname=grafout gsfmode=replace gsflen=60 border
16 hsize=7 vsize=4;
17 proc gplot data=outv2;
18 plot vari*distance /vaxis=axis2 haxis=axis1;
19 run;
20 *-----;
21 %let ic0= 30.0;
22 %let ics= 60.0;
23 %let ias= 5.0;
24 options ls=75 ps=65;
25 *****;
26 ** read data file and define dataset to be used; ** Note : INPUT
27 statement may change depend on the input file;
28 ** variogram model fitting : EXPONENTIAL MODEL;
29 data fits; set outv2;
30 npair = count;
31 avgdist=distance;
32 sv=variog; sva=variog;
33 index=1; run;
34 proc nlin data=fits method=marquardt maxiter=200 smethod=golden;
35 parameters c0=&ic0 cs=&ics as=&ias; if index=1 then
36 f=c0 + (cs-c0)*(1-exp(-3*avgdist/as));
37 if index=0 then
38 f=cs*(1-exp(-3*avgdist/as));
39 _weight_=(npair)/(f**2); model sv=f;
40 bounds cs>0, as>0;
41 der.cs=0;der.as=0;
42 if index=1 then do;
43 bounds c0>0;
44 der.c0=0;
45 der.c0=1-(1-exp(-3*avgdist/as));
46 end;
47 if index=1 then do;
48 der.cs=1-exp(-3*avgdist/as);
49 der.as=-(cs-c0)*3*avgdist*exp(-3*avgdist/as)/(as**2);
50 end;
51 if index=0 then do;
52 der.cs=1-exp(-3*avgdist/as);
53 der.as=-cs*3*avgdist*exp(-1*avgdist/as)/(as**2);
54 end;
```

```
55 output out=svf p=svfit r=svres
56 parms= c0 cs as;
57 run;
58 data outv3;
59 set outv;
60 set svf;
61 vari=variog; type='empirical'; output;
62 vari=c0 + (cs-c0)*(1-exp(-3*distance/as));
63 type='Exp'; output;
64 run;
65 symbol1 i=join l=1 v=star c=black;
66 symbol2 i=join l=1 v=square c=black;
67 filename grafout 'illsv2.eps';
68 goptions dev=PSLEPSF gaccess=sasgaedt gsfname=grafout
69 gsfname=replace gsflen=60 border hsize=7 vsize=4;
70 proc gplot data=outv3;
71 plot vari*distance=type /vaxis=axis2 haxis=axis1; run;
```

Appendix E

Non-linear procedure for estimating n, s, and r

This subroutine is used in section (5.4). They are a combination of Fortran program for generating the individual population, SAS procedures for estimating the parameters, and S plus for creating some graphs.

The Fortran program has three task, generating individual level observations, grouping the population, and calculating individual level and group level semivariogram. The SAS procedure is used for estimating individual level semivariogram parameters by non-linear method as discussed in section (5.4).

E.1 Fortran program

```
1 cc=====
2 cc filename      : vsim.for
3 cc author        : gandhi pawitan
4 cc date          : april 99
5 cc version       : a,b
6 cc address       : school of mathematics and applied stat.
7 cc               : university of wollongong
8 cc description   : the program will generate individuals
9 cc               : data values for a specified semivariogram
10 cc              : models.
11 cc              : The semivariogram models could be
12 cc              : a spherical, exponential, or gaussian.
13 cc              : There are three common parameters, nugget
14 cc              : sill, and range.
15 cc Note : some parameters are setted up by parameter file
16 cc       n, xmin,xmax,ymin,ymax,nug,sill,range,dinterval
17       program vsim
18       parameter (nbin=500)
19       real*8 x(2000),y(2000),z(2000),w(2000),c(2000,2000)
20       real*8 xmin,xmax,xmini,ymini,ymaxi,dmaxi,dmini,drange
21       real*8 xmin,xmax,xmean,xvar,ymin,ymax,ymean,yvar
22       real*8 zmin,zmax,zmean,zvar,zmin0,zmax0,zmean0,zvar0
23       real*8 nug,sill,range,dinterval, inug,isill,irange
24       real*8 h(nbin),sv(nbin),avgd(nbin),svmodel(nbin)
25       real*8 ivar,pct
26       integer f(nbin),nbin0,n,nobs
27       logical testfl,lbegin, testrun
28       character str*80
29       parameter (nbin1=200)
30       integer nbinu0,nbinw0,fuint(nbin1)
31       real*8 hu(nbin1),svu(nbin1),avgdu(nbin1),svumodel(nbin1)
32       real*8 hw(nbin1),fw(nbin1),svw(nbin1),avgdw(nbin1),svwmodel(nbin1)
33       real*8 dgmaxi,dgmini,dgrange,region,xgmean,xgvar,xgmini,xgmaxi
34       real*8 ygmean,ygvar,ygmini,ygmaxi
35       real*8 zgmean,zgvar,zgmini,zgmaxi,zgvarw
36       real*8 ngmean,ngvar,ngmini,ngmaxi
37       real*8 agmean,agvar,agmini,agmaxi
38       real*8 pgmean,pgvar,pgmini,pgmaxi
```

```

39     real*8 svumean,svuvar,svumini,svumaxi
40     real*8 svwmean,svwvar,svwmini,svwmaxi
41     parameter(maxgroup=2001)
42     integer ngroup, xdim, ydim
43     real*8 xgr(maxgroup),ygr(maxgroup),zgr(maxgroup),agr(maxgroup)
44     real*8 xc(maxgroup),yc(maxgroup),ngr(maxgroup),pgr(maxgroup)
45     real*8 wght(maxgroup),sillbar
46     real*8 n01,s01,r01,n02,s02,r02
47     real*8 inugu,isillu,irangeu,inugw,isillw,irangew
48 c read parameter file :
49     inquire(file='vsim.run',exist=testrun)
50     if(.not.testrun) then
51         str='vsim.par'
52         write(*,fmt='(a,$)')'?> Par file (vsim.par) :'
53         read(*,'(a20)')str(1:20)
54         if(str(1:1).eq.' ') str(1:20)='vsim.par'
55         inquire(file=str(1:20),exist=testfl)
56         if(.not.testfl) then
57             write(*,*)'*** Parameter file was not found ***'
58             stop
59         endif
60     else
61         str='vsim.run'
62     endif
63     open(8,file=str(1:20),status='OLD')
64     rewind(8)
65     lbegin=.FALSE.
66     dowhile (.not.lbegin)
67         read(8,'(a4)',end=97)str(1:4)
68         if(str(1:4).eq.'BEGI') lbegin=.TRUE.
69     enddo
70
71     read(8,*,end=97)n
72     read(8,*,end=97)xmini,xmaxi
73     read(8,*,end=97)ymini,ymaxi
74     read(8,*,end=97)nug,sill,range
75     read(8,*,end=97)dinterval
76     read(8,*,end=97)xdim,ydim
77
78     print '(a,i5)', '!>Number of data:',n
79     print '(a,2f9.3)', '!>X-min and X-max:',xmini,xmaxi
80     print '(a,2f9.3)', '!>Y-min and Y-max:',ymini,ymaxi
81     print '(a,3f9.3)', '!>Nugget, sill, and range:',nug,sill,range
82     print '(a,f9.3)', '!>Interval of distance:',dinterval
83     print '(a,2i5)', '!>Xdim and ydim:',xdim,ydim
84     close(8)
85 c start to simulate :
86     nobsn=n
87     print '(a)', '!>.....'
88     call xyrectangle(x,y,n,xmini,xmaxi,ymini,ymaxi)
89     call znorm(z,n)
90 c find performance of the generated data :
91     call statistics(x,n,xmean,xvar,xmin,xmax)
92     call statistics(y,n,ymean,yvar,ymin,ymax)
93     call statistics(z,n,zmean0,zvar0,zmin0,zmax0)
94     dmaxi=sqrt((xmin-xmax)**2+(ymin-ymax)**2)
95     print '(a,2f9.3)', '!>The original Z -- N(0,1):',zmean0,zvar0
96 c compute the V matrix :
97     call initvarcov(c,n)
98     call varcov(c,x,y,n,nug,sill,range,dmini,dmaxi)
99     drange=dmaxi-dmini
100 c choleski decomposition and find the Z :
101     call choles(c,n)
102     call mult(c,z,n)
103     call statistics(z,n,zmean,zvar,zmin,zmax)
104     print '(a,2f9.3)', '!>The generated Z :',zmean,zvar
105     print '(a,3f9.3)', '!>Distance (min,max,range):',
106         * dmini,dmaxi,drange
107 c write the generated data values to the file outvsim.xyz
108     call writexyz(n,x,y,z)
109 c take a sample of individuals of 10% population size
110     pct=0.10
111     call sample(pct,n,z,ivar)
112     print '(a,f9.3)', '!>Individual sample variance (Z):',ivar
113 c compute semivariogram :
114     nbin0=nbin
115     call semivar(h,f,sv,avgd,nbin0,x,y,z,n,dinterval,50.)
116     do i=1,nbin0
117         svmodel(i)=nug+(sill-nug)*(1.d0-dexp(-3.d0*avgd(i)/range))
118 c
119         w(i)=1.0/f(i)
120         w(i)=f(i)

```

```

120      enddo
121 c determine the initial value of nug, sill, and range.
122 c Initial value of the sill = zvar (individual pop. variance)
123      isill=zvar
124      call initstv(sv,avgd,nbin0,inug,isill,irange)
125      print '(a,3f9.3)', '!>Initial value :', inug, isill, irange
126 c write the semivariogram to the file outvsim.csv :
127      call writesv(nbin0,h,f,avgd,sv,svmodel)
128 c outputting the result
129      call desc(nobs,nug,sill,range,xmini,xmaxi,ymini,ymaxi,
130      *      dmini,dmaxi,xmin,xmax,xmean,xvar,ymin,ymax,ymean,yvar,
131      *      zmin0,zmax0,zmean0,zvar0,zmin,zmax,zmean,zvar,
132      *      nbin0,avgd,sv,f,h,inug,isill,irange)
133 c
134 CCCCCCCCCC=== GROUP LEVEL DATA DEFINITION ===CCCCCCCCCCCCCCCC
135 c calculate area of the region
136      region=dbl( (xmax-xmin) * (ymax-ymin) )
137      ngroun=xdim*ydim
138      call vgroup(x,y,z,n,xgr,ygr,zgr,ngr,agr,pgr,xc,yc,
139      *      ngroun,xdim,ydim,xmini,xmaxi,ymini,ymaxi)
140 c save the group data :
141      print '(a)', '!>===== '
142      print '(a)', '!>writing group level data ... '
143      call writexyzg(ngroun,ngr,agr,xgr,ygr,zgr,xc,yc)
144      print '(a,i5)', '!>Number of group Ng>0 :', ngroun
145 c calculate statistics
146      call statistics(xgr,ngroun,xgmean,xgvar,xgmini,xgmaxi)
147      call statistics(ygr,ngroun,ygmean,ygvar,ygmini,ygmaxi)
148      call statistics(zgr,ngroun,zgmean,zgvar,zgmini,zgmaxi)
149      call statistics(ngr,ngroun,ngmean,ngvar,ngmini,ngmaxi)
150      call statistics(agr,ngroun,agmean,agvar,agmini,agmaxi)
151      call statistics(pgr,ngroun,pgmean,pgvar,pgmini,pgmaxi)
152      do i=1,ngroun
153          wght(i)=1.0
154      enddo
155      call var(zgr,ngr,ngroun,zgvarw)
156      call var(zgr,wght,ngroun,zgvar)
157      call dminmax(xgr,ygr,ngroun,dgmini,dgmaxi)
158      dgrange=dgmaxi-dgmini
159 c print out the result
160      print '(a,f9.3)', '!>Unweighted group level variance :', zgvar
161      print '(a,f9.3)', '!>Weighted group level variance :', zgvarw
162      print '(a,3f9.3)', '!>Min, max, and range distance :', dgmini,
163      *      dgmaxi,dgrange
164      print '(a,f9.3)', '!>Average group size:', ngmean
165      print '(a,f9.3)', '!>Average group areal:', agmean
166 c calculate semivariogram unit
167 c the output will be save in : outunit.svg
168      call calcsv(ngr,agr,xgr,ygr,zgr,xc,yc,zgvarw,sillbar,ngroun)
169      print '(a,f9.3)', '!>Average sill :', sillbar
170 c compute initial value of nug, sill, and range
171      call methods(n,ngroun,ngmean,ivar,zgvarw,
172      *      agmean,sillbar,n01,s01,r01,n02,s02,r02)
173 c compute semivariogram :
174 c unweighted
175      nbinu0=nbin1
176      call semivar(hu,fuint,svu,avgd,nbinu0,xgr,ygr,zgr,
177      *      ngroun,dinterval,50.)
178      open(22,file='outvbiasu.svg')
179 c      write(22,*) '      h      npair      avgd      sv '
180      do i=1,nbinu0
181 c          svumodel(i)=vsvexp(nug,sill,range,avgd(i),0)
182          write(22,'(f10.3,i10,3f10.3)')hu(i),fuint(i),avgd(i),svu(i)
183      enddo
184      close(22)
185 c find the initial value of nug, sill, and range
186      irangeu= 0.51082 * sqrt(region) * 0.5
187      isillu= zgvar
188      call statistics(svu,nbinu0,svumean,svuvar,svumini,svumaxi)
189      inugu=svumaxi
190      do i=1,nbinu0
191          if(avgd(i).lt.irangeu) then
192              if(svu(i).lt.inugu) inugu=svu(i)
193          endif
194      enddo
195      open(33,file='svfitgu.par')
196      write(33,'(a,f9.3,a)') '%let ic0 = ', inugu, ';'
197      write(33,'(a,f9.3,a)') '%let ics = ', isillu, ';'
198      write(33,'(a,f9.3,a)') '%let ias = ', irangeu, ';'
199      close(33)
200 c weighted

```



```

201     nbinw0=nbin1
202     call gsemivar(hw,fw,svw,avgdw,nbinw0,xgr,ygr,zgr,ngr,
203 *       ngroup,dinterval,50.)
204     open(22,file='outvbiasw.svg')
205 c   write(22,*)'      h      npair      avgd      sv      svmodel'
206     do i=1,nbinw0
207         write(22,'(5f10.3)')hw(i),fw(i),avgdw(i),svw(i)
208     enddo
209     close(22)
210 c find the initial value of nug, sill, and range
211     irangew= 0.51082 * sqrt(region) * 0.5
212     isillw= zgvarw
213     call statistics(svw,nbinw0,svwmean,svwvar,svwmini,svwmaxi)
214     inugw=svwmaxi
215     do i=1,nbinw0
216         if(avgdw(i).lt.irangew) then
217             if(svw(i).lt.inugw) inugw=svw(i)
218         endif
219     enddo
220     open(33,file='svfitgw.par')
221     write(33,'(a,f9.3,a)')'%let ic0 = ',inugw',';'
222     write(33,'(a,f9.3,a)')'%let ics = ',isillw',';'
223     write(33,'(a,f9.3,a)')'%let ias = ',irangew',';'
224     close(33)
225     call descg(n,xmini,xmaxi,ymini,ymaxi,nug,sill,range,
226 *       ngroup,agmean,dgmini,dgmaxi,dgrange,xdim,ydim,
227 *       xgmini,xgmaxi,xgmean,xgvar,ygmini,ygmaxi,ygmean,ygvar,
228 *       zgmini,zgmaxi,zgmean,zgvar,zgvarw,ngmini,ngmaxi,ngmean,ngvar,
229 *       nbinu0,svu,avgdu,fuint,hu,nbinw0,svw,avgdw,fw,hw,
230 *       inugw,isillw,irange,inugu,isillu,irangeu)
231     stop
232 97 stop '***Error : in parameter file ***'
233     end
234 c end of main program
235 ccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
236 c subroutine definition:
237     subroutine writexyz(n,x,y,z)
238     integer n
239     real*8 x(n),y(n),z(n)
240     open(11,file='outvsim.xyz')
241     write(11,*)'x y z'
242     do i=1,n
243         write(11,'(3f8.3)')x(i),y(i),z(i)
244     enddo
245     close(11)
246     return
247     end
248 ccccccccccccccccccc
249     subroutine writesv(nbin0,h,f,avgd,sv,svmodel)
250     integer nbin0, f(nbin0)
251     real*8 h(nbin0),avgd(nbin0),sv(nbin0),svmodel(nbin0)
252     open(22,file='outvsim.svg')
253 c   write(22,*)'h npair avgd sv svmodel'
254     do i=1,nbin0
255         write(22,'(f9.3,i13,3f9.3)')h(i),f(i),avgd(i),sv(i),svmodel(i)
256     enddo
257     close(22)
258     return
259     end
260 ccccccccccccccccccc
261     subroutine writexyzg(ngroup,ngr,agr,xgr,ygr,zgr,xc,yc)
262     integer ngroup,ngroup0
263     real*8 ngr(ngroup),agr(ngroup),xgr(ngroup),ygr(ngroup)
264     real*8 zgr(ngroup),xc(ngroup),yc(ngroup)
265     real*8 ngr0(ngroup),agr0(ngroup),xgr0(ngroup),ygr0(ngroup)
266     real*8 zgr0(ngroup),xc0(ngroup),yc0(ngroup)
267 c
268     ngroup0=0
269     open(11,file='outvbias.xyz')
270     write(11,*)'Ng Ag Xg Yg Zg Xc Yc'
271     do i=1,ngroup
272         if(ngr(i).gt.0.0) then
273             ngroup0=ngroup0+1
274             ngr0(ngroup0)=ngr(i)
275             agr0(ngroup0)=agr(i)
276             xgr0(ngroup0)=xgr(i)
277             ygr0(ngroup0)=ygr(i)
278             zgr0(ngroup0)=zgr(i)
279             xc0(ngroup0)=xc(i)
280             yc0(ngroup0)=yc(i)
281             write(11,'(2f8.3,5f10.5)')ngr(i),agr(i),xgr(i),ygr(i),zgr(i),

```

```

282      *      xc(i),yc(i)
283      endif
284      enddo
285      close(11)
286      ngroup=ngroup0
287      do i=1,ngroup
288          ngr(i)=ngr0(i)
289          agr(i)=agr0(i)
290          xgr(i)=xgr0(i)
291          ygr(i)=ygr0(i)
292          zgr(i)=zgr0(i)
293          xc(i)=xc0(i)
294          yc(i)=yc0(i)
295      enddo
296      return
297      end
298      cccccccccccccccccccccccccc
299      subroutine sample(pct,nindi,z,ivar)
300      integer nindi,nsample, ismp(1000)
301      real*8 z(nindi),ivar,zsample(1000),wght(1000)
302      real*8 pct,sv(30)
303      nsample=int(pct * nindi)
304      ivar=0.0
305      do j=1,30
306          call iran(ismp,nsample,nindi)
307          do i=1,nsample
308              zsample(i)=z(ismp(i))
309              wght(i)=1.0
310          enddo
311          call var(zsample,wght,nsample,sv(j))
312          ivar=sv(j)+ivar
313      enddo
314      ivar=ivar/30.0
315      return
316      end
317      cccccccccccccccccccccccccc
318      subroutine initsv(sv,avgd,nbin0,inug,isill,irange)
319      real*8 sv(nbin0),avgd(nbin0),inug,isill,irange
320      integer nbin0, idist
321      real*8 svmini,svmaxi,svmean,svvar
322      call statistics(sv,nbin0,svmean,svvar,svmini,svmaxi)
323      inug=svmaxi
324      irange=0.0
325      idist=0
326      do i=1,nbin0
327          if(sv(i).gt.isill.and.idist.eq.0) then
328              irange=avgd(i-1)
329              idist=1
330          endif
331      enddo
332      do i=1,nbin0
333          if(avgd(i).lt.irange) then
334              if(sv(i).lt.inug) inug=sv(i)
335          endif
336      enddo
337      open(33,file='svfit1.par')
338      write(33,'(3a)')"%let bfinput='outvsim.sv';"
339      write(33,'(a,f9.3,a)')"%let ic0='inug,';"
340      write(33,'(a,f9.3,a)')"%let ics='isill,';"
341      write(33,'(a,f9.3,a)')"%let ias='irange,';"
342      close(33)
343      return
344      end
345      cccccccccccccccccccccccccc
346      subroutine desc(nobs,nug,sill,range,xmini,xmaxi,ymini,ymaxi,
347      *      dmini,dmaxi,xmin,xmax,xmean,xvar,ymin,ymax,ymean,yvar,
348      *      zmin0,zmax0,zmean0,zvar0,zmin,zmax,zmean,zvar,
349      *      nbin0,avgd,sv,f,h,inug,isill,irange)
350      integer nobs,nbin0, f(nbin0)
351      real*8 nug,sill,range,xmini,xmaxi,ymini,ymaxi,
352      *      dmini,dmaxi,xmin,xmax,xmean,xvar,ymin,ymax,ymean,yvar,
353      *      zmin0,zmax0,zmean0,zvar0,zmin,zmax,zmean,zvar,
354      *      avgd(nbin0),sv(nbin0),h(nbin0),
355      *      inug,isill,irange
356      open(33,file='outvsim.s')
357      call header(33)
358      write(33,'(a)')'# Semivariogram model : n + s Exp(r)'
359      write(33,'(a,3f8.3)')'# n,s,r : ',nug,sill,range
360      write(33,'(a,i6)')'# Number of points : ',nobs
361      write(33,'(a)')'# Region shape : rectangle'
362      write(33,'(a,2f8.3)')'# X-min,X-max : ',xmini,xmaxi

```

```

363 write(33,'(a,2f8.3)')'# Y-min,Y-max      : ',ymini,ymaxi
364 write(33,'(a,3f8.3)')'# dmin,dmax,drange : ',dmini,dmaxi,
365 *   dmaxi-dmini
366 write(33,'(a)')'#-----'
367 write(33,'(a)')'#Performance of the generated points : '
368 write(33,'(a)')'#           Min.      Max.      Mean  Vari.'
369 write(33,'(a,4f8.3)')', '#x ',xmin,xmax,xmean,xvar
370 write(33,'(a,4f8.3)')', '#y ',ymin,ymax,ymean,yvar
371 write(33,'(a,4f8.3)')', '#Z0 ',zmin0,zmax0,zmean0,zvar0
372 write(33,'(a,4f8.3)')', '#Z1 ',zmin,zmax,zmean,zvar
373 write(33,'(a,3f8.3)')'#Initial value n,s,r: ',inug,isill,irange
374 call plot(avgd,sv,f,h,nbin0,33)
375 write(33,'(a)')'#-----start of S code -----'
376 write(33,'(a)')'simxyz<-read.table("outvsim.xyz")'
377 write(33,'(a)')'simsv<-read.table("outvsim.sv")'
378 write(33,'(a)')'attach(simsv)'
379 write(33,'(a)')'postscript("vsimsv.eps",onefile=T)'
380 write(33,'(2a)')'plot(avgd,sv,type="n",ylim=c(0,26),' ,
381 *   'xlab="distance",ylab="semivariance")'
382 write(33,'(a)')'points(avgd,sv,pch=1,mkh=0,cex=0.6)'
383 c   write(33,'(a)')'points(avgd,svmodel,type="l")'
384 write(33,'(a)')'dev.off()'
385 write(33,'(a)')'detach()'
386 write(33,'(a)')'## plot xy location'
387 write(33,'(a)')'##attach(simxyz)'
388 write(33,'(a)')'##postscript("vsimxyz.eps",onefile=T)'
389 write(33,'(a)')'##plot(x,y,type="n")'
390 write(33,'(a)')'##points(x,y,pch=1,mkh=0,cex=0.4)'
391 write(33,'(a)')'##dev.off()'
392 write(33,'(a)')'##detach()'
393 write(33,'(a)')'##rm(simxyz,simsv)'
394 write(33,'(a)')'q()'
395 close(33)
396 return
397 end
398 ccccccccccccccccccccccc
399 subroutine methods(nindi,ngroup,ngmean,ivar,zgvarw,
400 *   agmean,sillbar,n01,s01,r01,n02,s02,r02)
401 integer nindi,ngroup
402 real*8 ngmean, ivar,zgvarw,agmean,sillbar
403 real*8 a1,a2,a3,c0,gw,n01,s01,delta,r01
404 real*8 r02,s02,n02
405 c calculate the properties for estimating the individual
406 a1=ngmean/(ngmean-1)
407 a2=(dble(nindi)-1.0)/dble(nindi)
408 a3=(dble(ngroup)-1.0)/dble(nindi)
409 c alt 1 : when the indiv. sample is available
410 c   the initial sill = ivar (indiv. sample variance)
411 c
412 c0=ivar
413 gw=a1*(a2*c0-a3*zgvarw)
414 n01=c0
415 s01=c0
416 delta=1.0-(gw/c0)
417 r01=(-1.0*0.5107*sqrt(agmean))/(dlog(delta))
418 c alt 2 : when the indiv. sample is not available
419 c   the initial sill = sillbar
420 c   see the paper for the procedure !!!!
421 c   ref: pawitan and steel. 1999. Aggregation Bias
422 c       in semivariogram analysis. under preparation.
423 c
424 c0=sillbar
425 gw=a1*(a2*c0-a3*zgvarw)
426 delta=1.0-(gw/c0)
427 r02=(-1.0*0.51082*sqrt(agmean))/(dlog(delta))
428 s02=c0
429 n02=c0
430 c
431 c write the gsx.inf file as input for gsx.sas
432 open(15,file='gsx1a.par')
433 write(15,'(a,f10.4,a)')"%let n0=",n01,";"
434 write(15,'(a,f10.4,a)')"%let s0=",s01,";"
435 write(15,'(a,f10.4,a)')"%let r0=",r01,";"
436 close(15)
437 open(15,file='gsx1b.par')
438 write(15,'(a)')"%let n0= 0.0;"
439 write(15,'(a,f10.4,a)')"%let s0=",s01,";"
440 write(15,'(a,f10.4,a)')"%let r0=",r01,";"
441 close(15)
442 c
443 open(15,file='gsx2a.par')

```

```

444 write(15,'(a,f10.4,a)')"%let n0=",n02,";"
445 write(15,'(a,f10.4,a)')"%let s0=",s02,";"
446 write(15,'(a,f10.4,a)')"%let r0=",r02,";"
447 close(15)
448 open(15,file='gsx2b.par')
449 write(15,'(a)')"%let n0= 0.0 ;"
450 write(15,'(a,f10.4,a)')"%let s0=",s02,";"
451 write(15,'(a,f10.4,a)')"%let r0=",r02,";"
452 close(15)
453 C
454 open(15,file='initab.nsr')
455 write(15,'(6f9.3)')n01,s01,r01,n02,s02,r02
456 close(15)
457 C
458 return
459 end
460 cccccccccccccccccccccccccccccccccc
461 subroutine descg(n,xmini,xmaxi,ymini,ymaxi,nug,sill,range,
462 * ngroup,agmean,dminig,dmaxig,drange,xgmini,xgmaxi,xgmean,xgvar,
463 * ygmini,ygmaxi,ygmean,ygvar,zgmini,zgmaxi,zgmean,zgvar,
464 * ngmini,ngmaxi,ngmean,ngvar,svu,zgvarw,ngmini,ngmaxi,ngmean,ngvar,
465 * nbinu0,svu,avgdu,fuint,hu,nbinw0,svw,avgdw,fw,hw,
466 * inugw,isillw,irangew,inugu,isillu,irangeu)
467
468 integer n,nbinu0,nbinw0,ngroup,fuint(nbinu0),xdim,ydim,
469 * fwint(nbinw0)
470 real*8 xmini,xmaxi,ymini,ymaxi,nug,sill,range,
471 * agmean,dminig,dmaxig,drange,xgmini,xgmaxi,xgmean,xgvar,
472 * ygmini,ygmaxi,ygmean,ygvar,zgmini,zgmaxi,zgmean,zgvar,zgvarw,
473 * ngmini,ngmaxi,ngmean,ngvar,svu(nbinu0),avgdu(nbinu0),
474 * hu(nbinu0),svw(nbinw0),avgdw(nbinw0),fw(nbinw0),hw(nbinw0),
475 * inugw,isillw,irangew,inugu,isillu,irangeu
476 C
477 open(33,file='outvbias.s')
478 call header(33)
479 write(33,'(a)')'# Semivariogram model : n + s Exp(r)'
480 write(33,'(a,3f8.3)')'# n,s,r : ',nug,sill,range
481 write(33,'(a,i6)')'# Number of points : ', n
482 write(33,'(a,2f8.3)')'# X-min,X-max : ',xmini,xmaxi
483 write(33,'(a,2f8.3)')'# Y-min,Y-max : ',ymini,ymaxi
484 write(33,'(a)')'#-----'
485 write(33,'(a)')'#Statistics of the group data : '
486 write(33,'(a,i5,a,2i5)')'#Number of group :',ngroup,'--',
487 * xdim,ydim
488 write(33,'(a,f8.3)')'#Average area of the group :',agmean
489 write(33,'(a,f8.3)')'#Min. dist between group :',dminig
490 write(33,'(a,f8.3)')'#Max. dist between group :',dmaxig
491 write(33,'(a,f8.3)')'#Range of dist between group :',drange
492 write(33,'(a)')'#=====
493 write(33,'(a)')'# Min Max Mean Var'
494 write(33,'(a,4f8.3)')'# Xg :',xgmini,xgmaxi,xgmean,xgvar
495 write(33,'(a,4f8.3)')'# Yg :',ygmini,ygmaxi,ygmean,ygvar
496 write(33,'(a,5f8.3,a)')'# Zg :',zgmini,zgmaxi,zgmean,zgvar,
497 * zgvarw,' (weighted var)'
498 write(33,'(a,4f8.3)')'# Ng :',ngmini,ngmaxi,ngmean,ngvar
499 write(33,*)
500 write(33,'(a)')'*** The weighted version of group SV ***'
501 write(33,*)
502 do i=1,nbinw0
503 fwint(i)=int(fw(i))
504 enddo
505 call plot(avgdw,svw,fwint,hw,nbinw0,33)
506 write(33,*)
507 write(33,'(a,3f9.3)')'# Rough parameter est. (n,s,r) :',
508 * inugw,isillw,irangew
509 C The unweighted group level semivariogram
510 write(33,*)
511 write(33,'(a)')'*** The unweighted version of group SV ***'
512 write(33,*)
513 call plot(avgdu,svu,fuint,hu,nbinu0,33)
514 write(33,*)
515 write(33,'(a,3f9.3)')'# Rough parameter est. (n,s,r) :',
516 * inugu,isillu,irangeu
517 close(33)
518 return
519 end

```

Note : this Fortran program needs to call library file "vlib2k.for" (F). The input for this program is stored in the file "vsim.run", that is

```

1 Parameter file for semivariogram simulation
2 in a rectangle region
3 ###
4 BEGIN
5 1500                n
6 20.0  90.0          xmin xmax
7 10.0  80.0          ymin ymax
8 5.0  20.0  13.0     nug sill range
9 0.5                lag distance
10 15 10              ydim ydim

```

E.2 SAS procedures

Semivariogram

```

1 *-----*
2 * Procedure name : gsx.sas (semivariogram) *
3 *   written by : gandhi pawitan *
4 *   for : non-linear estimation of individual level semivariogram *
5 *   parameters (n,s,r) from the group level semivariogram. *
6 *   reference : Pawitan, G. and Steel, D. G. (1998) *
7 *   Aggregation Bias in Variogram. *
8 *   date : december 1998 *
9 *-----*
10 *input initial value ;;
11 %let n0= 0.0;
12 %let s0= 0.015375;
13 %let r0= 20.1559;
14 *define input file;
15 %let infile='employb.svu';
16 options ls=90 ps=60 nodate;
17 data gammagh;
18 infile &infile;
19 input ng nh areag areah dgh Ggh sgh areaf;
20 run;
21
22 proc nlin data=gammagh method=marquardt maxiter=200 smethod=golden;
23 parameters n=&n0 s=&s0 r=&r0;
24 f=n*(.5*( (1/ng)+(1/nh) ))+(s-n)*( .5*( (1/ng)+(1/nh) )
25      -exp(-3*dgh/r)
26      +0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
27      +0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) );
28 model Ggh=f;
29 *****;
30 bounds n>=0;
31 bounds r>0;
32 bounds s>0;
33 *****;
34 der.n=0;
35 der.s=0;
36 der.r=0;
37 *****;
38 der.n=exp(-3*dgh/r)
39      -0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
40      -0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) ;
41 der.s=( .5*( (1/ng)+(1/nh) )-exp(-3*dgh/r)
42      +0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
43      +0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) );
44 der.r=-((s-n)*3*dgh*exp(-3*dgh/r)/(r**2)
45      +(s-n)*0.5*(1-(1/ng))*3*0.5107778455*sqrt(areag)
46      *exp(-3*0.5107778455*sqrt(areag)/r)/(r**2)
47      +(s-n)*0.5*(1-(1/nh))*3*0.5107778455*sqrt(areah)
48      *exp(-3*0.5107778455*sqrt(areah)/r)/(r**2));
49 output out=svgrpl p=svgfit r=svgres parms=n1 s1 r1;
50 run;
51 *****
52 ** This part will assume the sill is given and equal to sample **
53 ** variance of individual data. **
54 ** Hence the procedure will estimate for the nugget and range. **
55 *****;
56 proc nlin data=gammagh method=marquardt maxiter=200 smethod=golden;
57 parameters n=&n0 r=&r0;
58 s2=&s0;
59 f=n*(.5*( (1/ng)+(1/nh) ))+(s2-n)*( .5*( (1/ng)+(1/nh) )

```

```

60      -exp(-3*dgh/r)
61      +0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
62      +0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) );
63 model Ggh=f;
64 *****;
65 bounds  n>=0;
66 bounds  r>0;
67 *****;
68 der.n=0;
69 der.r=0;
70 *****;
71 der.n=exp(-3*dgh/r)
72      -0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
73      -0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) ;
74
75 der.r=-(s2-n)*3*dgh*exp(-3*dgh/r)/(r**2)
76      +(s2-n)*0.5*(1-(1/ng))*3*0.5107778455*sqrt(areag)
77      *exp(-3*0.5107778455*sqrt(areag)/r)/(r**2)
78      +(s2-n)*0.5*(1-(1/nh))*3*0.5107778455*sqrt(areah)
79      *exp(-3*0.5107778455*sqrt(areah)/r)/(r**2);
80 output out=svgrp2 p=svgfit r=svgres parms=n2 r2;
81 run;

```

Cross semivariogram

```

1  *-----*
2  * Procedure name : gcx.sas (cross semivariogram) *
3  *   written by : gandhi pawitan *
4  *   for : non-linear estimation of individual level semivariogram *
5  *   parameters (n,s,r) from the group level semivariogram. *
6  *   reference : Pawitan, G. and Steel, D. G. (1998) *
7  *   Aggregation Bias in Variogram. *
8  *   date : december 1998 *
9  *-----*
10 *define initial value and input file ;
11 %let n3= 10.0;
12 %let s3= 13.0;
13 %let r3= 15.5;
14 %let infile='xyac001g.cvu';
15 *-----;
16 options ls=90 ps=60 nodate;
17 data gammagh;
18 infile &infile;
19 input ng nh areag areah dgh Ggh areaf;
20 run;
21 proc nlin data=gammagh method=marquardt maxiter=200 smethod=golden;
22 parameters n=&n3 s=&s3 r=&r3;
23 f=n*(.5*( (1/ng)+(1/nh) ))+(s-n)*( .5*( (1/ng)+(1/nh) )
24      -exp(-3*dgh/r)
25      +0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
26      +0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) );
27 model Ggh=f;
28 *****;
29 bounds  n>=0;
30 bounds  r>0;
31 *****;
32 der.n=0;
33 der.s=0;
34 der.r=0;
35 *****;
36 der.n=exp(-3*dgh/r)
37      -0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
38      -0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) ;
39 der.s=( .5*( (1/ng)+(1/nh) )-exp(-3*dgh/r)
40      +0.5*(1-(1/ng))*exp(-3*0.5107778455*sqrt(areag)/r)
41      +0.5*(1-(1/nh))*exp(-3*0.5107778455*sqrt(areah)/r) );
42 der.r=-(s-n)*3*dgh*exp(-3*dgh/r)/(r**2)
43      +(s-n)*0.5*(1-(1/ng))*3*0.5107778455*sqrt(areag)
44      *exp(-3*0.5107778455*sqrt(areag)/r)/(r**2)
45      +(s-n)*0.5*(1-(1/nh))*3*0.5107778455*sqrt(areah)
46      *exp(-3*0.5107778455*sqrt(areah)/r)/(r**2);
47 output out=svgrp1 p=svgfit r=svgres parms=n1 s1 r1;
48 run;

```

Appendix F

Description of Fortran codes and SAS procedure

This chapter gives a description of the Fortran codes involved in the simulation which were discussed in the chapter 5, 6, and 7. The simulations are mainly written in Fortran, but the semivariogram model fitting procedures are done using the SAS. The source codes are available in the CDROM attached.

F.1 Simulation of semivariogram and cross-semivariogram

The objective of this program is to generate the individual level population under a particular semivariogram or cross-semivariogram model within the defined region. The applied model are exponential, spherical and Gaussian.

| File names | par file | Description |
|--------------|--------------|-----------------------|
| vsim2k.forv | sim2k.par | univariate simulation |
| vsimbi2k.for | vsimbi2k.par | bivariate simulation |

F.2 The grouping process

The objective of these programs is the creation of the group level data based on the geographical partition of the region.

| File names | par file | Description |
|----------------|----------------|---------------------------------|
| vgroup2k.for | vgroup2k.par | grouping procedure (univariate) |
| vbigroup2k.for | vbigroup2k.par | grouping procedure (bivariate) |

F.3 Empirical semivariogram computation

The objective of this program is to compute the empirical semivariogram, either the individual level semivariogram or the group level semivariogram. In the group semivariogram will also calculated the weighted version of the group level semivariogram.

| File names | par file | Description |
|--------------|--------------|---|
| vsemiv2k.for | vsemiv2k.par | empirical semivariogram computation |
| vcross2k.for | vcross2k.par | empirical cross-semivariogram computation |
| vempiri2.for | vempiri2.par | empirical relationship of the group level semivariogram and components of the individuals semivariogram |

F.4 Exploring the MAUP

The objective of this program is to illustrate the MAUP in relation with the semivariogram.

| File names | par file | Description |
|-------------|-------------|---------------------|
| vmaup2k.for | vmaup2k.par | illustrate the MAUP |

F.5 Model fitting procedures

This procedures are written in SAS and intended to fit the model of semivariogram and cross-semivariogram.

```

1  %let ic0      = 15;      /* initial value of nugget */
2  %let ics      = 30;      /* initial value of sill */
3  %let ias      = 20;      /* initial value of range */
4  options ls=75 ps =65;
5  ****
6  ** read data file and define dataset to be used;
7  ** Note : INPUT statement may change depend on the input file;
8  ****
9  ** variogram model fitting : EXPONENTIAL MODEL;
10 data fits; infile '0svarg.dat';
11 input bin npair avgdist svg svw1 svw2 svw3 svw4 svw5;
12 *index=0 : without nugget, index=1 : with nugget;
13 index=1; run;
14 ** group level ;
15 proc nlin data=fits method=marquardt maxiter=200 smethod=golden;
16 parameters c0=&ic0 cs=&ics as=&ias;
17 if index=1 then
18   f=c0 + (cs-c0)*(1-exp(-3*avgdist/as));
19 if index=0 then
20   f=cs*(1-exp(-3*avgdist/as));
21 _weight_=(npair)/(f**2);
22 model svg=f;
23   bounds cs>0, as>0;
24   der.cs=0;der.as=0;
25 if index=1 then do;
26   bounds c0>0;
27   der.c0=0;
28   der.c0=1-(1-exp(-3*avgdist/as));
29 end;
30 if index=1 then do;
31   der.cs=1-exp(-3*avgdist/as);
32   der.as=- (cs-c0)*3*avgdist*exp(-3*avgdist/as)/(as**2);
33 end;

```



```

34  if index=0 then do;
35      der.cs=1-exp(-3*avgdist/as);
36      der.as=-cs*3*avgdist*exp(-1*avgdist/as)/(as**2);
37  end;
38  output out=svf p=svfit r=svres parms= c0 cs as; run;
39  data a; set svf; filename out '0svarg0.out'; file out;
40  put (c0 cs as) (14.5); run;

```

Note : This procedure can be used as well for the weighted semivariogram and also the unweighted or weighted cross semivariogram. This can be done by setting the appropriate initial value and change **svg** in the line **model svg=f**; with the proper variable names, for example

```

model svw1=f; /* this is for weighted #1 */
model svw2=f; /* this is for weighted #2 */

```

F.6 Miscellaneous subroutine

The subroutines are needed in compilation of the main program.

| File names | Description |
|-------------|---|
| vlib2k.for | collection of the Fortran subroutine, which is needed to compile the above programs |
| vdist2k.for | program to illustrate the empirical properties of the distance distribution within and between groups |

F.7 Some notes of Fortran

The Fortran codes was compiled in the UNIX and DOS system. In the DOS system, the *g77* compiler is used. The compact *G77* for Win32 (Windows 95/NT) package is available from

<http://www.geocities.com/Athens/Olympus/5564>

Appendix G

Data sets

There are three main data sets are used in this thesis. The data sets contain some social characteristics. The data are based on the 1991 Australian census of Population and Housing. The first data sets is the records from the Illawarra region, and the second data sets is from the Adelaide region. They will be called subsequently the Illawarra data set and Adelaide data set. The third data set was come from the simulation result, containing 10,000 points.

G.1 Illawarra data set

| Variables name | Description |
|----------------|--|
| cdid | collection district identification |
| dpc | derived post code |
| area | collection district area in km ² |
| lon | longitude of the CD centroid |
| lat | latitude of the CD centroid |
| easting | the distance (km) to the most west point |
| northing | the distance (km) to the most south point |
| base1 | number of people of 15 years old and more |
| base2 | number of people within 15-65 years old |
| employ | number of people employed |
| unemploy | number of people unemployed |
| formal | number of people with formal qualification |
| informal | number of people with informal qualification |

G.2 Adelaide data set

| Variables name | Description |
|----------------|--------------------------------------|
| cdid | collection district identification |
| lga | local government area identification |

| | |
|----------|---|
| dpc | derived post code |
| ssc | collection district derived suburb id. |
| km2 | area of the collection district |
| easting | the distance (km) to the most west point |
| northing | the distance (km) to the most south point |
| ng | number of peoples between 15-65 years old |
| employ | employment rate |
| unemp | unemployment rate |
| labor | labor participation rate |
| sal020 | rate of income below AUD 20000 |
| sal2040 | rate of income between AUD 20000-40000 |
| sal40ov | rate of income over AUD 40000 |
| owned | rate of housing by owned |
| bpch | rate of housing being purchased |
| rent | rate of housing rent |
| wage | rate of wage or salary earner |
| self | rate of self employed |
| empyer | rate of employer |
| formal | formal qualification rate |
| inform | informal qualification rate |
| notql | rate of not qualified person |

G.3 Simulated data set

The data contains only x, y, z variables, where x is the easting, y is the northing, and z is observations. The random variable z is originally distributed by $N(0, 1)$, but by the exponential semivariogram model with ($n = 5$, $s = 20$, and $r = 15$), it becomes a random variable with mean=0.758 and variance=18.397. The properties of this population can be looked as follow

[illegible]

References

- ABS, & MapInfo. (1993). *CDA91 with MapInfo User's Manual*. Australia: MapInfo Australia.
- Amrhein, C. G. (1995). Searching for the elusive aggregation effect: evidence from statistical simulation. *Environment and Planning A*, 27, 105–119.
- Amrhein, C. G., & Reynolds, H. (1996). Using Spatial Statistics to Assess Aggregation Effects. *Geographical Systems*, 3, 143–158.
- Anselin, L. (1988). *Spatial Econometrics, Methods and Models*. Dordrecht: Kluwer.
- Anselin, L. (1992). Space and Applied Econometrics. *Regional Science and Urban Economics*, 22, 307–316.
- Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis*, 27, 93–115.
- Anselin, L., & Getis, A. (1992). Spatial Statistical Analysis and Geographic Information Systems. *The Annals of Regional Science*, 26, 19–33.
- Arbia, G. (1989a). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.
- Arbia, G. (1989b). Statistical effect of spatial data transformation : a proposed general framework. In M. Goodchild & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 249–259). London: Taylor & Francis.
- Arbia, G. (1993). The Use of GIS in Spatial Statistical Surveys. *International Statistical Review*, 61(2), 339–359.
- Bailey, T. C. (1994). A review of statistical spatial analysis in geographical information systems. In S. Fotheringham & P. Rogerson (Eds.), *Spatial analysis and gis* (pp. 13–43). UK.
- Bartlett, M. S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika*, 51(3 and 4), 299–311.
- Birkin, M., Clarke, G., Clarke, M., & Wilson, A. (1990). Elements of a model-based Geographic Information System for the Evaluation of Urban Policy. In L. Worral (Ed.), *Geographic information systems : Developments and applications* (pp. 133–162). London: Belhaven Press.
- Bond, D., & Devine, P. (1991). The Role of Geographic Information Systems in Survey Analysis. *The Statistician*, 40, 209–215.
- Bonham-Carter, G. F. (1994). *Geographic Information Systems for Geoscientists : Modelling with GIS*. New York: Elsevier Science, Inc.

- Brown, P. J. B. (1991). Exploring Geodemographics. In I. Masser & Blakemore (Eds.), *Handling geographical information : Methodology and potential application* (pp. 221–258). London: Taylor & Francis.
- Burrough, P. A. (1986). *Principles of Geographical Information Systems for Land Resources Assessment*. London: Oxford University Press.
- Burrough, P. A., & McDonnel, R. A. (1998). *Principles of Geographical Information Systems*. London: Oxford University Press.
- Carr, J. R. (1995). *Numerical Analysis for Geological Sciences*. Englewood Cliffs: Prentice Hall.
- Carrat, F., & Valleron, A. (1992). Epidemiologic mapping using the kriging method : application to an influenza-like illness epidemic in France. *American Journal of Epidemiology*, 135(11), 1293–1300.
- Castles, I. (1991). *How Australia Takes A Census*. Canberra: Australian Bureau of Statistics.
- Christensen, R. (1991). *Linear Models for Multivariate, Time Series, and Spatial Data*. New York: Springer-Verlag.
- Christensen, R., Johnson, W., & Pearson, L. M. (1993). Covariance function diagnostics for spatial linear models. *Mathematical Geology*, 25(2), 145–160.
- Clark, I. (1982). *Practical geostatistics*. London: Applied Science Publishers Ltd.
- Cliff, A. D., Hagget, P., Ord, J. K., Basset, K. A., & Davies, R. B. (1975). *Elements of Spatial Structure A Quantitative Approach*. London: Cambridge University Press.
- Cliff, A. D., & Ord, J. K. (1981). *Spatial Processes Models and Applications*. London: Pion.
- Clifford, P., Richardson, S., & Hémon, D. (1989). Assessing the Significance of the Correlation Between Two Spatial Processes. *Biometrics*, 45, 123–134.
- Cressie, N. (1985). Fitting Variogram Models by Weighted Least Squares. *Mathematical Geology*, 17(5), 563–586.
- Cressie, N. (1989). Geostatistics. *The American Statistician*, 43(4), 197–202.
- Cressie, N. (1991). *Spatial Statistics*. New York: Jon Wiley & Sons.
- Cressie, N. (1996). Change of Support and The Modifiable Areal Unit Problem. *Geographical Systems*, 3, 159–180.
- Cressie, N., & Aldworth, J. (1997). Spatial Statistical Analysis and Its Consequences for Spatial Sampling. In E. Y. Baafi & N. A. Schofield (Eds.), *Geostatistics wollongong '96 vol. 1* (pp. 126–137). Netherlands: Kluwer Academic Publisher.
- Curtis, A. J., & MacPherson, A. D. (1996). The Zone Definition Problem in Survey Research : An Empirical Example from New York State. *Professional Geographer*, 48(3), 310–320.
- Deutsch, C. V., & Journel, A. G. (1992). *GSLIB : Geostatistical software library and user's guide*. Oxford: Oxford University Press.
- Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics*, 47(3), 299–350.
- Ding, Y., & Fotheringham, A. S. (1992). The Integration of Spatial Analysis and GIS. *Computers, Environment, and Urban Systems*, 16, 3–19.
- Flowerdew, R., & Goldstein, W. (1989). Geodemographics in Practice : Development in North America. *Environment and Planning A*, 21, 605–616.
-

- Flowerdew, R., & Green, M. (1989). Statistical Methods for Inference between incompatible Zonal Systems. In M. Goodchild & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 239–247). London: Taylor & Francis.
- Flowerdew, R., & Green, M. (1991). Data Integration : Statistical Methods for Transferring Data Between Zonal Systems. In I. Masser & Blakemore (Eds.), *Handling geographical information : Methodology and potential application* (pp. 38–54). London: Taylor & Francis.
- Flowerdew, R., & Green, M. (1992). Developments in Areal Interpolation Methods and GIS. *The Annals of Regional Science*, 26, 67–78.
- Flowerdew, R., & Green, M. (1994). Areal Interpolation and Types of Data. In S. Fotheringham & P. Rogerson (Eds.), *Spatial analysis and gis* (pp. 121–145). London: Taylor & Francis.
- Fotheringham, A. S., Charlton, M., & Brunsdon, C. (1996). The Geography of Parameter Space : an Investigation of Spatial Non-Stationarity. *International Journal of Geographical Information Systems*, 10(5), 605–627.
- Fotheringham, A. S., & Rogerson, P. A. (1993). GIS and Spatial Analytical Problems. *International Journal of Geographical Information Systems*, 7(1), 3–19.
- Gehlke, C. E., & Biehl, K. (1934). Certain effects of Grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 24(supplement), 169–170.
- Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3), 189–206.
- Getis, A., & Ord, J. K. (1996). Local spatial statistics : an overview. In P. Longley & M. Batty (Eds.), *Spatial analysis : Modelling in a gis environment* (pp. 261–277). Cambridge: GeoInformation International.
- Ghosh, B. (1951). Random distances within a rectangle and between two rectangles. *Bull. Calcutta Math. Soc.*, 43, 17–24.
- Goodchild, M. F. (1980). Algorithm 9: Simulation of autocorrelation for aggregated data. *Environment and Planning A*, 12, 1073–81.
- Green, M., & Flowerdew, R. (1996). New evidence on the modifiable areal unit problem. In P. Longley & M. Batty (Eds.), *Spatial analysis : Modelling in a gis environment* (pp. 41–54). Cambridge: GeoInformation International.
- Griffith, D. A. (1988). *Advanced Spatial Statistics*. Dordrecht: Kluwer Academic Publishers.
- Griffith, D. A. (1993). Which Spatial Statistics Techniques Should be Converted to GIS Functions? In M. M. Fischer & P. Nijkamp (Eds.), *Geographic information systems, spatial modelling and policy evaluation*. New York: Springer-Verlag.
- Griffith, D. A. (1996). Introduction: the need for spatial statistics. In S. L. Arlinghaus, D. A. Griffith, W. C. Arlinghaus, W. D. Drake, & J. D. Nystuen (Eds.), *Practical handbook of spatial statistics* (pp. 1–15). Boca Raton: CRC Press.
- Griffith, D. A., & Amrhein, C. G. (1991). *Statistical Analysis for Geographers*. Englewood Cliffs: Prentice-Hall.
- Griffith, D. A., Haining, R., & Arbia, G. (1994). Heterogeneity of Attribute Sampling Error in Spatial Data Sets. *Geographical Analysis*, 26(4), 300–320.
- Grondona, M. O., & Cressie, N. (1991). Using Spatial Considerations in the Analysis of Experiments. *Technometrics*, 33(4), 381–392.
- Hagget, P., Cliff, A. D., & Frey, A. (1977). *Locational Methods*. London: Edward Arnold.

- Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press.
- Haining, R. (1994). Designing spatial data analysis modules for geographical information systems. In S. Fotheringham & P. Rogerson (Eds.), *Spatial analysis and gis* (pp. 45–63). UK: Taylor & Francis.
- Haining, R., Griffith, D. A., & Bennet, R. (1983). Simulating two-dimensional autocorrelated surfaces. *Geographical analysis*, 15(3), 247–255.
- Haining, R., Wise, S., & Ma, J. (1998). Exploratory spatial data analysis in a geographic information system environment. *The Statistician*, 47(3), 457–469.
- Harvey, D. (1969). *Explanation in Geography*. London: Edward Arnold.
- Haslett, J. (1992). Spatial Data Analysis – Challenges. *The Statistician*, 41, 271–284.
- Haslett, J. (1997). On the sample variogram and the sample autocovariance for non-stationary time series. *The Statistician*, 46(4), 475–485.
- Hepple, L. W. (1976). A Maximum Likelihood Model for Econometric Estimation with Spatial Series. In I. Masser (Ed.), *Theory and practice in regional science* (pp. 90–104). London: Pion.
- Hepple, L. W. (1996). Directions and Opportunities in Spatial Econometrics. In P. Longley & M. Batty (Eds.), *Spatial analysis : Modelling in a gis environment* (pp. 231–246). Glasgow: Geoinformation International.
- Holt, D., Steel, D. G., & Tranmer, M. (1996). Area Homogeneity and the Modifiable Areal Unit Problem. *Geographical System*, 2, 83–101.
- Kaluzny, S. P., Vega, S. C., Cardoso, T. P., & Shelly, A. A. (1998). *S+ Spatial Stats : User's manual for Windows and Unix*. New York: Springer-Verlag.
- Kelejian, H. H., & Robinson, D. P. (1992). Spatial Autocorrelation : A new computationally simple test with an application to per capita county police expenditures. *Regional Science and Urban Economics*, 22, 317–331.
- King, G. (1997). *A solution to the ecological inference problem : reconstructing individual behavior from aggregate data*. Princeton, NJ: Princeton University Press.
- Langbein, L. I., & Lichtman, A. J. (1978). *Ecological Inference*. Beverly Hills, Cal: Sage.
- Littel, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *Sas System for Mixed Models*. Cary, NC: Sas Institute Inc.
- Maling, D. H. (1992). *Coordinate System and Map Projections*. England: Pergamon Press.
- Martin, D. (1996). *Geographic Information System Socioeconomic Applications*. London: Routledge.
- Matérn, B. (1986). *Spatial Variation*. Berlin Heidelberg: Springer-Verlag.
- McCracken, K. W. J. (1983). Dimensions of Social Well-being : Implications of Alternative Spatial Frames. *Environment and Planning A*, 15, 579–592.
- McLennan, W. (1995). *Australian Standard Geographical Classification*. Canberra: Australian Bureau of Statistics.
- Müller, H. G., Stadtmüller, U., & Tabnak, F. (1997). Spatial Smoothing of Geographically Aggregated Data, with Application to the Construction of Incidence Maps. *Journal of the American Statistical Association*, 92(437), 61–71.
- Morisette, J. (1997). Short note : Examples using SAS to fit the model of linear coregionalization. *Computers & Geosciences*, 23(3), 317–323.
-

- Myers, D. E. (1982). Matrix formulation of cokriging. *Journal of Mathematical Geology*, 14(3), 249–257.
- Myers, D. E. (1988). Some Aspects of Multivariate Analysis. In C. F. Chung, A. G. Fabbri, & R. Sinding-Larsen (Eds.), *Quantitative Analysis of Mineral and Energy Resources* (pp. 669–687). Dordrecht: D. Reidel Publishing Company.
- O'Brien, L. G. (1990). Small Area Information Systems : Problems and Prospect. In L. Worral (Ed.), *Geographic Information Systems : Developments and Applications* (pp. 215–244). London: Belhaven Press.
- Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers, New Series*, 2, 459–472.
- Openshaw, S. (1978). An optimal zoning approach to the study of spatially aggregated data. In I. Masser & P. J. B. Brown (Eds.), *Spatial representation and spatial interaction* (pp. 95–113). Leiden: Martinus Nijhoff Social Sciences Division.
- Openshaw, S. (1984). Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A*, 6, 17–31.
- Openshaw, S., & Taylor, P. J. (1979). A Million or so Correlation Coefficients : Three Experiments on the Modifiable Areal Unit Problem. In N. Wrigley (Ed.), *Statistical Applications in the Spatial Sciences* (pp. 127–144). London: Pion Ltd.
- Openshaw, S., & Taylor, P. J. (1981). Modifiable areal unit problem. In N. Wrigley & R. J. Bennet (Eds.), *Quantitative geography : a british view* (pp. 60–69). London: Routledge & Kegan Paul.
- Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics : Distributional Issues and an Application. *Geographical Analysis*, 27(4), 286–306.
- Pawitan, G. (1993). *Area frame construction utilizing Geographic Information System*. Unpublished manuscript master thesis, University of Philippines at Los Banõs, Los Banõs–Philippines.
- Pettit, A. N., & McBratney, A. B. (1993). Sampling design for estimating spatial variance components. *Applied Statistics*, 42(1), 185–209.
- Rajagopalan, S. (1992). "Lambert-Grid"—A Program for Converting Geographic Coordinates to Grid Coordinates and Vice-versa. *Computers & Geosciences*, 18(2/3), 349–366.
- Ripley, B. D. (1979). Test of randomness for spatial point patterns. *Journal of the Royal Statistical Society, Series B*(41), 368–374.
- Ripley, B. D. (1981). *Spatial Statistics*. New York: John Wiley & Sons.
- Robinson, G. K. (1990). A role for variograms. *Australian Journal of Statistics*, 32(3), 327–335.
- Robinson, W. S. (1950). Ecological Correlations and the Behaviour of Individuals. *American Sociological Review*, 15, 351–357.
- SAS. (1996). *SAS/STAT Technical Report : Spatial prediction using the SAS system*. Cary, North Carolina: SAS Institute Inc.
- Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.
- Steel, D. G., & Holt, D. (1996a). Analysing and Adjusting Aggregation Effects : The Ecological Fallacy Revisited. *International Statistical Review*, 64(1), 39–60.
- Steel, D. G., & Holt, D. (1996b). Rules for Random Aggregation. *Environment and Planning A*, 28, 957–978.
-

- Steel, D. G., Holt, D., & Tranmer, M. (1996). Making Unit-Level Inferences From Aggregated Data. *Survey Methodology*, 22(1), 2–15.
- Tobler, W. R. (1989). Frame independent spatial analysis. In M. Goodchild & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 115–122). London: Taylor & Francis, Ltd.
- Tranmer, M., & Steel, D. G. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, 30, 817–831.
- Unwin, A. (1998). Exploratory spatial data analysis with local statistics. *The Statistician*, 47(3), 415–421.
- Vaughan, R. (1984). Approximate formulas for Average distances associated with zones. *Transportation Science*, 18(3), 231–244.
- Venables, W. N., & Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.
- Ver Hoef, J. M., & Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology*, 25(2), 219–240.
- Wackernagel, H. (1988). Geostatistical techniques for interpreting multivariate spatial information. In C. F. Chung, A. G. Fabbri, & R. Sinding-Larsen (Eds.), *Quantitative Analysis of Mineral and Energy Resources* (pp. 393–409). Dordrecht: D. Reidel Publishing Company.
- Wackernagel, H. (1998). *Multivariate Geostatistics : An Introduction with Application*. Berlin: Springer.
- Watson, S. (1997). Evaluation of semivariance estimator under normal conditions. *The Statistician*, 46(4), 495–503.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. New York: John & Wiley and Sons.
- Wilhelm, A., & Steck, R. (1998). Exploring spatial data by using interactive graphics and local statistics. *The Statistician*, 47(3), 423–430.
- Wilson, R. M. (1990). The average distance between two zones. *Geographical Analysis*, 22(4), 348–351.
- Wong, D. (1996). Aggregation Effects in Geo-Referenced Data. In S. L. Arlinghaus, D. A. Griffith, W. C. Arlinghaus, W. D. Drake, & J. D. Nystuen (Eds.), *Practical Handbook of Spatial Statistics*. Boca Raton: CRC Press.
- Wong, D. W. S., & Fotheringham, A. S. (1990). Urban Systems as Examples of Bounded Chaos : Exploring the Relationship Between Fractal Dimension, Rank-Size, and Rural-to-Urban Migration. *Geografiska Annaler*, 72-B(2-3), 89–99.
- Wrigley, N., Holt, T., Steel, D., & Tranmer, M. (1996). Analysing, modelling, and resolving the ecological fallacy. In P. Longley & M. Batty (Eds.), *Spatial analysis : Modelling in a gis environment* (pp. 25–40). Cambridge: GeoInformation International.
- Zimmerman, D. L., & Zimmerman, M. B. (1991). A Comparison of Spatial Semivariogram Estimators and Corresponding Ordinary Kriging Predictors. *Technometrics*, 33(1), 77–91.