

# University of Wollongong - Research Online

## Thesis Collection

Title: Scale effects in multilevel modeling

Author: Russell R Familiar

Year: 2008

Repository DOI:

### 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: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. 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.

**Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.**

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

*University of Wollongong Thesis Collections*

*University of Wollongong Thesis Collection*

---

*University of Wollongong*

*Year 2008*

---

Scale effects in multilevel modeling

Russell R. Familiar  
University of Wollongong

Familiar, Russell R, Scale effects in multilevel modeling, PhD thesis, School of Mathematics and Applied Statistics, University of Wollongong, 2008. <http://ro.uow.edu.au/theses/566>

This paper is posted at Research Online.  
<http://ro.uow.edu.au/theses/566>

## **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.

# Scale Effects in Multilevel Modeling

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

**Doctor of Philosophy**

*from*

**University of Wollongong**

*by*

**Russell R. Familiar**    B.Sc.Mathematics, MS Applied Statistics.  
MSU-IIT

**School of Mathematics and Applied Statistics**

**2008**

This thesis is submitted to the University of Wollongong in fulfillment of the requirements for the award of Doctor of Philosophy, in the School of Mathematics and Applied Statistics. I hereby declare that the work described here is my own unless otherwise referenced and has not been submitted for a degree to any other University or Institution.

Russell R. Familiar

May 2008

# Acknowledgements

Many people are involved in one way or another in finishing this research. Particularly, Professor David Steel who had given me the opportunity to do this project and believed that I can make it. Thank you very much sir.

I am grateful to the staff of the School of Mathematics and Applied Statistics for their help during the research period. Special thanks to Nick von Sanden who had shared his knowledge in computer programming.

To my family Mama Raquel, Tutut, Nico, Sweetrat, Panot and the whole Familiar and Tabuzo clans who had given me support in one way or another in the making of this thesis.

Special thanks to Jojie my ever patient wife.

# Abstract

In many instances data are available as aggregated measurements for a set of areal units that are arbitrarily defined in terms of number and boundaries. Analysis using spatial data is a multi-disciplinary subject attracting the attention of statisticians, geographers, physical and social scientists. The Modifiable Areal Unit Problem (MAUP) is the sensitivity of results of statistical analysis to the definition of areal units for which the data are available. The results vary with the level of aggregation and the configuration of the zoning system. Multilevel models offer an approach to the MAUP. Multilevel modeling is potentially subject to the MAUP, since different estimates of the variance components can be obtained if boundaries are changed or a different scale is used.

This thesis presents results of experiments conducted to look into the scale effects of statistics calculated directly from aggregated data and statistics derived from a simple multilevel under different initial conditions. The analysis of spatial data is usually affected by the complex relationships between variables and the existence of spatial autocorrelation. A reason for multilevel models being subject to the MAUP is that, while the data available may be hierarchical, the population structure may be more complex. Theoretical and empirical investigations to link a simple multilevel model and spatial autocorrelation and the implications for the MAUP are conducted.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Modifiable Areal Unit Problem . . . . .	1
1.2	Spatial Autocorrelation . . . . .	4
1.3	Multilevel Modeling . . . . .	5
1.4	The Problem . . . . .	6
<b>2</b>	<b>Review of Research into the MAUP</b>	<b>8</b>
2.1	Introduction . . . . .	8
2.2	Research on the MAUP . . . . .	9
2.3	Some Methods Employed to Solve the MAUP . . . . .	17
2.3.1	Data Manipulation Approach . . . . .	17
2.3.2	Technique-Oriented Approaches . . . . .	18
2.3.3	Error Modeling Approaches . . . . .	19
2.3.4	Some comments and recommendations on how to find a solu- tion of the MAUP . . . . .	20
<b>3</b>	<b>The Causes of the MAUP</b>	<b>21</b>
3.1	Basic Theory . . . . .	21
3.1.1	Spatial Aggregation . . . . .	21
3.1.2	Intra-Area Correlation and Cross Correlation . . . . .	24
3.1.3	Pure Correlation . . . . .	28
3.1.4	Pure Regression . . . . .	29
3.1.5	Moran's I . . . . .	29



3.1.6	Cross-Moran's I . . . . .	30
3.1.7	Relationships Between Pure Coefficients and the Intra-area Correlation . . . . .	31
3.1.8	The relationship between intra-area correlation and the Moran's I . . . . .	33
<b>4</b>	<b>Multilevel Modeling and the MAUP</b>	<b>36</b>
4.1	Is Multilevel modeling a possible solution to the MAUP? . . . . .	36
4.2	Experiment 1: Scale effects of some statistics from simulated data . .	39
4.2.1	Data Set 1: Both variables have low autocorrelation . . . . .	40
4.2.2	Data Set 2: Both variables have medium autocorrelation . . . .	59
4.2.3	Data Set 3: Both variables have high autocorrelation . . . . .	76
4.2.4	Discussion of Experiment 1 . . . . .	90
4.3	Experiment 2: Scale effects when both variables are not autocorrelated	96
4.4	Comments on Experiments 1 and 2 . . . . .	108
4.5	Experiment 3: Scale effects when the variables do not have the same levels of autocorrelation . . . . .	122
4.6	Summary . . . . .	129
<b>5</b>	<b>Analysis of Real Data from UK Census</b>	<b>132</b>
5.1	Data from two sources . . . . .	132
5.1.1	Case 1: Individual level from SAR and second level is Enu- meration District (ED) . . . . .	137
5.1.2	Case 2: Individual level from SAR and second level is Ward .	148
5.1.3	Linear Regressions and Pure Regressions of Percentage of Full- Time Workers and Other variables . . . . .	156
5.2	Data from one source (SAS) . . . . .	160
5.2.1	Some Statistics from 1991 UK Census . . . . .	160
5.2.2	Regression Analysis of Variable Ftw and Other variables . . .	164
5.2.3	Spatial autocorrelation of the Variables . . . . .	164
5.3	Summary . . . . .	173

<b>6</b>	<b>Data generation based on actual boundaries</b>	<b>176</b>
6.1	Data Set Generator . . . . .	176
6.2	Case 1: Variables have the same spatial autocorrelation . . . . .	178
6.2.1	Behavior of Some Statistics . . . . .	180
6.2.2	Aggregation Effects . . . . .	187
6.2.3	Intra-Area Correlations and Intra-Area Cross-Correlations . .	191
6.3	Different Initial correlations . . . . .	194
6.3.1	Relationship between some statistics and the Moran's I with different proximity matrices: . . . . .	203
6.4	Case 2: Variables have different spatial autocorrelation . . . . .	208
6.5	Summary . . . . .	214
<b>7</b>	<b>Conclusion</b>	<b>217</b>
7.1	Summary and Conclusions . . . . .	218
7.1.1	The Mean and the variance . . . . .	220
7.1.2	The direct correlation and regression coefficients . . . . .	222
7.1.3	The pure coefficients . . . . .	223
7.1.4	Approach to Aggregation . . . . .	225
7.2	Further Research and Development . . . . .	225
<b>A</b>	<b>Dataset Simulation Codes</b>	<b>227</b>
A.1	Data Set Generator for a square grid . . . . .	227
A.2	Data sets for a region . . . . .	234
A.3	Proximity Weights (WtEDLag01Queen.txt) . . . . .	243
A.4	Weight (NumEdPerWard.txt) . . . . .	244
A.5	Proximity Weights used in the computation of Morans I (Block Prox- imity) . . . . .	244
	<b>Glossary of Terms</b>	<b>246</b>
	<b>References</b>	<b>247</b>

# List of Figures

4.1	Unweighted Variance of X and Covariance(X,Y), X and Y both have low autocorrelation . . . . .	47
4.2	Weighted Variance of X and Covariance(X,Y), X and Y both have low autocorrelation . . . . .	49
4.3	Pearson Correlation, X and Y both have low autocorrelation . . . . .	50
4.4	Regression Coefficient, X and Y both have low autocorrelation . . . . .	53
4.5	Variance Components of X, X have low autocorrelation . . . . .	54
4.6	Intra-area correlation X, X have low autocorrelation . . . . .	56
4.7	Intra-area cross-correlation of X and Y, X and Y both have low autocorrelation . . . . .	57
4.8	Pure Correlation, X and Y both have low autocorrelation . . . . .	58
4.9	Pure Regression, X and Y both have low autocorrelation . . . . .	60
4.10	Unweighted Variance of X and Covariance(X,Y), X and Y both have medium autocorrelation . . . . .	65
4.11	Weighted Variance of X and Covariance(X,Y), X and Y both have medium autocorrelation . . . . .	66
4.12	Pearson Correlation (The horizontal axis denotes number of groups), X and Y both have medium autocorrelation . . . . .	69
4.13	Regression Coefficient, X and Y both have medium autocorrelation . . . . .	70
4.14	Variance Components of X, X have medium autocorrelation . . . . .	71
4.15	Intra-Area Correlation X, X have medium autocorrelation . . . . .	72
4.16	Pure Correlation, X and Y both have medium autocorrelation . . . . .	74
4.17	Pure Regression, X and Y both have medium autocorrelation . . . . .	75

4.18 Unweighted Variance of X and Covariance (X,Y) . . . . .	81
4.19 Weighted Variance of X and Covariance(X,Y), X and Y both have high autocorrelation . . . . .	83
4.20 Pearson Correlation, X and Y both have high autocorrelation . . . .	84
4.21 Regression Coefficient, X and Y both have high autocorrelation . . .	85
4.22 Variance Components of X, X have high autocorrelation . . . . .	86
4.23 Intra-Area correlation of X, X have high autocorrelation . . . . .	87
4.24 Intra-Area Cross-Correlation, X and Y both have high autocorrelation	88
4.25 Pure Correlation, X and Y both have high autocorrelation . . . . .	89
4.26 Pure Regression, X and Y both have high autocorrelation . . . . .	91
4.27 Pearson Correlation, X and Y both are not autocorrelated . . . . .	101
4.28 Variance Components of X, X not autocorrelated . . . . .	103
4.29 Intra-Area Correlation X, X not autocorrelated . . . . .	105
4.30 Intra-Area Correlation Y, Y not autocorrelated . . . . .	106
4.31 Pure Correlation, X and Y both are not autocorrelated . . . . .	107
4.32 Level 2 and Level 1 Pure Regression, X and Y both are not autocor- related . . . . .	108
4.33 Correlations at different levels of aggregation and degrees of autocor- relation . . . . .	110
4.34 Regression Coefficient at Different Levels of Aggregation and Degrees of Autocorrelation . . . . .	112
4.35 Unweighted Covariance at Different Levels of Aggregation and Auto- correlation . . . . .	113
4.36 Level 2 Pure Correlation at Different levels of Aggregation and De- grees of Autocorrelation . . . . .	114
4.37 Level 1 Pure Correlation at Different levels of Aggregation and De- grees of Autocorrelation . . . . .	115
4.38 Level 2 Pure Regression at Different levels of Aggregation and Degrees of Autocorrelation . . . . .	116

4.39	Level 1 Pure Regression at Different levels of Aggregation and Degrees of Autocorrelation . . . . .	117
4.40	Relationship between Intra-Area Correlation and N-Bar at different degrees of autocorrelation for the two variable: (a) very low, (b) low, (c) medium, and (d) high . . . . .	119
4.41	Relationship between the mean of Intra-Area Correlation and N-Bar at different degrees of autocorrelation for the two variable . . . . .	120
4.42	Relationship between Intra-Area Cross-Correlation and N-Bar at different degrees of autocorrelation for the two variable: (a) very low, (b) low, (c) medium, and (d) high . . . . .	121
4.43	Relationship between the mean of Intra-Area Cross-Correlation and N-Bar at different degrees of autocorrelation for the two variable . . . . .	122
4.44	Pearson Correlation variable X (low autocorrelation) and variables Y (different autocorrelation) . . . . .	123
4.45	Regression Coefficient variable X (low autocorrelation) and variables Y (different autocorrelation) . . . . .	124
4.46	Level 2 Pure Correlation, variable X (low autocorrelation) and variables Y (different autocorrelation) . . . . .	125
4.47	Level 1 Pure Correlation, variable X (low autocorrelation) and variables Y (different autocorrelation) . . . . .	126
4.48	Level 2 Pure Regression, variable X (low autocorrelation) and variables Y (different autocorrelation) . . . . .	127
4.49	Level 1 Pure Regression, variable X (low autocorrelation) and variables Y (different autocorrelation) . . . . .	128
5.1	Location of the four districts . . . . .	133
5.2	The region with its boundaries . . . . .	134
5.3	Correlations: Individual level (SAR) and ED level (SAR) . . . . .	144
5.4	Individual level Correlations(SAR) versus Level 1 Pure Correlation . . . . .	146
5.5	ED Level Correlations versus Level 2 Pure Correlation . . . . .	147
5.6	Individual Level Correlations versus Ward Level Correlation . . . . .	150

5.7	Ward Level Correlations versus Level 2 Pure Correlations . . . . .	152
5.8	Individual level correlations vs Ward and ED level correlations . . . .	153
5.9	Individual level correlations vs Ward and ED Level 1 Pure correlations	154
5.10	Individual level correlations vs Ward and ED Level 2 Pure correlations	156
5.11	ED level correlations vs. Ward level correlations . . . . .	161
5.12	ED level correlations vs. Level 1 and Level Pure Correlations . . . . .	163
5.13	Bivariate Moran vs. Intra-area Cross-correlation . . . . .	167
5.14	Graphical representation of the variable AGE at different levels . . .	168
5.15	Graphical representation of the variable FTW at different levels . . .	169
5.16	Graphical representation of the variable UNEMP at different levels .	170
5.17	Graphical representation of the variables at different levels . . . . .	171
5.18	Graphical representation of the variable NOCAR at different levels .	172
6.1	The region with its boundaries . . . . .	177
6.2	Some realizations of the data generator . . . . .	179
6.3	Some realizations of the data generator the same degree of autocorrelation $\rho=0.8$ . . . . .	180
6.4	Unweighted Variance of X and Covariance of X and Y at Ward level .	181
6.5	Weighted Variance X at Ward level . . . . .	182
6.6	Weighted Covariance at Ward level . . . . .	183
6.7	Correlations at Ward level . . . . .	184
6.8	Level 1 and Level 2 Variance components of X . . . . .	185
6.9	Level 1 and Level 2 Pure correlations . . . . .	187
6.10	Level 1 and Level 2 Pure Regression Coefficients . . . . .	189
6.11	Variance and Covariance Aggregation Effects at Different degrees of Autocorrelations . . . . .	190
6.12	Variance Aggregation Effects versus Moran's I with Proximity matrix (a) Lag1 (b) "block" . . . . .	191
6.13	Combined Variance Effects versus Moran's I with Proximity matrix (a) Lag1 (b) "block" . . . . .	192
6.14	Intra-area correlations versus Moran's I with block proximity . . . . .	193

6.15	Intra-area cross-correlations versus Bivariate Moran's I with block proximity . . . . .	194
6.16	Unweighted Variance of X at Ward level with different initial correlations at different degrees of autocorrelations . . . . .	195
6.17	Unweighted Covariance of X and Y at Ward level with different initial correlations at different degrees of autocorrelations . . . . .	196
6.18	Aggregated weighted covariance with different initial correlations at different degrees of autocorrelations . . . . .	196
6.19	Weighted Correlations at Ward level with at different initial correlations and different degrees of autocorrelations . . . . .	197
6.20	Level 1 Pure Correlations at different initial correlations at different degrees of autocorrelations . . . . .	197
6.21	Level 2 Pure Correlations at different initial correlations at different degrees of autocorrelations . . . . .	198
6.22	Level 1 Pure Regression at different initial correlations at different degrees of autocorrelations . . . . .	199
6.23	Level 2 Pure Regression at different initial correlations at different degrees of autocorrelations . . . . .	200
6.24	Distributions of the three statistics at initial correlations of 0.3 at different degrees of autocorrelations . . . . .	201
6.25	Aggregation Effects at different initial correlations at different degrees of autocorrelations . . . . .	202
6.26	Aggregation Effects at different degrees of autocorrelations at different initial correlations . . . . .	202

6.27	Relationship between Unweighted Variance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right $r=0.1$ , $r=0.3$ , $r=0.5$ , $r=0.7$ , $r=0.9$ ; The vertical axes are the variances and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity . . . . .	203
6.28	Relationship between Weighted Variance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right $r=0.1$ , $r=0.3$ , $r=0.5$ , $r=0.7$ , $r=0.9$ ; The vertical axes are the Weighted Variance and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity . . . . .	204
6.29	Relationship between Unweighted Covariance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right $r=0.1$ , $r=0.3$ , $r=0.5$ , $r=0.7$ , $r=0.9$ ; The vertical axes are the Unweighted Covariance and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity . . . . .	205



6.30	Relationship between Weighted Covariance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right $r=0.1$ , $r=0.3$ , $r=0.5$ , $r=0.7$ , $r=0.9$ ; The vertical axes are the Weighted Covariance and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity . . . . .	206
6.31	Relationship between Correlations and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right $r=0.1$ , $r=0.3$ , $r=0.5$ , $r=0.7$ , $r=0.9$ ; The vertical axes are the Correlations and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity . . . . .	207
6.32	Unweighted Variance of Y and Covariance of X and Y at Ward level .	209
6.33	Correlation at Ward level with different autocorrelation . . . . .	210
6.34	Level 1 and Level 2 Variance component of variable Y . . . . .	211
6.35	Level 1 Pure Correlation at different degrees of autocorrelation in Y .	212
6.36	Level 1 and Level Pure Regression . . . . .	213
6.37	Scatter Plot of the Different Moran's I and the Corresponding IAC .	214

# List of Tables

4.1	Range of values for the categories . . . . .	39
4.2	Moran's I . . . . .	41
4.3	Correlation and regression coefficients at different scales, X and Y both have low autocorrelation . . . . .	42
4.4	Intra-Area correlations and variance components of X, X have low autocorrelation . . . . .	43
4.5	Intra-Area correlations and variance components of Y using moments, Y have low autocorrelation . . . . .	44
4.6	Intra-Area cross-correlations and covariance components at two lev- els, X and Y both have low autocorrelation . . . . .	44
4.7	Pure correlations and regressions, X and Y both have low autocorre- lation . . . . .	45
4.8	Description of Unweighted Variance of X and Covariance of X and Y, X and Y both have low autocorrelation . . . . .	48
4.9	Description of Weighted Variance of X and Covariance of X and Y, X and Y both have low autocorrelation . . . . .	48
4.10	Description of the Pearson Correlation, X and Y both have low au- tocorrelation . . . . .	50
4.11	Expected correlation and the variance and SD at different levels of aggregation assuming no autocorrelation . . . . .	51
4.12	Description of Regression Coefficient, X and Y both have low auto- correlation . . . . .	53

4.13	Description of the Level 2 and Level 1 Variance Components, X and Y both have low autocorrelation . . . . .	55
4.14	Description of the Intra-Area Correlation, X have low autocorrelation	55
4.15	Description of the Intra-Area Cross-Correlation, X and Y both have low autocorrelation . . . . .	56
4.16	Description of the Level 2 and Level 1 Pure Correlation, X and Y both have low autocorrelation . . . . .	59
4.17	Description of the Level 2 and Level 1 Pure Regression, X and Y both have low autocorrelation . . . . .	61
4.18	Moran's I . . . . .	61
4.19	Unweighted variance and covariance, X and Y both have medium autocorrelation . . . . .	61
4.20	Weighted variance and covariance, X and Y both have medium autocorrelation . . . . .	62
4.21	Correlation and regression coefficients at different scales, X and Y both have medium autocorrelation . . . . .	62
4.22	Intra-Area correlations and variance components of X, X have medium autocorrelation . . . . .	63
4.23	Intra-Area correlations and variance components of Y, Y have medium autocorrelation . . . . .	63
4.24	Intra-Area cross-correlations at two levels, X and Y both have medium autocorrelation . . . . .	63
4.25	Pure correlations and pure regression at two levels, X and Y both have medium autocorrelation . . . . .	64
4.26	Description of Unweighted Variance of X and Covariance of X and Y, X and Y both have medium autocorrelation . . . . .	67
4.27	Description of Weighted Variance of X and Covariance of X and Y, X and Y both have medium autocorrelation . . . . .	68
4.28	Description of Pearson Correlation, X and Y both have medium autocorrelation . . . . .	68

4.29 Description of Regression Coefficient, X and Y both have medium autocorrelation . . . . .	69
4.30 Description of the Level 2 and Level 1 Variance Components, X and Y both have medium autocorrelation . . . . .	72
4.31 Description of Intra-Area Correlation (X), X both have medium autocorrelation . . . . .	73
4.32 Description of the Level 2 and Level 1 Pure Correlation, X and Y both have medium autocorrelation . . . . .	73
4.33 Description of the Level 2 and Level 1 Pure Regression, X and Y both have medium autocorrelation . . . . .	76
4.34 Moran's I . . . . .	76
4.35 Unweighted variance and covariance, X and Y both have high autocorrelation . . . . .	77
4.36 Weighted variance and covariance, X and Y both have high autocorrelation . . . . .	77
4.37 Correlation and regression coefficients at different scales, X and Y both have high autocorrelation . . . . .	78
4.38 Intra-Area correlations and variance components of X, X have high autocorrelation . . . . .	78
4.39 Intra-Area correlations and variance components of Y, Y have high autocorrelation . . . . .	79
4.40 Intra-Area cross-correlations at two levels, X and Y both have high autocorrelation . . . . .	79
4.41 Pure coefficients at two levels, X and Y both have high autocorrelation	80
4.42 Description of the Unweighted Variance(X) and Covariance(X,Y), X and Y both have high autocorrelation . . . . .	80
4.43 Description of the Weighted Variance(X) and Covariance(X,Y), X and Y both have high autocorrelation . . . . .	82
4.44 Description of Pearson Correlation, X and Y both have high autocorrelation . . . . .	84

4.45	Description of Regression Coefficient, X and Y both have high autocorrelation . . . . .	85
4.46	Description of the Level 2 and Level 1 Variance Component X, X have high autocorrelation . . . . .	87
4.47	Description of Intra-Area Correlation of X, X have high autocorrelation	88
4.48	Description of Intra-Area Cross-Correlation, X and Y both have high autocorrelation . . . . .	88
4.49	Description of the Level 2 and Level 1 Pure Correlation, X and Y both have high autocorrelation . . . . .	90
4.50	Description of the Level 2 and Level 1 Pure Regression, X and Y both have high autocorrelation . . . . .	92
4.51	Moran's I and Unweighted Variance . . . . .	92
4.52	Theoretical Expected Value and Standard Deviation of Correlation of Data sets 1, 2, and 3 . . . . .	94
4.53	Summary of Correlation of Data sets 1, 2, and 3 . . . . .	94
4.54	Summary of Pure Correlation of Data sets 1, 2, and 3 . . . . .	95
4.55	Summary of Pure Regression of Data sets 1, 2, and 3 . . . . .	96
4.56	Moran's I . . . . .	97
4.57	Unweighted Variance and Covariance, X and Y both are not autocorrelated . . . . .	97
4.58	Weighted Variance and Covariance, X and Y both are not autocorrelated . . . . .	98
4.59	Correlation and regression coefficients at different scales, X and Y both are not autocorrelated . . . . .	98
4.60	Intra-Area correlations and variance components of X, X not autocorrelated . . . . .	99
4.61	Intra-Area correlations and variance components of Y, Y both not autocorrelation . . . . .	99
4.62	Intra-Area Cross-Correlation, X and Y both are not autocorrelated .	100
4.63	Pure correlations at two levels, X and Y both are not autocorrelated .	100

4.64	Pure Regressions at two levels, X and Y both are not autocorrelated .	101
4.65	Description of Pearson Correlation, X and Y both are not autocorrelated	102
4.66	Theoretical Mean, Standard deviation, Lower and Upper Limits of group-level correlation . . . . .	102
4.67	Description of the Level 2 and Level 1 Variance Component of X, X not autocorrelated . . . . .	104
4.68	Description of Intra-Area Correlation of X, X not autocorrelated . . .	104
4.69	Description of Intra-Area Correlation of Y, Y not autocorrelated . . .	105
4.70	Description of the Level 2 and Level 1 Pure Correlation, X and Y both are not autocorrelated . . . . .	109
4.71	Description of the Level 2 and Level 1 Pure Regression, X and Y both are not autocorrelated . . . . .	109
4.72	Summary of Correlations at Different Degrees of Autocorrelation and Levels of Aggregation . . . . .	111
4.73	Summary of Regression Coefficient at Different Degrees of Autocor- relation and Levels of Aggregation . . . . .	111
4.74	Summary of Covariance of X and Y at Different Degrees of Autocor- relation and Levels of Aggregation . . . . .	113
4.75	Summary of Level 2 Pure Correlation at Different Degrees of Auto- correlation and Levels of Aggregation . . . . .	114
4.76	Summary of Level 1 Pure Correlation at Different Degrees of Auto- correlation and Levels of Aggregation . . . . .	116
4.77	Summary of Level 2 Pure Regression at Different Degrees of Auto- correlation and Levels of Aggregation . . . . .	117
4.78	Summary of Level 1 Pure Regression at Different Degrees of Auto- correlation and Levels of Aggregation . . . . .	118
4.79	Summary of correlations when X have low autocorrelation and Y have different levels of autocorrelation at different levels of aggregation . .	124
5.1	Individuals counts from SAR and SAS and number of EDs and Wards for each district . . . . .	135

5.2	Mean and Individual Level Variances from SAR and different levels from Census . . . . .	140
5.3	Variance-Covariance matrix Individual Level: SAR . . . . .	140
5.4	Correlations at Individual level . . . . .	140
5.5	Weighted Mean and Variances from SAS (ED level) . . . . .	141
5.6	Variance-Covariance matrix ED Level . . . . .	142
5.7	Correlations at ED level . . . . .	142
5.8	Variable combinations and correlations at different levels . . . . .	143
5.9	Variance components and Intra-area correlation . . . . .	143
5.10	Aggregation Effect and Intra-Area Correlation . . . . .	144
5.11	Intra-area Cross-correlation . . . . .	145
5.12	Level 1 Pure correlation . . . . .	145
5.13	Level 2(ED level) Pure correlation . . . . .	146
5.14	Weighted Mean and Variances from SAS (Ward level) . . . . .	148
5.15	Variance-Covariance matrix Ward Level . . . . .	149
5.16	Correlations at Ward level . . . . .	149
5.17	Variance components and Intra-area correlation . . . . .	150
5.18	Intra-Ward Cross-Correlations . . . . .	151
5.19	Level 1 Pure correlation . . . . .	151
5.20	Level 2 (Ward level) Pure correlation . . . . .	151
5.21	Correlations at Different Levels . . . . .	154
5.22	Correlations at Individual Level (from SAR) and Level 1 Pure Cor- relations when level 2 are ED and Ward Levels (fromSAS) . . . . .	155
5.23	Correlations at Ward Level (from SAS) and Level 2 Pure Correlations when level 2 is Ward (from SAS) and level 1 is Individual (from SAR)	157
5.24	Correlation and regression coefficients at different scales . . . . .	157
5.25	Some Statistics derived from multilevel model . . . . .	158
5.26	Correlation and regression coefficients at different scales . . . . .	158
5.27	Some Statistics derived from multilevel model . . . . .	158
5.28	Correlation and regression coefficients at different scales . . . . .	159

5.29	Some Statistics derived from multilevel model . . . . .	159
5.30	Pearson Correlations at ED and Ward Levels . . . . .	160
5.31	Pure Correlations . . . . .	162
5.32	Aggregation effects on the variances (diagonal, bold) and covariances (off-diagonal) . . . . .	162
5.33	Variance component and Intra-Ward Correlation . . . . .	163
5.34	Intra-Area Cross Correlation . . . . .	164
5.35	Regression coefficients at ED and Ward levels . . . . .	165
5.36	Level 1 and Level 2 Pure Regression Coefficients . . . . .	165
5.37	Moran's I at different weight definition . . . . .	166
5.38	Moran's I with different proximity matrices Ward level . . . . .	166
5.39	Bivariate Moran using GeoDa (EDs within Ward) . . . . .	167
6.1	Description of the distribution of Weighted Variances of X at Ward level . . . . .	182
6.2	Description of the Weighted Covariances at Ward level . . . . .	183
6.3	Description of the direct correlations at different levels of autocorre- lations . . . . .	184
6.4	Description of the Level 1 Pure Correlation at different levels of au- tocorrelations . . . . .	188
6.5	Description of the Level 2 Pure Correlation at different levels of au- tocorrelations . . . . .	188
6.6	Description of Variance of X and Covariance (Y,X) at different degrees of autocorrelations . . . . .	188



# Chapter 1

## Introduction

This chapter introduces some definitions of key terms and the Modifiable Areal Unit Problem (MAUP), spatial autocorrelation, and multilevel model. Also the chapter describes the problems to be tackled in the thesis and the flow of presentation.

### 1.1 The Modifiable Areal Unit Problem

Analysis using spatial data is a multi-disciplinary subject attracting the attention of statisticians, geographers, physical and social scientists. A geographical region may be completely covered by a number of mutually exclusive zones referred to as areal units. In many instances data are available as aggregated measurements for a set of areal units that are arbitrarily defined in terms of number and boundaries. The areal units can be partitioned into smaller subareas or grouped into larger areas in a hierarchical manner (Wong, 1996), or boundaries can be changed for some reason. In Australia an example of a hierarchical geographical structure is, from smallest areal unit to a larger unit; Census Collectors Districts (CDs), Local Government Areas (LGAs), Statistical Divisions (SDs) and States/Territories. In the United Kingdom an example of an hierarchical structure is; Enumeration Districts (EDs), Wards, and Districts.

The results of statistical analyses based on the data available for areal units vary according to the definition of the areal units. Any statistical relationship may be manipulated by the choice of areal units (Openshaw, 1984b). This phenomenon is referred to as the *modifiable areal unit problem* (MAUP). The term was first used and defined by Openshaw and Taylor (1979) (see Fotheringham and Wong, 1991). The modifiable areal unit problem reflects not only the properties of the variables under consideration but also the properties of the zoning system itself (Yule and Kendal, 1950). Statistical analysis based on data aggregated over spatial units often produce results that are very different from those obtained from analyzing corresponding individual or household level data (Steel, Holt and Tranmer, 1996). One approach would then seem to be to only use data at the lowest possible level of aggregation (Goodchild, 1992), which may be the individual level. There are often reasons that it is necessary to aggregate data. One reason is to reduce the volume of data to be processed. Another reason is that it protects the confidentiality of personal data (Openshaw and Albanides, 1996). A further reason is that there may be no interest in purely individual level relationships, but in relationships at some higher level of aggregation. Many analyses are tied to arbitrarily defined areal units and the results apply only for the particular areal units that have been used. Methods that eliminate or minimize the impact of the MAUP or have predictable qualities when areal units are changed will be of enormous value.

The MAUP is the sensitivity of results to the definition of the areal units for which the data are available. These results may vary with the level of aggregation and the configuration of the zoning system. The MAUP consists of two sub-problems: the scale problem and the zoning problem (Openshaw and Taylor, 1979). The scale problem refers to the variation in results that may be obtained when the same areal units are combined into sets of increasingly larger areal units for analysis (Openshaw and Taylor, 1979). It is the change in results that occurs as the number of areal units into which the population is partitioned changes. The zoning problem refers to the variability in results when different boundaries are used at the same scale, that is, for the same number of areal units (see Wrigley, Holt, Steel, and Tran-

mer, 1996). The term "modifiable" is used because the choice of area boundaries and the number of areas used to cover the population are often not fundamental and other choices could have been made (Holt, Steel and Tranmer, 1996a). Usually the areal units used to identify the geographical location of the objects being studied have no special significance having been constructed for reasons of cost, operational or administrative convenience (Steel, Holt and Tranmer, 1996a).

Another issue that is related to the MAUP is the *ecological fallacy*. It occurs when spatially aggregated data are analyzed and the results are assumed to apply to relationships at the individual level. It arises when group or area level data are the only source of information available to the researcher but the objective of the study are individual level characteristics and relationships (Wrigley et al., 1996). In ecological analysis, the main data available consist of group or area level means or totals from a census or sample but the targets of inference are at the unit level (Holt, Steel, and Tranmer, 1997).

Socio-economic differences between arbitrarily defined areal units contribute to the effects of the MAUP on statistical analysis. In practice, individuals who live in the same area tend to be more alike in terms of a variety of socio-economic variables than individuals in different areas. This is referred to as positive clustering and is characterized by positive intra-area correlation, a statistic which measures the homogeneity of individuals within areas or groups (Holt et al., 1996a). Choices of areal unit boundaries may create areas that are relatively homogeneous, whereas other choices of boundaries may result in areas that are less homogeneous and thus the MAUP occurs. The MAUP will usually affect different variables to varying degrees, leading to the unpredictable scale and zoning effects on the relationships between variables (Holt, et. al., 1996a). The analysis of spatial data is usually complicated by the complex relationship between variables, the spatial pattern of variables and the existence of spatial autocorrelation. There are three kinds of effects that can lead to spatial clustering being important. One is the tendency for people with similar attributes to choose to live near each other. Another is that people in the same area experience the same effects of characteristics of the area. Lastly, the

tendency for people living nearby to interact and develop common characteristics (Steel et al., 1994).

## 1.2 Spatial Autocorrelation

Spatial autocorrelation is a measure of the correlation between values of a variable with regard to spatial location. It measures the level of spatial interdependence of the characteristic and strength of the dependence. Spatial autocorrelation can be categorized as either positive or negative. A positive spatial autocorrelation implies that similar values appear close together and a negative autocorrelation has dissimilar values appearing close together. Spatial autocorrelation basically measures correlation of a single variable for all pairs of points at a particular distance or some other category. Standard global and some new local spatial statistics have been developed to detect spatial autocorrelation and spatial association. These measures include: Moran's I, Geary C, G Statistic, LISA (Anselin, 1995), GLISA (Bao and Henry, 1996).

The analysis of spatial data is usually complicated by the complex relationships between variables and the existence of spatial autocorrelation. The smoothing effect that results from averaging is a contribution to the scale problem in the MAUP. As heterogeneity among units is reduced through aggregation the similarity among units is also reduced. Another factor is spatial autocorrelation. The decrease in the variance is moderated by the positive autocorrelation of the original observations and is worsened by negative autocorrelation (Gotway and Young, 2002). This means that the more the variable is positively autocorrelated, the more the chance that similar values are grouped together when aggregated so that less variance is lost at the aggregate level. As the level of autocorrelation decreases and approaches a negative autocorrelation, the chances increase that non-similar values are grouped together resulting in a greater loss of variance in the aggregate level.

In this thesis Moran's I will be used to describe the spatial autocorrelation of a given variable. Moran's I was the first measure of spatial autocorrelation and was

introduced by Moran (1950) to study stochastic phenomena that are distributed in space in two or more dimension. It has been used in almost all studies employing spatial autocorrelation. The value of Moran's I range from 1 indicating strong positive correlation, to 0 indicating a random pattern, to -1 which implies strong negative spatial autocorrelation. This statistic can be used to measure spatial autocorrelation of ordinal, interval or ratio data. The Moran's I provides a one number overall measure of spatial autocorrelation.

### 1.3 Multilevel Modeling

Over the past 20 years multilevel modeling has been used in many applications. Researchers in social, geographical, education and medical sciences utilize multilevel modeling when the data have a heirarchical structure. Examples are school/children, grouped into schools which may then be grouped into districts. In spatial analysis, data may be collected for areal such as EDs and Wards. In this case the lowest level maybe the household unit and the possible next level will be EDs and then Wards.

The fundamental principle of multilevel modeling is the existence of different levels of variation. The methodology is an extension of multivariate regression in which lower level (say, level-1) and higher level (say, level-2) effects are combined in a model so that both lower level and higher level variation can be investigated. Multilevel modeling can be used to isolate variation resulting from the variability in the lower level from variation resulting from differences between zones. If one is interested in examining lower level data, variation of a particular variable is not only a function of attributes at that level but also that of higher level factors. Goldstein (1998) noted that the application of multilevel modeling has begun to produce new insights in several areas because relevant software has become more widely available. He described multilevel approaches to research in education and other areas of applications. Goldstein (1998) introduced some of the more recent extensions of multilevel modeling and illustrated their potential for analysing social processes. Multilevel modeling is another approach that allows for across unit correlations.

In a series of papers Steel and Tranmer (1998) have developed an approach that tackles the MAUP and ecological fallacy using a multilevel modeling framework. They applied the approach to investigate scale effects. The basic model they used involves assuming that individuals within an areal unit are all equally correlated with each other and that there is no correlation between individuals in different areal units. Basically, the approach considers the average within areal unit correlation between individuals and ignores the association across areal units. A more general approach can be developed which allows correlation between different individuals to depend on their spatial location.

Multilevel modeling was originally developed in situations when the levels correspond to non-spatial groups such as schools and hospitals. It has also become a popular approach for analysing geographic data (Jones and Duncan, 1996). For geographical data the levels in a multilevel model can correspond to individual, neighbourhoods, and other higher level of geographic units such as administrative areas, regions or provinces. The effects at a particular level may reflect many influences that operate at that level such as local policies, physical features, or interactions between people. The impact of such factors may not be so clear cut. Therefore a basic issue is how multilevel modelling itself is affected by the MAUP.

## 1.4 The Problem

This thesis is going to address some issues on how a simple multilevel model is affected by the MAUP. In Chapter 2 a review of research on the MAUP, including empirical investigations, theoretical studies, and some methods previously employed, will be conducted.

In Chapter 3 some theoretical background will be discussed to investigate the causes of the MAUP. Definitions of some statistics that are relevant to the study will be presented in this chapter as well as relationships between pertinent statistics.

Initial investigation of the possibility of multilevel modeling as a solution to one aspect of MAUP, the scale effect, will be presented in Chapter 4. Several experiments

injecting some specified conditions such as spatial autocorrelations in simulated data sets in a square grid will be conducted to look into the scale effects on pertinent statistics.

In Chapter 5 real data from the 1991 UK Census will be used to investigate the scale effect of pertinent statistics. Several scenarios will be investigated when individual level data are available and when no individual level data are available.

In Chapter 6 an actual region divided into Enumeration Districts (EDs) and Wards will be used to generate data sets with different initial conditions. The behaviors of various statistics, including statistics derived from the simple multilevel model will be examined in this chapter. The actual region used in Chapter 6 will be used to generate some arbitrary boundaries to examine the other aspect of the MAUP, the zonation effect.

Chapter 7 provides a summary and conclusions.

To sum it up, the thesis will look into the MAUP effects of some statistics derived from the simple multilevel model using real data and simulated data sets. This thesis will give insight into how simple multilevel modeling and other pertinent statistics are affected by the MAUP under various conditions.

In particular, the experiments in these thesis investigate a number of questions concerning scale effects.

## Chapter 2

# Review of Research into the MAUP

The research described in this thesis builds on some previous results in the study of the MAUP. This chapter briefly reviews some previous research relevant to this thesis.

### 2.1 Introduction

After the initial discovery of the MAUP, several lines of research have been followed. A large amount of the research on the MAUP focused on revealing the problem and was devoted to assessing the magnitude and impact on standard statistics such as correlation and regression coefficients. Various researchers have conducted studies to examine the effect of varying scale and aggregation on correlation and regression coefficients. However, the approaches were mainly empirical and there was little effort to provide a theoretical explanation or solution.



## 2.2 Research on the MAUP

Aspects of the MAUP were first raised by Gehlke and Biehl (1934) when they conducted an empirical study that was motivated by an issue pointed out by Dr. Henry Sheldon in 1931 that stated: "*a tendency for the correlation coefficient to increase in size as the units of census tract areas increase in size from one tract to several, and decrease in number of tracts from 188 to 23*" (page 169, Gehlke and Biehl, 1934). There are three parts of the study and one concerned the grouping effects in census tract data. The 252 census tracts of Cleveland were successively grouped into areas in such a way, that as much as possible, they had approximately the same size and were made up of contiguous territory. They grouped the 252 census tracts into 200, 175, 125, 100, 50, and 25 areas. They found that the correlation coefficients between median monthly rental payment and juvenile delinquency increased when areal units became larger. One of their conclusions was that the magnitude of the correlation coefficient seems to be affected by the changes of the size of the unit used in such a way that a smaller value was associated with the use of the smaller areal units.

Robinson (1950) provides empirical evidence that an ecological correlation is not equal to its corresponding individual correlation in his studies of illiteracy and colour and illiteracy and foreign birth. Yule and Kendal (1950) noted the values of correlation coefficients depended on the size of the unit and the tendency for the correlations to increase with the size of groups. Blalock (1964) assessed the impact on the correlation and slope estimates of a bivariate linear model under four different aggregation criteria; random, by the dependent variable, by the independent variable, and by proximity. There are originally 150 counties in Blalock's study. He then formed artificial groupings of 75, 30, 15, and 10 groups. The results showed that random grouping had no impact on the correlation coefficients and regression coefficient. The counties are then ranked according to scores of the independent variable and then grouped so that the first group had the first  $n$  lowest scores, the next group had the next  $n$  lowest scores, and so on, where  $n$  is the number of coun-

ties in a group. The correlation and regression coefficients are then computed for the different groupings. The correlation coefficient was observed to increase with scale but no effect on the slope coefficient was observed. For grouping by the dependent variable, both the correlation coefficients and the slope coefficient increased with scale. The increase in the magnitude of the correlation coefficient was of the same magnitude as for grouping by the dependent variable. Grouping by proximity resulted in increases in the correlation coefficient and the slope coefficient and the result was closer to the results of grouping by the independent variable than grouping by the dependent variable. He commented that grouping by proximity may in some degree involve units being put together to maximize variation in either variables and will affect the correlation and regression coefficients (Blalock, 1964).

Clark and Avery (1976) conducted investigations into bivariate relationships by examining the scale effect in a simple regression model. Part of their study was to compare correlation and regression coefficients at individual level and the census tract groupings of the data. The independent variable was a measure of the level of education of the head of the household and the dependent variable was a measure of family income. Individual level household data (952 households) were obtained from the Los Angeles Metropolitan Area Survey (LAMAS) conducted in 1972. They were able to use 1556 Census tract units in Los Angeles County for 1970. In addition, they used two government groupings as aggregate units: 134 Welfare Planning Council Study areas and 35 Regional Planning Commission Statistical Areas. They found that the correlation coefficient and regression coefficient of the spatially aggregated data tends to increase in comparison with the individual household level and that the coefficients tended to vary at different levels of aggregation. To investigate aggregation effects they used the data derived from LAMAS. The groups were formed using the criterion of spatial proximity and made as spatially compact as possible and include the following; 136 groups of 7 individuals, 68 groups of 14 individuals, 34 groups of 28 individuals, and 17 groups of 56 individuals. The correlation and the regression coefficient between the two variables considered tended to increase with the level of aggregation, but irregularity happened in the fifth level when both coeffi-

cients decreased below their corresponding values at the fourth level. In conclusion, they claimed that *"from the empirical evidence of their study, spatial aggregation of data has significant consequences in the correlation and regression analysis of areally distributed phenomena"* (Clark and Avery, 1976, p 436). They also suggested *"that the deviations of the observed from the expected behavior of the coefficients are related directly to the manner in which the covariation between the independent and dependent variables changes with increased aggregation, and indirectly to the way in which spatial autocorrelation is exhibited among the micro- and macrolevel data"* (Clark and Avery, 1976, p 436).

Taylor (1977) reviewed the work of Blalock (1964) and suggested that the effect of rising correlation with rising scale was related to spatial autocorrelation. Openshaw and Taylor (1979) reported the results from three closely related experiments on the variation in the correlation coefficients under different spatial and statistical conditions. The purpose of the experiments was to increase the understanding of the MAUP from both geographical and statistical perspectives. The first experiment was carried out with the use of a set of data describing Iowa, USA. They used the 99 counties of Iowa as the basic areal units and for each unit they used two measures: the percentage vote for Republican candidates in the congressional election of 1968 as the dependent variable and the percentage of population over sixty-years old recorded in the 1970 US census as the independent variable. The 99 counties were combined into five different areal arrangements with six areal units. Correlation coefficients for the five areal arrangements were computed and only one of the five coefficients is below the correlation coefficient computed from the basic areal units. To identify the limits of the scale and aggregation problem they apply an automatic zoning algorithm that identifies zonings or groupings of data that approximately optimize any general function defined in terms of the aggregated data. To produce zoning and grouping distributions of correlation coefficients they used a random zoning and grouping system generator. One of the main observations of the first experiment was *"There seems to be very distinct differences between zoning and grouping systems in many situations and these seem to be caused by the*

*interaction of the contiguity in the zoning with the spatial autocorrelation in the data.*" (Openshaw and Taylor, 1979, page 137). The second experiment was to investigate the effects of spatial autocorrelation on the correlation coefficient and the sum-of-squares terms. They used a data set generator based on the quadratic loss function to construct artificial data. The data set generator was used to produce new variables with the known properties of the 99 Iowa counties. Two sets of data were generated in their simulation with the properties of the Iowa data and were designed in such a way that they differed only in terms of autocorrelation. One data set having maximum positive spatial autocorrelation and the other normally distributed. The conclusion drawn from the second experiment was *"that zoning and spatial autocorrelation do interact in quite predictable ways and that this interaction explains much of the variety of results previously obtained from the original autocorrelated Iowa data."*(Openshaw and Taylor, 1979, page 140). The third experiment's objective was to obtain a more thorough understanding of the relationship between sum-of-squares and correlations that can be obtained from random arrangements. The third experiment resulted in their claim *"the expected relationship between the sum-of-squares term and the correlation coefficient was found to be more illusive than initially expected."* (Openshaw and Taylor, 1979, page 142). They felt that the MAUP is much more complex than had been previously believed.

Arbia (1989) considered the relationship between the MAUP and the spatial configuration of the data for both univariate and bivariate statistical analysis. He looked at a framework that not only takes into account the size of the area but also the interconnectedness and dependence of areal units.

The effects of the MAUP on multivariate statistical analysis were investigated by Fotheringham and Wong (1991), who examined the impact of scale and zoning effects on two multivariate models, a multiple linear regression model and a multiple logit regression model. Data were from 871 block groups in the Buffalo Metropolitan Area for the 1980 US census. For the scale investigation, the 871 block groups were aggregated randomly and contiguously to scales of 800, 400, 200, 100, 50 and 25 areal units and 20 different aggregations were used at each scale. Several statistics

were examined, including the regression parameters, the standard errors of these parameters, confidence intervals, Moran's Coefficients, t values, and  $r^2$ . For both models when scale was varied, the regression parameters increased or decreased with scale. The increase or decrease depending on the relationship (negative or positive) between dependent and independent variables and the model. The standard errors of parameter estimates increased as the number of zones is decreased. There was no obvious relationship between the level of spatial autocorrelation and the severity of the MAUP. To investigate the zoning problem, the 871 block groups were randomly aggregated 150 different ways using contiguity constraints, at the same scale of 218 zones. Regression coefficients values varied according to the zoning system so that values ranged from positive to negative values. They claimed that it is important for the multivariate analysis that further analysis of the MAUP be presented to uncover insights into the sensitivity of the analytical results to both scale and zoning variations (Fotheringham and Wong 1991).

Amrhein (1995) explored the nature and extent of the scale effect and zonation effect and to challenge the notion in the literature that aggregation effects are pervasive and unpredictable. The paper focused on the following question: *"Are the effects currently described as aggregation effect at least partly a result of methodological considerations relating to the appropriateness of the statistics chosen and their application?"* (page 108, Amrhein, 1995). He used simulated data from predetermined distributions. Values for locations on a continuous region containing 10,000 locations, which represent addresses for individuals, were generated. The addresses were generated using a uniform distribution for the variables x and y and then a normal  $N(0,1)$  distribution. Each location was then given values from randomly generated values from a uniform distribution and then a normal distribution. This resulted in four sets of data based on the selected distribution for addresses and values of the variables. The 10,000 observations were taken as the population. To investigate the scale effect, the individuals were aggregated into 100, 49, and 9 square areal units. He also investigated the zonation effect, the effect when different definition of the boundaries are used while holding the scale constant. From the

result of the experiments Amrhein (1998) came up with some "*aggregation rules*":

1. The mean does not display any pronounced aggregation effects (scale or zonation) at any level of aggregation used in the study.
2. The variance does not display any pronounced scale effect beyond those expected from the decrease in the number of observations. However, it was noted that scale-specific variance values cannot be imputed to other scales without adjusting for the change in the number of reporting units.
3. Populations with high variances tend to exhibit more pronounced zonation effects than populations with smaller variance.
4. The regression coefficient does not display scale effects that increase systematically with decreasing number of zones.
5. The standard deviations of the regression coefficient display pronounced zonation effects. The standard deviations of the regression coefficient increases to a point at which it fails to provide reliable information (based on the expectation).
6. The Pearson correlation coefficient exhibits systematically increasing aggregation effects as the number of groups decreases. The range and standard deviations of coefficients calculated in the experiment ultimately span the range of the statistics.

Steel and Holt (1996) presented both theoretical and empirical results on random aggregation. They used the term *aggregation effect* as the effects observed when individuals are allocated into spatial groups and the group means are used. They derived aggregation effects on some common statistics when the individuals are randomly grouped and the variate values are independent of the group membership. To investigate aggregation effects, a population of  $N$  individuals with the associated variables  $X$  and  $Y$  was divided into  $M$  random groups. To look deeper into their theoretical results, they used the same simulation design used by Amrhein (1995). They generated 10000 locations by using uniform and normal distributions and values for the variables. The region was then divided into 100, 49, and 9 zones. From the results, both theoretical and empirical, they formulated rules for random aggregation and some of them are presented below:

1. The expected value of weighted group-level statistics are not affected by

aggregation and that any observed change is due to random variation.

2. The variance of the weighted group-level statistics are affected mainly by the number of groups in the analysis. The variation will be high when the number of groups is small.

3. The weighted correlation and regression coefficients calculated using  $m$  areas have the same properties as coefficients calculated from  $m$  individuals.

Green and Flowerdew (1996) investigated the effects of aggregation on correlation and regression analysis of spatially correlated data generated using three types of aggregation; random, systematic, and spatial. The effects of aggregation on the correlation coefficient and regression coefficient were recorded. In comparison to results obtained from raw data, random aggregation did not change the values of the correlation and regression coefficients but the standard errors increased. Systematic aggregation caused a large increase in the correlation coefficient but did not affect the regression coefficient or the standard error. Spatial aggregation caused a large increase in the correlation coefficient and smaller increases in the regression coefficient and standard error. They concluded from the results that if  $X$  was spatial autocorrelated, the correlation coefficient displayed the scale effect but the regression coefficient was not affected. They also conducted another experiment where the data were simulated with spatial autocorrelation in  $Y$  and not in  $X$ . The correlation coefficient decreased for the aggregated data while the regression coefficient was unchanged. The experiment was continued using an extension to the standard regression equation to incorporate regional and local effects. The result from this experiment showed that in the presence of autocorrelation, the correlation coefficients demonstrated the MAUP effects. The regression coefficients do not exhibit the MAUP unless there is spatial cross-correlation between the independent and the dependent variables. From their simulation results and theoretical considerations, the sum of the two coefficients  $b(\text{local})$  and  $b(\text{regional})$  displayed no inconsistency between analyses at different levels of aggregation (Green and Flowerdew, 1996). They then applied the ideas in real data. They used data from 5 counties from the 1991 Census for Great Britain. The variables considered in their empirical study

were male unemployment rate and ethnicity.

Amrhein (1995) used simulated data to show the effects of the MAUP on various statistics including weighted statistics, means, standard deviations, variances, regression and Pearson correlation coefficients. Amrhein (1995) discussed the concept of an ideal number of aggregates that would reduce computational burden but would not incur too high aggregation effects. Amrhein and Reynolds (1996) used data from the British Census for the county of Lancashire to demonstrate the ability of a modified G (Getis) statistic to predict the effects of aggregation on several variables. They extended this study and confirmed results by using a much larger data set from Toronto Census Metropolitan Area (Amrhein and Reynolds, 1997).

The study by Flowerdew et al.(2001) concerned the relevance of the MAUP to multiple regression. They claimed that the effect of the MAUP on regression coefficients when the response variable is regressed on a set of explanatory variables is dependent on the spatial distribution of all variable involved. Some results of their research are:

1. The MAUP effects on regression results may be generated when there is cross-correlation between the values Y in one zone and the values of X in the zones in the immediate vicinity.
2. From the results of their study they suggested, not unreasonably, that compact zones capture the regional effect better than less compact zones.
3. They also claimed that the results suggested that defining regions individually for each enumeration districts (ED) by taking the average overall its neighbors excluding the ED itself might be better at capturing the regional effect than defining a complete coverage of wards or pseudo-wards for the whole study area.

The paper by Fotheringham and Wong (1991) was the starting point of the study by Flowerdew et al.(2001)



## 2.3 Some Methods Employed to Solve the MAUP

Most of the studies of the MAUP concentrated on identifying parts of the problem rather than providing an overall solution. Fotheringham (1989) suggested methods to get around the MAUP which include: *"(i) the derivation of the "optimal" zoning systems; (ii) the identification of basic entities; (iii) sensitivity analysis; (iv) abandonment of traditional statistical analysis; (v) shifting the emphasis of spatial analysis towards relationships that focus on rates of change"* (page 222, Fotheringham, 1989). Wong (1996) summarized the suggested methods into three categories: (i) data manipulation approach; (ii) technique-oriented approach and (iii) error-modeling approach.

The following subsections contain reviews of some papers that try to find a solution to the MAUP as categorized by Wong (1996).

### 2.3.1 Data Manipulation Approach

The data manipulation approach is based on the belief that if the selected zoning system can be justified in some way instead for administrative convenience, the MAUP would vanish (Wong, 1996).

Molering and Tobler (1972) used analysis of variance techniques to partition the total variation between the lowest level of geographic areas into components attributable to various aggregation levels in situations where they had nested hierarchical geographic data. The paper presented a method for examining geographical scale effects in data available from sources such as the census. They claimed that the most disaggregated level data are a linear combination of the mean at the disaggregate level and the effects from the different levels of aggregation and that the variance can be partitioned into parts attributable to the different aggregation levels. The method can assign variances to different levels of aggregation starting from the most disaggregate level. The variances can then indicate at which scale the action is taking place and thus isolate the most important level or levels of aggregation. This method is not a complete solution to the MAUP because the technique is not

capable of accommodating multivariate situations and it fails to deal with zoning or aggregation effects as it requires an a priori definition of the hierarchy to identify the aggregation level with the most action (Wong, 1996).

Another approach to deal with the MAUP is the concept of optimal zoning first proposed by Openshaw (1977). In this empirical approach, an ideal or optimal zonal configuration could be achieved. Basic areal units should be aggregated to maximize or minimize whatever criteria are used to evaluate the performance of the model (Openshaw, 1977a, 1977b).

### 2.3.2 Technique-Oriented Approaches

The technique-oriented approach is based on the belief that the MAUP effects might have been caused by using inappropriate models or techniques and thus new techniques should be developed (Wong, 1996).

One of the first proposed solutions to the MAUP was suggested by Robinson in 1956. He proposed that weighting areal units by the areas of the units when computing the regression coefficient is necessary. He claimed that significant discrepancies in size of areal units should be taken into account and this can be accomplished by using the actual areas of the statistical units. The simple weighing scheme proposed by Robinson (1956) fails to correct for the errors propagated by aggregation (Wong, 1996).

Goodman (1959) was the first to seriously consider a model under which ecological inference could validly be used to make inferences considering relationships at an individual level. He considered regression analysis with separate and different regression slopes and intercepts for each group (Holt and Steel, 1996a).

Amrhein and Flowerdew (1989) used Poisson regression model to describe migration flows in Canada. *"The model failed to capture the aggregation effect typical of the MAUP, perhaps because the data were not subject to the MAUP"* (Amrhein and Flowerdew, 1989, p 237).

Tobler (1989) argued that there is no MAUP when the correct analysis procedure

is used and there should be a technique which does not depend on the areal units resulting in frame independent spatial analysis. He also cast doubt on the correlation coefficient as an appropriate measure of association between spatial units.

Several other methods had been used to try to deal with scale and zoning problems. The list of models enumerated by Openshaw (1977a) includes; spatial variate differencing (Curry, 1971), spectral analysis (Rayner, 1971), space-time versions of Box-Jenkins model (Cliff and Ord, 1975).

### 2.3.3 Error Modeling Approaches

Error modeling approaches are based on the idea that *"when analysis moves from one spatial scale to another, relationships among variables and among spatial entities also change"*(Wong, 1996, page 100,). Thus it is necessary to document explicitly these changes and include them in the modeling and analyses.

Steel et al., (1994) attempted to model the error created by the aggregation process so that individual information can be estimated from regional data. This model depends on decomposing the conditional expectation of the variance-covariance matrix at the regional level into a variance-covariance matrix at the individual level, a bias component accounting for aggregation effect on a set of variables called grouping variables, and the residuals from within-group correlation. The grouping variables are a set of variables that characterize the way in which the individuals are clustered within a population of interest. According to Steel et al.(1996a), this model is based on the concept of positive clustering, that is, individuals within areas or groups are usually more alike than between areas. Wong (1996) is unconvinced about this model because it seems *"to apply only classical statistical concepts while failing to deal with the spatial aspect of the MAUP, except in the process of deriving regional level data"* (page 102, Wong, 1996).

### 2.3.4 Some comments and recommendations on how to find a solution of the MAUP

*"The MAUP was regarded as the most stubborn problem in geography and spatial science"* (page 104, Wong, 1996). Wong (1996) stated that the role of spatial autocorrelation in producing the MAUP is evident and that it is expected that the solution of the MAUP is probably *"to depend on how to model the multivariate spatial autocorrelation effect in the multi-scale situations"* (page 105, Wong, 1996).

The early research on the MAUP focused on empirical demonstrations of its existence. Later research has identified the potential role of population structure tied to the areal units and spatial autocorrelation. One popular approach to handling the hierarchical nature of geographical data is multilevel modeling.

Multilevel modeling is used in many projects and most of them look into the variation at different levels. In a series of papers Steel and Tranmer developed an approach to tackle the MAUP using a multilevel framework. This framework reflects a very simple spatial autocorrelation structure, with equal autocorrelation structure, within groups and zero autocorrelation across groups. However, the impacts of more complex spatial autocorrelation were not included in their analysis. This thesis will try to fill that gap. This thesis will look into the impact of spatial autocorrelation on the scale effect of statistics derived from a simple multilevel model.

# Chapter 3

## The Causes of the MAUP

This chapter describe some definitions and theoretical relationships between relevant statistics that are keys to explaining the causes of the MAUP.

### 3.1 Basic Theory

#### 3.1.1 Spatial Aggregation

Suppose we have a region R with N individuals and associated with the individuals are two variable Y and X. The region is divided into M groups or areas by some process. An individual can only belong to one group. The number of individuals in the  $g$ th group is  $N_g$ , where  $g = 1, 2, \dots, M$ .

Given the situation above, we can define some statistics. The means of Y and X are

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i \tag{3.1}$$

and

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i. \quad (3.2)$$

The corresponding population variances are

$$S_{YY}^{(1)} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (3.3)$$

and

$$S_{XX}^{(1)} = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2. \quad (3.4)$$

The population covariance between Y and X is

$$S_{YX}^{(1)} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}). \quad (3.5)$$

When the data are aggregated across the groups, the data available are  $(\bar{Y}_g, \bar{X}_g)$ , for  $g = 1, 2, \dots, M$ , where,

$$\bar{Y}_g = \frac{1}{N_g} \sum_{i \in g} Y_i \quad (3.6)$$

and

$$\bar{X}_g = \frac{1}{N_g} \sum_{i \in g} X_i \quad (3.7)$$

are the group means. From these aggregated data, we can define some unweighted statistics. The means are

$$\tilde{Y} = \frac{1}{M} \sum_{g=1}^M \bar{Y}_g \quad (3.8)$$

and

$$\tilde{X} = \frac{1}{M} \sum_{g=1}^M \bar{X}_g. \quad (3.9)$$

The corresponding variances are

$$\tilde{S}_{YY} = \frac{1}{M-1} \sum_{g=1}^M (\bar{Y}_g - \tilde{Y})^2 \quad (3.10)$$

and

$$\tilde{S}_{XX} = \frac{1}{M-1} \sum_{g=1}^M (\bar{X}_g - \tilde{X})^2. \quad (3.11)$$

The covariance between the groups is

$$\tilde{S}_{YX} = \frac{1}{M-1} \sum_{g=1}^M (\bar{Y}_g - \tilde{Y})(\bar{X}_g - \tilde{X}). \quad (3.12)$$

The aggregated data can also be analyzed using weighted statistics where the weights are the corresponding group populations sizes  $N_g$ . For the weighted statistics, we have,

$$\bar{Y} = \frac{1}{N} \sum_{g=1}^M N_g \bar{Y}_g \quad (3.13)$$

and

$$\bar{X} = \frac{1}{N} \sum_{g=1}^M N_g \bar{X}_g \quad (3.14)$$

are the weighted means at group level. These are exactly the same as the individual level means defined by (3.1) and (3.2) respectively.

The corresponding weighted variances are,

$$S_{YY}^{(2)} = \frac{1}{M-1} \sum_{g=1}^M N_g (\bar{Y}_g - \bar{Y})^2 \quad (3.15)$$

and

$$S_{XX}^{(2)} = \frac{1}{M-1} \sum_{g=1}^M N_g (\bar{X}_g - \bar{X})^2. \quad (3.16)$$

The weighted covariance for the group means is

$$S_{YX}^{(2)} = \frac{1}{M-1} \sum_{g=1}^M N_g (\bar{Y}_g - \bar{Y})(\bar{X}_g - \bar{X}). \quad (3.17)$$

Given the statistics above further analytical statistics can be produced. The correlation of the two variable at the individual level can be computed using

$$r_{YX}^{(1)} = \frac{S_{YX}^{(1)}}{\sqrt{S_{YY}^{(1)}S_{XX}^{(1)}}}. \quad (3.18)$$

The regression coefficients can also be computed. The slope of the regression of Y on X is

$$b_{YX}^{(1)} = \frac{S_{YX}^{(1)}}{S_{XX}^{(1)}} \quad (3.19)$$

and the intercept of the regression of Y on X is

$$a_{YX}^{(1)} = \bar{Y} - b_{YX}^{(1)}\bar{X}. \quad (3.20)$$

Further statistics can also be computed from aggregated data. The unweighted correlation is

$$\tilde{r}_{YX}^{(2)} = \frac{\tilde{S}_{YX}^{(2)}}{\sqrt{\tilde{S}_{YY}^{(2)}\tilde{S}_{XX}^{(2)}}}. \quad (3.21)$$

The slope of the regression of  $\tilde{Y}$  on  $\tilde{X}$  is

$$\tilde{b}_{YX}^{(2)} = \frac{\tilde{S}_{YX}^{(2)}}{\tilde{S}_{XX}^{(2)}} \quad (3.22)$$

and the intercept is

$$\tilde{a}_{YX}^{(2)} = \tilde{Y} - \tilde{b}_{yx}^{(2)}\tilde{X}. \quad (3.23)$$

Corresponding population weighted statistics can be calculated from the group means, giving  $r_{YX}^{(2)}$ ,  $b_{YX}^{(2)}$ ,  $a_{YX}^{(2)}$ .

### 3.1.2 Intra-Area Correlation and Cross Correlation

Tobler's First Law of Geography (Tobler, 1970) states: *"Everything is related to everything else, but near things are more related than distance things"*. In a similar



way, Tranmer and Steel (2001) claimed that individuals in the same area tend to be a little more alike than individuals in different areas and used the term '*with-area homogeneity*' to describe this phenomenon. A measure of within-area homogeneity of a single variable is the intra-area correlation as described by Holt *et. al.*(1996). "*The higher the value of the intra-area correlation, the more similar the values of the variable are for different individuals within the same areas*". Consider the model for a single variable of interest Y:

$$Y_i = \mu_Y + \alpha_{Y_g} + \epsilon_{Y_i} \quad \text{for } i \in g \quad (3.24)$$

where

$Y_i$  represents the value of Y for the  $i$ th individual in area  $g$

$\mu_Y$  is the expectation of Y across the region of interest

$\alpha_{Y_g}$  is a random variable representing the area effect for the  $g$ th area

$\epsilon_{Y_i}$  is a random variable representing the pure individual effect.

Similarly for another variable X, we have:

$$X_i = \mu_X + \alpha_{X_g} + \epsilon_{X_i} \quad \text{for } i \in g \quad (3.25)$$

where

$X_i$  represents the value of X for the  $i$ th individual in area  $g$

$\mu_X$  is the expectation of X across the region of interest

$\alpha_{X_g}$  is a random variable representing the area effect for the  $g$ th area

$\epsilon_{X_i}$  is a random variable representing the pure individual effect.

### Assumptions:

(i) The random variables have population means equal to zero and variance-covariance matrix

$$\Lambda^{(l)} = \begin{bmatrix} \Lambda_{XX}^{(l)} & \Lambda_{YX}^{(l)} \\ \Lambda_{YX}^{(l)} & \Lambda_{YY}^{(l)} \end{bmatrix} \quad (3.26)$$

where  $l=1,2$  indicates the level. Individuals are level 1 and areas that are groups are the level 2 units.

(ii) The random effects are not correlated between levels.

Thus, for variables Y and X the overall variance-covariance matrix is:

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{YX} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} = \begin{bmatrix} \Lambda_{XX}^{(2)} & \Lambda_{YX}^{(2)} \\ \Lambda_{YX}^{(2)} & \Lambda_{YY}^{(2)} \end{bmatrix} + \begin{bmatrix} \Lambda_{XX}^{(1)} & \Lambda_{YX}^{(1)} \\ \Lambda_{YX}^{(1)} & \Lambda_{YY}^{(1)} \end{bmatrix} \quad (3.27)$$

From the model above, important statistics can be formulated.

By considering the expectation of  $S_{YY}^{(1)}$  and  $S_{YY}^{(2)}$  under model (3.24) Tranmer and Steel (1998) show that the group level variance component, which we shall call level 2 variance component, can be approximately unbiasedly estimated by

$$\hat{\Lambda}_{YY}^{(2)} = \frac{S_{YY}^{(2)} - S_{YY}^{(1)}}{\bar{N}^* - 1} \quad (3.28)$$

where

$$\bar{N}^* = \bar{N} + \frac{\bar{N} - \bar{N}^0}{M - 1}, \quad \bar{N}^0 = \frac{1}{N} \sum_{g=1}^M N_g^2, \quad \bar{N} = \frac{N}{M}.$$

Proof:

Tranmer and Steel (1998 ) show

$$E \left[ S_{YY}^{(1)} \right] = \Lambda_{YY}^{(1)} + \left( 1 - \frac{\bar{N}^0 - 1}{N - 1} \right) \Lambda_{YY}^{(2)} \quad (3.29)$$

$$E \left[ S_{YY}^{(2)} \right] = \Lambda_{YY}^{(1)} + \bar{N}^* \Lambda_{YY}^{(2)} \quad (3.30)$$

$$\text{where } \bar{N}^0 = \frac{1}{N} \sum_g N_g^2 = \bar{N}(1 - C_N^2) \quad , \quad C_N^2 = \frac{d_N^2}{\bar{N}^2} \quad \text{and} \quad d_N^2 = \frac{1}{M} \sum_g (N_g - \bar{N})^2$$

Hence

$$\begin{aligned} \frac{\bar{N}^0 - 1}{\bar{N} - 1} &= \frac{\bar{N}(1 - C_N^2) - 1}{M\bar{N}} \\ &\approx \frac{1}{M} (1 + C_N^2) - \frac{1}{M\bar{N}} \quad \text{if } N \text{ is large} \\ &= \frac{1}{M} \left( 1 + C_N^2 - \frac{1}{\bar{N}} \right) \\ &\approx \frac{1}{M} \quad \text{unless } C_N^2 \text{ is large and } \bar{N} \text{ is small} \end{aligned}$$

Rearranging (3.29) and (3.30) gives an unbiased estimate of  $\hat{\Lambda}_{YY}^{(2)}$ ,

$$\hat{\Lambda}_{YY}^{(2)} = \frac{S_{YY}^{(2)} - S_{YY}^{(1)}}{\bar{N}^* - 1 + \frac{\bar{N}^o - 1}{N-1}} \approx \frac{S_{YY}^{(2)} - S_{YY}^{(1)}}{\bar{N}^* - 1} \quad \text{provided} \quad \frac{1}{M} \ll \bar{N}^*.$$

From (3.29) an estimate of  $\Lambda_{YY}^{(1)}$  is

$$\hat{\Lambda}_{YY}^{(1)} = S_{YY}^{(1)} - \left(1 - \frac{\bar{N}^o - 1}{N-1}\right) \hat{\Lambda}_{YY}^{(2)} \approx S_{YY}^{(1)} - \hat{\Lambda}_{YY}^{(2)} \quad \text{provided} \quad \frac{\bar{N}^o - 1}{N-1} \approx \frac{1}{M} \quad \text{is negligible.}$$

Thus an approximately unbiased estimate of the level 1 variance component is

$$\hat{\Lambda}_{YY}^{(1)} = S_{YY}^{(1)} - \hat{\Lambda}_{YY}^{(2)}. \quad (3.31)$$

Similarly, for X,

$$\hat{\Lambda}_{XX}^{(2)} = \frac{S_{XX}^{(2)} - S_{XX}^{(1)}}{\bar{N}^* - 1} \quad \text{and} \quad \hat{\Lambda}_{XX}^{(1)} = S_{XX}^{(1)} - \hat{\Lambda}_{XX}^{(2)}. \quad (3.32)$$

The estimate of the level 2 and level 1 covariance are, respectively,

$$\hat{\Lambda}_{YX}^{(2)} = \frac{S_{YX}^{(2)} - S_{YX}^{(1)}}{\bar{N}^* - 1}, \quad \text{and} \quad \hat{\Lambda}_{YX}^{(1)} = S_{YX}^{(1)} - \hat{\Lambda}_{YX}^{(2)} \quad (3.33)$$

Also to  $O\left(\frac{1}{M}\right)$ ,  $S_{YY}^{(1)}$  is unbiased for  $\Lambda_{YY}^{(1)} + \Lambda_{YY}^{(2)} = \Sigma_{YY}$ .

The *intra-area correlation* for a variable Y is the correlation between the value of Y for two different units within the same group. For the model defined by (3.24), this is equal to

$$\delta_{YY} = \frac{\Lambda_{YY}^{(2)}}{\Sigma_{YY}} \quad (3.34)$$

and can be estimated by

$$\hat{\delta}_{YY} = \frac{\hat{\Lambda}_{YY}^{(2)}}{S_{YY}^{(1)}}. \quad (3.35)$$

Similarly, for variable X,

$$\delta_{XX} = \frac{\Lambda_{XX}^{(2)}}{\Sigma_{XX}} \quad \text{and can be estimated by} \quad \hat{\delta}_{XX} = \frac{\hat{\Lambda}_{XX}^{(2)}}{S_{XX}^{(1)}}. \quad (3.36)$$

These estimates are method of moments estimates. Alternatively, Maximum Likelihood (ML) estimates can also be used if the random variable have a Normal

distribution. From initial empirical results there is not much difference between the results obtained from the multilevel modeling software MLWiN (Goldstein, 1998) that utilized ML and those obtained from the moments approach used by Tranmer and Steel (1998) as described above for areal unit data. In this thesis we focus on the moments approach for convenience.

A measure of the within-area homogeneity for a pair of variables is the *intra-area cross-correlation*. Similarity of the values of two different variables within areas can be measured using the intra-area cross-correlation. For the model described by (3.24) and (3.25), the intra-area cross correlation,  $\delta_{YX}$  is:

$$\delta_{YX} = \frac{\Lambda_{YX}^{(2)}}{\sqrt{\Sigma_{XX}\Sigma_{YY}}} \quad (3.37)$$

and can be estimated by

$$\hat{\delta}_{YX} = \frac{\hat{\Lambda}_{YX}^{(2)}}{\sqrt{S_{XX}^{(1)}S_{YY}^{(1)}}}. \quad (3.38)$$

There was not much difference when the moments approach and ML approach was used for the type of data considered in this thesis. The computation of the intra-area-cross correlation was done using the method by Tranmer and Steel(1998).

### 3.1.3 Pure Correlation

The term *pure correlation coefficient* is used to describe the correlation of two variables where the effect of the other level is removed and thus reflect effects at a pertinent level (Tranmer and Steel, 2001). Based on the models given by (3.24) to (3.27), for levels  $l=1,2$ .

$$\rho_{YX}^{(l)} = \frac{\Lambda_{YX}^{(l)}}{\sqrt{\Lambda_{XX}^{(l)}\Lambda_{YY}^{(l)}}} \quad (3.39)$$

Estimates of the *pure correlation coefficient* are obtained by using  $\hat{\Lambda}_{YX}^{(l)}$ ,  $\hat{\Lambda}_{XX}^{(l)}$  and  $\hat{\Lambda}_{YY}^{(l)}$ . Thus,

$$\hat{\rho}_{YX}^{(l)} = \frac{\hat{\Lambda}_{YX}^{(l)}}{\sqrt{\hat{\Lambda}_{XX}^{(l)} \hat{\Lambda}_{YY}^{(l)}}}, \quad l = 1, 2. \quad (3.40)$$

### 3.1.4 Pure Regression

Similarly, a *pure regression coefficient* refers to the regression coefficient when the effect of the other level is removed. For level  $l=1,2$  the pure regression coefficient is defined as:

$$\beta_{YX}^{(l)} = \frac{\Lambda_{YX}^{(l)}}{\Lambda_{XX}^{(l)}} \quad (3.41)$$

The estimates are computed using  $\hat{\Lambda}_{YX}^{(l)}$  and  $\hat{\Lambda}_{XX}^{(l)}$ . Thus,

$$\hat{\beta}_{YX}^{(l)} = \frac{\hat{\Lambda}_{YX}^{(l)}}{\hat{\Lambda}_{XX}^{(l)}}, \quad l = 1, 2. \quad (3.42)$$

### 3.1.5 Moran's I

Spatial autocorrelation basically measures correlation of a single variable for all pairs of points at a particular distance or some other category. Spatial autocorrelation can be categorized as either positive or negative. A positive spatial autocorrelation implies that similar values appear close together and a negative autocorrelation has dissimilar values appearing close together. Standard global and some new local spatial statistics have been developed to detect spatial autocorrelation and spatial association. The global spatial autocorrelation measure most often used is the Moran's I coefficient.

The Moran's I is defined:

$$I_{YY}^{(1)} = \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}} \quad (3.43)$$

where

$$\frac{1}{N} \sum_i^N (Y_i - \bar{Y})^2 = \frac{N-1}{N} S_{YY}^{(1)}$$

$Y_i$  denotes the observed value at location  $i$

$\bar{Y}$  is the average of  $Y_i$  over  $N$  locations

$w_{ij}$  is the spatial weight measure.

The definition of  $w_{ij}$ , the spatial weight is an important issue. Different definitions of the spatial weight result in different values of the Moran's  $I$  aimed at detecting different types of spatial relationships and resulting in different conclusions. Weights can be based on contiguity; if say, location  $i$  is adjacent to location  $j$ , it is given a weight of 1, otherwise it is given a weight of 0. Weights can also be based on distance using distance between points or between centroids of polygons. Another way of defining the spatial weight is based on lagged contiguity. Thus, different choices of  $w_{ij}$  are possible, depending on what type of feature of the spatial relationships we are attempting to assess. There are software that can be used to calculate the weight matrix  $\mathbf{W}$  with elements  $w_{ij}$ . In this thesis the Moran's  $I$  was computed using SPLUS and S+Spatial and R.

### 3.1.6 Cross-Moran's $I$

Suppose we let,

$$Z_i = Y_i + X_i \tag{3.44}$$

$$I_{ZZ}^{(1)} = \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} (Y_i + X_i - \bar{Y} - \bar{X})(Y_j + X_j - \bar{Y} - \bar{X})}{\left(\frac{N-1}{N}\right) S_{ZZ}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}} \tag{3.45}$$

Note that

$$S_{ZZ}^{(1)} = S_{YY}^{(1)} + S_{XX}^{(1)} + 2S_{YX}^{(1)}$$

thus,

$$\begin{aligned}
I_{ZZ}^{(1)} &= \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} [(Y_i - \bar{Y}) + (X_i - \bar{X})][(Y_j - \bar{Y}) + (X_j - \bar{X})]}{\left(\frac{N-1}{N}\right) S_{ZZ}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}} \\
&= \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} [(Y_i - \bar{Y})(Y_j - \bar{Y}) + (X_i - \bar{X})(X_j - \bar{X}) + 2(Y_i - \bar{Y})(X_j - \bar{X})]}{\left(\frac{N-1}{N}\right) S_{ZZ}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}} \\
&= \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\frac{N-1}{N}\right) S_{ZZ}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}} + \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\left(\frac{N-1}{N}\right) S_{ZZ}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}} \\
&\quad + 2 \frac{\sum_i^N \sum_{i \neq j}^N w_{ij} (Y_i - \bar{Y})(X_j - \bar{X})}{\left(\frac{N-1}{N}\right) S_{ZZ}^{(1)} \sum_i^N \sum_{i \neq j}^N w_{ij}}
\end{aligned}$$

So we have,

$$S_{ZZ}^{(1)} I_{ZZ}^{(1)} = I_{YY}^{(1)} S_{YY}^{(1)} + I_{XX}^{(1)} S_{XX}^{(1)} + 2 I_{YX}^{(1)} \sqrt{S_{YY}^{(1)} S_{XX}^{(1)}}.$$

Solving for  $I_{YX}$ , which is Cross-Moran's I,

$$I_{YX}^{(1)} = \frac{\left(S_{YY}^{(1)} + S_{XX}^{(1)} + 2S_{YX}^{(1)}\right) I_{ZZ}^{(1)} - S_{YY}^{(1)} I_{YY}^{(1)} - S_{XX}^{(1)} I_{XX}^{(1)}}{2\sqrt{S_{YY}^{(1)} S_{XX}^{(1)}}} \quad (3.46)$$

Hence, we can calculate  $I_{YX}$  using any method that calculates Moran's I for a variable by creating  $Z_i$  and using (3.46).

### 3.1.7 Relationships Between Pure Coefficients and the Intra-area Correlation

The overall correlation,  $\rho_{YX}$  can be expressed in terms of the intra-area correlation and the *pure correlations*,

$$\begin{aligned}
\rho_{YX} &= \frac{\Sigma_{YX}}{\sqrt{\Sigma_{YY}\Sigma_{XX}}} \\
&= \frac{\Lambda_{YX}^{(1)} + \Lambda_{YX}^{(2)}}{\sqrt{\Sigma_{YY}\Sigma_{XX}}} \\
&= \rho_{YX}^{(1)} \sqrt{\frac{\Lambda_{YY}^{(1)}\Lambda_{XX}^{(1)}}{\Sigma_{YY}\Sigma_{XX}}} + \rho_{YX}^{(2)} \sqrt{\frac{\Lambda_{YY}^{(2)}\Lambda_{XX}^{(2)}}{\Sigma_{YY}\Sigma_{XX}}} \\
&= \rho_{YX}^{(1)} \sqrt{(1 - \delta_{YY})(1 - \delta_{XX})} + \rho_{YX}^{(2)} \sqrt{\delta_{YY}\delta_{XX}}
\end{aligned}$$

this shows that

$$\rho_{YX} = \rho_{YX}^{(1)} \sqrt{(1 - \delta_{YY})(1 - \delta_{XX})} + \rho_{YX}^{(2)} \sqrt{\delta_{YY}\delta_{XX}}. \quad (3.47)$$

Thus from (3.47) the level 2 *pure correlation* coefficient  $\rho_{YX}^{(2)}$  is:

$$\rho_{YX}^{(2)} = \frac{\rho_{YX} - \rho_{YX}^{(1)} \sqrt{(1 - \delta_{YY})(1 - \delta_{XX})}}{\sqrt{\delta_{YY}\delta_{XX}}}. \quad (3.48)$$

The level 1 *pure correlation* coefficients is:

$$\rho_{YX}^{(1)} = \frac{\rho_{YX} - \rho_{YX}^{(2)} \sqrt{\delta_{YY}\delta_{XX}}}{\sqrt{(1 - \delta_{YY})(1 - \delta_{XX})}} \quad (3.49)$$

Similarly, the regression coefficient  $\beta_{YX}$  can be expressed in terms of intra-area correlation and pure regression coefficients.

$$\begin{aligned}
\beta_{YX} &= \frac{\Sigma_{YX}}{\Sigma_{XX}} \\
&= \frac{\Lambda_{YX}^{(1)} + \Lambda_{YX}^{(2)}}{\Sigma_{XX}} \\
&= \frac{\beta_{YX}^{(1)} \Lambda_{XX}^{(1)}}{\Sigma_{XX}} + \frac{\beta_{YX}^{(2)} \Lambda_{XX}^{(2)}}{\Sigma_{XX}} \\
&= \frac{\beta_{YX}^{(1)} (\Sigma_{XX} - \delta_{XX} \Sigma_{XX})}{\Sigma_{XX}} + \frac{\beta_{YX}^{(2)} \delta_{XX} \Sigma_{XX}}{\Sigma_{XX}} \\
&= \beta_{YX}^{(1)} (1 - \delta_{XX}) + \beta_{YX}^{(2)} \delta_{XX}
\end{aligned}$$



Thus, we have

$$\beta_{YX} = \beta_{YX}^{(1)} (1 - \delta_{XX}) + \beta_{YX}^{(2)} \delta_{XX}. \quad (3.50)$$

From 3.50 the level 2 *pure regression* is

$$\beta_{YX}^{(2)} = \frac{\beta_{YX} - \beta_{YX}^{(1)} (1 - \delta_{XX})}{\delta_{XX}} \quad (3.51)$$

and the level 1 pure regression is

$$\beta_{YX}^{(1)} = \frac{\beta_{YX} - \beta_{YX}^{(2)} \delta_{XX}}{(1 - \delta_{XX})}. \quad (3.52)$$

The relationships described above will be used later to look into the behavior of the pure coefficients given the initial correlation or regression coefficients and the initial intra-area correlation (or the Moran's I at level 1).

### 3.1.8 The relationship between intra-area correlation and the Moran's I

Assume that  $N_g = \bar{N}$ , so  $\bar{N}^* = \bar{N}$ . Since  $N$  is large,  $N-1 \approx N$ . Also since  $M$  is large,  $M-1 \approx M$ . The definition of Moran's I is:

$$I_{YY}^{(1)} = \frac{\sum_i^N \sum_{j \neq i}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} \sum_i^N \sum_{j \neq i}^N w_{ij}} \quad (3.53)$$

Here  $w_{ij}$  is the spatial weight measure that is equal to 1 if  $i$  and  $j \in g$  and 0 otherwise.

Since we assume that  $N_g = \bar{N}$ , we have,

$$\sum_i^N \sum_{j \neq i}^N w_{ij} = \sum_g^M N_g (N_g - 1) = M \bar{N} (\bar{N} - 1) \quad (3.54)$$

$$\begin{aligned}
I_{YY}^{(1)} &= \frac{\sum_i^N \sum_{j \neq i}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{\sum_g^M \sum_{i \in g} \left[ \sum_{j \neq i \in g} (Y_i - \bar{Y})(Y_j - \bar{Y}) \right]}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{\sum_g^M \sum_{i \in g} \left[ \sum_{j \in g} (Y_i - \bar{Y})(Y_j - \bar{Y}) - (Y_i - \bar{Y})^2 \right]}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{\sum_g^M \sum_{i \in g} [(Y_i - \bar{Y}) N_g (\bar{Y}_g - \bar{Y})] - \sum_g^M \sum_{i \in g} (Y_i - \bar{Y})^2}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{\sum_g^M N_g^2 (\bar{Y}_g - \bar{Y})^2 - \sum_g^M \sum_{i \in g} (Y_i - \bar{Y})^2}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{\bar{N} \sum_g^M N_g (\bar{Y}_g - \bar{Y})^2 - \sum_g^M \sum_{i \in g} (Y_i - \bar{Y})^2}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{\bar{N} (M - 1) S_{YY}^{(2)} - (N - 1) S_{YY}^{(1)}}{\left(\frac{N-1}{N}\right) S_{YY}^{(1)} M \bar{N} (\bar{N} - 1)} \\
&= \frac{(N - 1) S_{YY}^{(2)} - (N - 1) S_{YY}^{(1)}}{(N - 1) (\bar{N} - 1) S_{YY}^{(1)}} \\
&= \frac{S_{YY}^{(2)} - S_{YY}^{(1)}}{(\bar{N} - 1) S_{YY}^{(1)}}.
\end{aligned}$$

Thus, we have,

$$\begin{aligned}
I_{YY}^{(1)} &= \frac{(\bar{N}^* - 1) \hat{\Lambda}_{YY}^{(2)}}{(\bar{N} - 1) S_{YY}^{(1)}} \\
&= \frac{\hat{\Lambda}_{YY}^{(2)}}{S_{YY}^{(1)}} \\
&= \hat{\delta}_{YY}
\end{aligned}$$

This shows that we can regard the intraclass correlation as a measure of the average spatial correlation within groups. A similar relationship hold for the intra-area cross correlations and the cross-Moran's I.

Equation (3.30) is the key to explaining the MAUP under a simple multilevel model. Comparing (3.30) with (3.29) we see that in aggregating the data and calculating a weighted variance the contribution of the level 1 variance component is unchanged, whereas the contribution of the level 2, an area level, variance component from approximately 1 to  $\bar{N}^*$ . Similar results hold for  $S_{XX}^{(2)}$  and  $S_{YX}^{(2)}$  leading to the scale effect. Examining (3.30), we see that even if the spatial correlation produces quite small intra-area correlation, the presence of  $\bar{N}^*$ , which is effectively the average number of people per areal units, implies that aggregation may have substantial effects on the variances and covariances and coefficients calculated from them. ‘

## Chapter 4

# Multilevel Modeling and the MAUP

This chapter describes the first of a series of experiments and results on the scale effects of relevant statistics derived from a simple multilevel model, as well as the standard statistics. Different degrees of spatial autocorrelations are considered in the experiments.

### 4.1 Is Multilevel modeling a possible solution to the MAUP?

Multilevel models offer an approach to a number of issues, including the MAUP. It can provide estimates of both the average effects of a variable over a number of settings, and the extent to which that effect varies over settings (Jones and Duncan, 1996). Multilevel modeling allows for effects at different levels.

Multilevel Modeling has been suggested for use with hierarchical data. An example of the consequence if a hierarchy is ignored in analysis is given by Aitkin et al

(1981) who reanalyzed the study of Bennet (1976) on primary school children. Bennet (1976) claimed that formal styles of teaching reading produced greater progress among pupils than any other methods. In this study, the grouping of pupils within teachers and classes were ignored. Aitkin et al. (1981) took these groupings into account and the statistically significant difference between teachers' styles disappeared and they concluded that the formally taught pupils could not be shown to differ from others (Gleave et al, 2000). When applied to areal data, multilevel modeling is still potentially subject to the MAUP, since different estimates of the variance components can be obtained if boundaries are changed or a different scale is used. A possible reason for multilevel models still being subject to the MAUP, is that while the data available may be hierarchical, the population correlation structure may be more complex. In particular the spatial pattern of correlations between units may be more complex than that implied by a standard multilevel model. Multilevel modeling provides an approach to analysing spatially aggregated data but itself may be affected by the MAUP. We will examine how the results of multilevel modeling are affected by the MAUP and whether it can produce results that are less affected than standard analysis methods.

To evaluate the potential effectiveness of multilevel modeling as a possible solution or approach to the MAUP several experiments were conducted. As an initial investigation of this possibility, the scale effects of some statistics that can be derived from multilevel models were computed. These include the intra-area correlation, intra-area cross correlation, pure correlation and pure regression coefficients.

The following experiments also include single level analyses of the data sets at different levels of aggregation.

It is well known that the results of individual level analyses are different from those conducted using group level data. The usual result for correlation coefficients is that they increase as the level of aggregation increases. If the individuals are grouped together in a non-random way, the correlation coefficient at the individual level is usually less than the correlation at group level. The population mean using the appropriate population weights is not affected by aggregation. The sample

variance in general is affected by aggregation (Holt, Steel and Tranmer, 1996).

To examine aggregation effects say between individual level and group level analysis, Holt, Steel and Tranmer (1996) proposed a sample variance components model wherein the variance can be partitioned into the area and individual level covariance matrices. Recall the model in Chapter 3. Consider the model for a single variable of interest  $Y$ :

$$Y_i = \mu_Y + \alpha_{Y_g} + \epsilon_{Y_i} \quad (4.1)$$

where

$\mu_Y$  is the expectation of  $Y$  across the region of interest

$\alpha_{Y_g}$  is a random variable representing the area effect for the  $g$ th area

$\epsilon_{Y_i}$  is a random variable representing the pure individual effect.

**Assumptions:**

$$\begin{aligned} E(\alpha_{Y_g}) &= 0, E(\epsilon_{Y_i}) = 0, \text{ and } \text{var}(\alpha_{Y_g}) = \Lambda_{YY}^{(2)}, \text{var}(\epsilon_{Y_i}) = \Lambda_{YY}^{(1)}, \\ \text{cov}(\alpha_{Y_g}, \epsilon_{Y_i}) &= 0, \text{cov}(\epsilon_{Y_i}, \epsilon_{Y_{i'}}) = 0, \text{ for } i \neq i'. \end{aligned}$$

**Properties:**

$$\begin{aligned} E(Y_i) &= \mu_Y \\ \text{var}(Y_i) &= \Lambda_{YY}^{(1)} + \Lambda_{YY}^{(2)} \\ \text{cov}(Y_i, Y_j) &= \Lambda_{YY}^{(2)} \text{ if } i \in g, j \in g \\ &= 0 \text{ otherwise.} \end{aligned}$$

The following section will investigate the scale effects of some standard statistics and some statistics derived from the multilevel model described by (4.1) when there are different degrees of spatial autocorrelation present. The aim is to see if statistics and analyses associated with a simple multilevel model are less affected than the standard statistics and if the effects are more predictable as scale changes. Three experiments are conducted. In experiment 1, each variable has the same level of autocorrelation, which is set to high, medium and low. In experiment 2 neither variables is autocorrelated. Experiment 3 considers the case when the level of autocorrelation is different for the two variables. A summary of results is provided in section 4.6.

## 4.2 Experiment 1: Scale effects of some statistics from simulated data

To initially investigate the MAUP effects on analysis based on the multilevel model described above, various sets of data are generated.

The first three data sets are generated in such a way that both variables, Y and X are autocorrelated but with different degree of autocorrelation. The degree of autocorrelation will be categorized as ‘*low*’, ‘*medium*’, and ‘*high*’. In section 4.3 we look at the case when both variables have no autocorrelation. In section 4.5 we look into the effects when two variables have different degrees of autocorrelations then two more data sets were generated, one variable ‘*low*’ autocorrelated and one ‘*high*’ autocorrelated. Table 4.1 shows ranges of the categories used in the experiments. The measure of autocorrelation used in this study is the Moran’s I described in equation (3.42). The connectivity matrix used in determining the Moran’s I is the queen’s case. In a square grid, queen contiguity implies that adjacent cell with common borders and common vertex are considered neighbors. The queen’s case is used as we start at the individual level and this case creates local neighborhoods of individuals at a similar distance apart. It also correspond to a rational way to form larger scale areal units from individuals on smaller scale areal units with both simulated and real data. The computation of the Moran’s I was done using GeoDa (Anselin, 1996) and Rookcase (Sawada, M., 1999).

	Moran’s I
Low	0.1 - 0.3
Medium	0.4 - 0.63
High	0.7 - 0.83

**Table 4.1: Range of values for the categories**

Data Set 1: Y is ‘low’ autocorrelated and X is ‘low’ autocorrelated

Data Set 2: Y is ‘medium’ autocorrelated and X is ‘medium’ autocorrelated

Data Set 3: Y is ‘high’ autocorrelated and X is ‘high’ autocorrelated

### 4.2.1 Data Set 1: Both variables have low autocorrelation

The data generation process is similar to that used by Green and Flowerdew (1996) to generate data with a known pattern of spatial autocorrelation.

Values for two variables Y and X are assigned to each cell of a 100x100 square grid. Two cells are neighbors if they have one common side or common vertex. Initially, a set of normally distributed random numbers with mean 0 and variance 16, denoted by  $\sim N(0,16)$  are generated and are assigned to the spatial locations. These values are transformed into autocorrelated data by taking the average of the neighboring values for each data points, an error is then introduced that is independent and identically distributed (iid)  $\sim N(0,4)$ . The results are the values of variable X.

The values of variable Y are then generated using a similar procedure. Data are generated using  $\{ 10 + (\text{original set of random numbers}) + \text{error} \}$ , the error is iid  $\sim N(0,4)$ . The results are then transformed into autocorrelated data by taking the average of the neighboring values for each data point. The results are the values of the variable Y.

To summarize the data generation,

1. Generate  $A \sim \text{iid } N(0,16)$ .
2. Let  $A^*$  be the average of the neighbors of A as described above.
3. Let  $B = 10 + A + e$ , where  $e \sim \text{iid } N(0,16)$ .
4. Let  $B^*$  be the average of the neighbors of B as described above.
5. Variables X and Y has values,  $X = A^* + e_1$  and  $Y = B^* + e_2$ ,  
where  $e_1 \sim \text{iid } N(0,4)$  and  $e_2 \sim \text{iid } N(0,4)$  and  $e_1$  and  $e_2$  are independent.
6. The mean and variance of X and Y are then changed to desired values.

To change the mean of a variable, add  $(m_2 - m_1)$  for each observation where  $m_1$  is current mean and  $m_2$  is the desired mean. The variance of a variable can be changed by multiplying each observation by  $(\delta_2/\delta_1)$  where  $\delta_2$  is the desired standard deviation and  $\delta_1$  is the current standard deviation.

The means for variables X and Y were set at 0.005 and 10, respectively. The desired variances were set at 6 and 8 for X and Y respectively.



We have now a set of data (Y,X) with corresponding locations and a certain level of spatial autocorrelation.

The data sets are then aggregated spatially by contiguous blocks of mxm cells being grouped together, where  $m = 2, 5, 10, 20, 25, 50$ . Thus, the number of zones are 2500, 625, 400, 100, 25, 4, respectively. This means that when  $m=2$  the 100x100 grid is divided into 2500 zones each containing 4 of the original cells. When  $m=5$ , the 100x100 grid is divided into 625 zones each containing 25 of the original units, and so on.

#### Analysis of one realization:

Initially one realization of the data set generation will be used to examine pertinent statistics. Table 4.2 shows the Moran's I and the cross-Moran at individual level and different levels of aggregation of one realization. The connectivity matrix used in determining the Moran's I is the queen's case. In a square grid, queen contiguity implies that units with common borders and common vertex are considered neighbors. The computation of the Moran's I was done using GeoDa (Anselin, 1996) and Rookcase (Sawada, M., 1999). Because of the way the data are generated, variable Y has higher autocorrelation than variable X as shown in Table 4.2. The cross-Moran is computed using Equation (3.44) in Chapter 3 sub-section (3.1.6).

Level	$I_{XX}^l$	$I_{YY}^l$	$I_{YX}^l$
Individual	0.1222	0.2051	0.1102
Z2500	0.1608	0.2292	0.1348
Z625	0.0493	0.0675	0.0459
Z400	0.0171	0.0455	0.0342
Z100	0.0054	0.0612	0.0657
Z25	-0.0125	0.0698	-0.0432
Z4	-	-	-

Table 4.2: Moran's I

The unweighted variances and the covariances decrease with scale. The weighted variances and covariances increased with scale. It can be noted that for the case of equal group population sizes the unweighted variance can be obtained using

$$Variance(unweighted) = \frac{Variance(weighted)}{n_{zone}}$$

where  $n_{zone}$  is the number of elements in each zone.

Table 4.3 shows the unweighted correlation and regression coefficients. Note that the weighted and unweighted coefficient are the same because of the equal cell sizes used. The increase of the coefficients with scale can be attributed to the aggregation effects of the variances and the covariance of the two variables.

	Correlation Coefficient	Regression Coefficient
Individual Data	0.2944	0.3399
Number of Zones		
2500	0.4280	0.5398
625	0.5157	0.6813
400	0.5558	0.7586
100	0.6051	0.9084
25	0.7768	1.1115
4	0.7437	1.4561

**Table 4.3: Correlation and regression coefficients at different scales, X and Y both have low autocorrelation**

The aggregation effects of the weighted variance of Y ( $S_{YY}^{(l)}/S_{YY}^{(1)}$ ) is greater than the aggregation effect of the weighted variance of X ( $S_{XX}^{(l)}/S_{XX}^{(1)}$ ) because of the way the data are generated. Variable Y will have greater autocorrelation than variable X. The unweighted covariance decreases as the number of zones decreases. Reynolds (1998) suggested that the unweighted covariance tend to decrease when the data are aggregated because the change in spatial arrangements of the two variables is more likely make the association random than it is to make it more related. The aggregation effect of the weighted covariance ( $S_{YX}^{(l)}/S_{YX}^{(1)}$ ) is greater than the aggregation effects of variables X and Y in all levels of aggregation except the last one. Because of this, the correlations as the data are aggregated are as shown in Table 4.3. The correlation increases with scale except for the last correlation.

Correlation and regression analysis was conducted using ordinary least squares (OLS) at each scale. From this point we call statistics calculated from the data directly at any scale as *direct* coefficients (or statistics). Thus, the results in Table 4.3 are direct coefficients. It can be observed that both the correlation and regression coefficients display scale effects. They tend to increase with scale, that is, the

estimated coefficient increases as the number of zones decreases and therefore the number of observations increases in each zone.

To look at some statistics that can be derived from the multilevel model described in section (3.1.2), further statistics were computed. One purpose of using the multilevel model is to use the components of the model for further computations of some useful statistics. Table 4.4 shows the estimated intra-area correlation and the variance components using the moments approach (Tranmer and Steel(1998) method) and using MLWiN, respectively for the variable X. The estimates of level 2 variance components derived from MLWiN are larger than the corresponding estimates using the moments estimation approach, resulting in larger estimates of the intra-area correlation. However, the descending trend as the number of groups decrease are similar. In this study the moments approach is used because it can easily be used when individual unit level data with group indicators are not available, provided data for group means and a unit level sample without group indicators are available (see Tranmer and Steel, 1998).

A. Moments				
Level 1	Level 2	$\hat{\Lambda}_{XX}^{(2)}$	$\hat{\Lambda}_{XX}^{(1)}$	$\hat{\delta}_{XX}$
Individual	Z2500	0.8476	5.1524	0.1413
	Z625	0.5170	5.4830	0.0862
	Z400	0.3871	5.6128	0.0654
	Z100	0.0906	5.9094	0.0151
	Z25	0.0300	5.9700	0.0050
	Z4	0.0034	5.9966	0.0006
B. MLWiN				
Level 1	Level 2	$\hat{\Lambda}_{XX}^{(2)}$	$\hat{\Lambda}_{XX}^{(1)}$	$\hat{\delta}_{XX}$
Individual	Z2500	0.8787	5.1213	0.1465
	Z625	0.7174	5.2826	0.1196
	Z400	0.4228	5.5772	0.0705
	Z100	0.1387	5.8613	0.0231
	Z25	0.0487	5.9513	0.0081
	Z4	0.0046	5.9954	0.0007

**Table 4.4: Intra-Area correlations and variance components of X, X have low auto-correlation**

Table 4.5 shows the intra-area (IAC) correlation of the Y variable and the estimated level 1 and level 2 variance components. The results presented are obtained

using the Tranmer and Steel method(1998). For both variables the intra-area correlations decreases with scale. The level 2 variance components decrease with scale and approach zero. The level 1 variance components increase as the number of zones decrease and approaches the individual level variance. Generally, as groups become larger, more dissimilar units are included leading to the average within-area homogeneity, which is what  $\hat{\delta}$  measures, to decrease.

Level 1	Level 2	$\hat{\Lambda}_{YY}^{(2)}$	$\hat{\Lambda}_{YY}^{(1)}$	$\hat{\delta}_{YY}$
Individual	Z2500	1.8632	6.1337	0.2328
	Z625	1.0670	6.9330	0.1334
	Z400	0.8533	7.1467	0.1067
	Z100	0.2593	7.7407	0.0324
	Z25	0.0718	7.9282	0.0090
	Z4	0.0180	7.9820	0.0023

**Table 4.5: Intra-Area correlations and variance components of Y using moments, Y have low autocorrelation**

The estimated intra-area cross-correlation (IACC) of the two variables, denoted by  $\hat{\delta}_{YX}$  is shown in Table 4.6. The estimated level 2 covariance components decrease with scale, the level 1 covariance components increase and intra-area cross correlation decrease with scale.

Level 1	Level 2	$\hat{\Lambda}_{YX}^{(2)}$	$\hat{\Lambda}_{YX}^{(1)}$	$\hat{\delta}_{YX}$
Individual	Z2500	0.8570	1.1824	0.1237
	Z625	0.4885	1.5509	0.0705
	Z400	0.3981	1.6413	0.0575
	Z100	0.1164	1.9231	0.0168
	Z25	0.0445	1.9949	0.0064
	Z4	0.0072	2.0323	0.0010

**Table 4.6: Intra-Area cross-correlations and covariance components at two levels, X and Y both have low autocorrelation**

Table 4.7 shows the estimated pure correlation and regression coefficients based on a simple multilevel model. Pure correlation, as defined in Chapter 3 equation (3.38), is the correlation of two variables where the effect of the other level is removed reflecting the correlation at the pertinent level. The level 1 pure correlation

(denoted by  $\hat{\rho}_{YX}^{(1)}$ ) increases with scale, starting with a value lower than the correlation coefficient (0.2944) at the individual level but approaches that value as the number of zones decreases. The change of the pure correlation as the scale changes is slow compared with the direct correlation shown in Table 4.3. This is a sign of more stability or being less affected by the MAUP. Estimated pure regressions at two levels and different scales are also shown in Table 4.7. Pure regression coefficients are the regression coefficient when the effect of the other level is removed. Pure regression at level 1 (denoted by  $\hat{b}_{YX}^{(1)}$ ) approaches the regression coefficient at the individual level (0.3399) as the number of zones decreases. Level 2 pure coefficients, while different from the individual level, show less scale effect than the direct coefficients and a general tendency to increase with scale, except when  $m=4$  where they may be affected by the small number of groups.

Level 1	Level 2	$\hat{\rho}_{YX}^{(2)}$	$\hat{\rho}_{YX}^{(1)}$	$\hat{b}_{YX}^{(2)}$	$\hat{b}_{YX}^{(1)}$
Individual	Z2500	0.6821	0.2103	0.4602	0.1926
	Z625	0.6578	0.2515	0.4578	0.2237
	Z400	0.6926	0.2592	0.4665	0.2297
	Z100	0.7594	0.2843	0.4487	0.2484
	Z25	0.9290	0.2900	0.6202	0.2516
	Z4	0.9088	0.2937	0.3976	0.2546

**Table 4.7: Pure correlations and regressions, X and Y both have low autocorrelation**

### Analysis of Distribution of Statistics

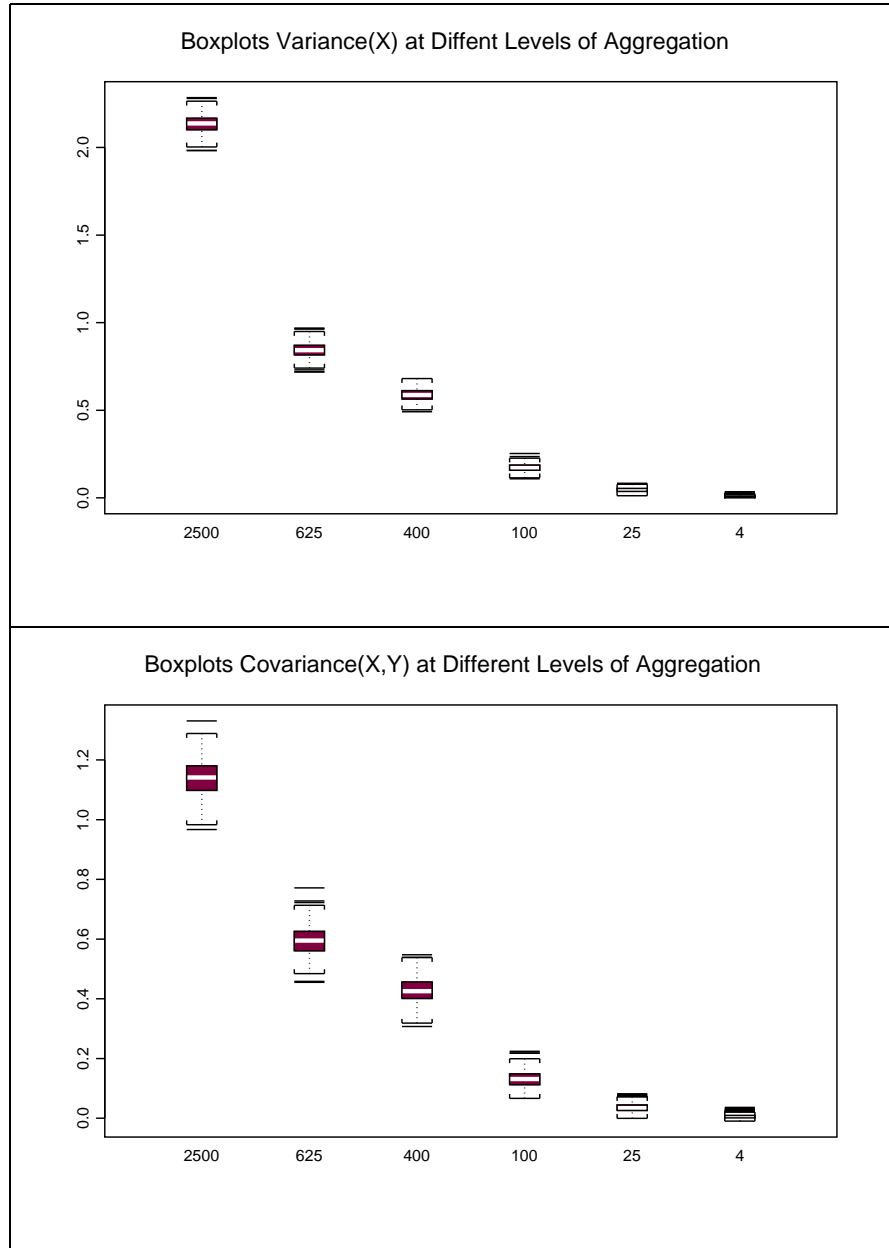
The results above arise from just one realization of the data generation. To examine the distribution of the statistics derived from the MLM, the data generation is repeated 500 times. In each repetition, the mean of X and mean of Y is scaled to 0.005 and 10, respectively and the variance of X and Y are also scaled to 6 and 8, respectively. This makes the different realizations directly comparable. Although we cannot completely control or change the initial Pearson correlation, the standard deviation of the initial individual level correlation is not large, the values ranging from 0.2599 to 0.3247 with mean equal to 0.2929.

Figure 4.1 shows the ranges of the values of the unweighted variance at dif-

ferent levels of aggregation when the data generation is repeated 500 times. The unweighted variance and covariance are observed to decrease as the number of zones decreases. Note that the horizontal axis shows the number of zones in the square grid, 2500 means that the original 100x100 square grid was made into a 50x50 square grid giving 2500 square zones; 625 means that the original 100x100 was made into a 25x25 square grid with 625 square zones; and so on. Recall that the initial variance for X is 6. For the rest of the thesis, this notation applies. The effect on the variance of Y is similar to that of the variable X. Thus, given data set 1, we can say that when a variable has low autocorrelation, the mean of the unweighted variance decreases with the decrease in the number of zones.

The standard deviation of the variance decreases with the decrease of the number of zones as shown in the last column of the upper portion of Table 4.8. At this point the claim of Reynolds that *"When significantly positive autocorrelated variables are aggregated, increasing the number of regions per cell increases the likelihood that more widely differing values will be included in each cell, so one would expect the variability of possible aggregate variance values to increase with the decrease in the number of cells."* (page 23, Reynolds, 1998) is observed in Figure 4.1. The standard deviation of the variance increases as the number of individuals included in the group decreases. In other words, the standard deviation decreases as the number of groups decreases.

Figure 4.2 shows the distribution of the weighted variance of X and weighted covariance of X and Y. The boxplots for the variance of variable Y is not shown because it is similar to the boxplots of the variance of X. The scale effect of the weighted covariance is similar to that of the effect of the variance. Recall that all zones have equal number of units included in each group so that the (weighted variance) = (number of units per group) × (unweighted variance). Thus, the weighted variance includes the factors of the number of units per group  $n_{units}$  and unweighted variance. The variability of the unweighted variance decreases with aggregation but in Figure 4.2 the standard deviation increases with aggregation. This is because of the factor  $n_{units}$ . Recall that the different levels of aggregations have 2500, 625,



**Figure 4.1: Unweighted Variance of X and Covariance(X,Y), X and Y both have low autocorrelation**

400, 100, 25, and 4 groups, which means that the number of units per group for the different levels of aggregation are 4, 16, 25, 100, 400, and 2500 respectively. Thus, the standard deviation of the weighted variance is expected to rise because of the increasing values of  $n_{units}$ .

Table 4.9 summarizes the distribution of the weighted variance of X and the

Unweighted Variance of X	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	6.0000	6.0000	6.0000	6.0000	0.0000
2500	2.1682	2.1622	1.8998	2.5105	0.1007
625	0.8565	0.8539	0.7220	1.0474	0.0580
400	0.5964	0.5949	0.4750	0.7335	0.0451
100	0.1757	0.1744	0.1103	0.2732	0.0237
25	0.0464	0.0454	0.0129	0.0838	0.0130
4	0.0076	0.0063	0.0001	0.0361	0.0056
Unweighted Covariance of X and Y	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	2.0292	2.0298	1.8007	2.2493	0.0738
2500	1.1644	1.1671	0.9712	1.4121	0.0800
625	0.6070	0.6052	0.4452	0.8060	0.0556
400	0.4368	0.4311	0.3083	0.5921	0.0476
100	0.1351	0.1329	0.0663	0.2369	0.0274
25	0.0369	0.0352	-0.0001	0.0862	0.0151
4	0.0061	0.0049	-0.0010	0.0380	0.0066

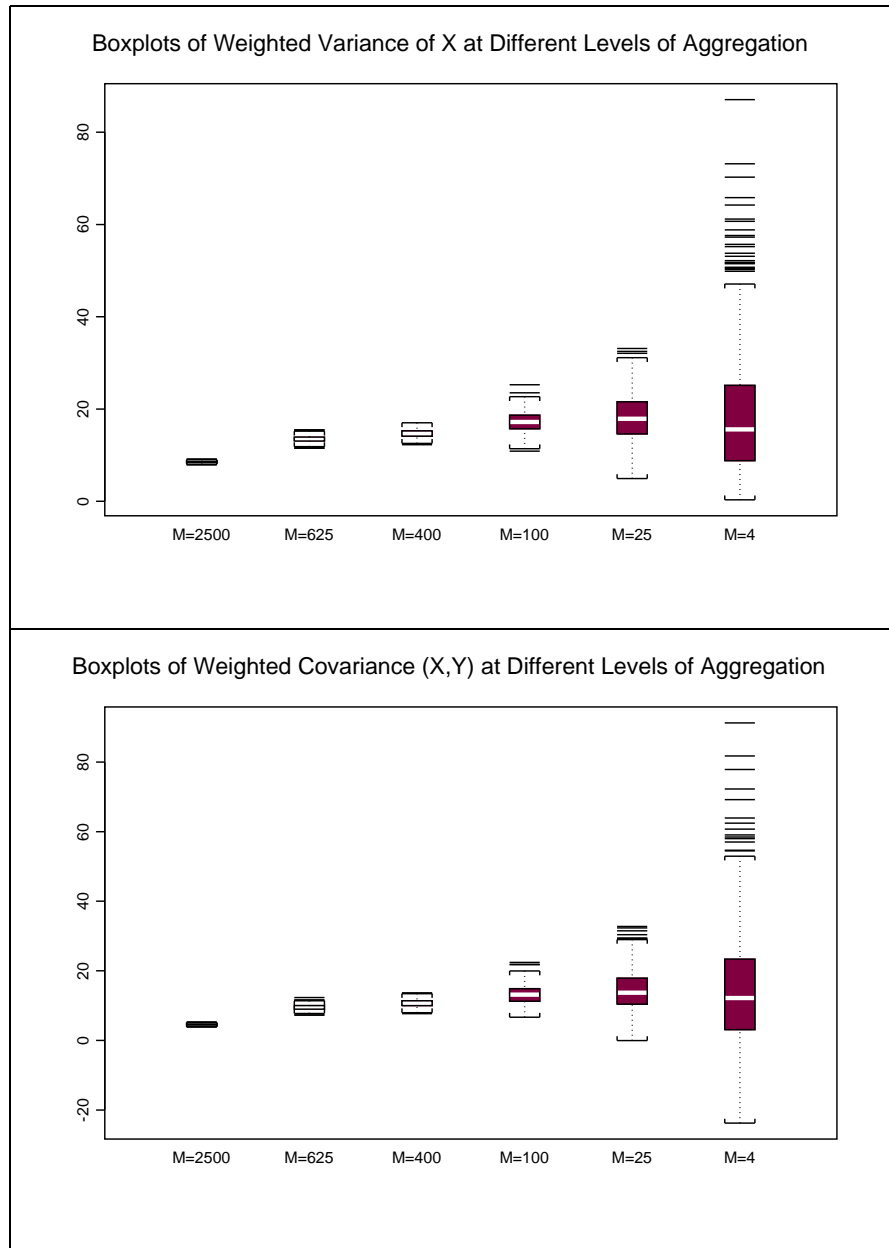
**Table 4.8: Description of Unweighted Variance of X and Covariance of X and Y, X and Y both have low autocorrelation**

weighted covariance of X and Y and shows the mean and the median increases with aggregation. In both the weighted variance and weighted covariance, the standard deviation increase with aggregation.

Weighted Variance of X	Mean	Median	Minimum	Maximum	Std. Dev.
2500	8.5416	8.5474	7.9298	9.1371	0.2004
625	13.4936	13.4804	11.4946	15.5029	0.6857
400	14.6797	14.6807	12.2834	17.0218	0.8635
100	17.3023	17.2079	10.8643	25.2772	2.2145
25	18.2766	17.9044	4.9363	33.1333	5.0246
4	18.6786	15.5848	0.3227	87.0727	13.6853
Weighted Covariance of X and Y	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	2.0292	2.0298	1.8007	2.2493	0.0738
2500	4.5654	4.5653	3.8670	5.3221	0.2380
625	9.5185	9.5159	7.2845	12.3398	0.7401
400	10.7004	10.6384	7.6858	13.6834	1.0274
100	13.2360	13.1340	6.6786	22.4103	2.6029
25	14.4596	13.6990	-0.0417	32.7256	5.8346
4	15.0037	12.1987	-23.7587	91.2361	15.9994

**Table 4.9: Description of Weighted Variance of X and Covariance of X and Y, X and Y both have low autocorrelation**

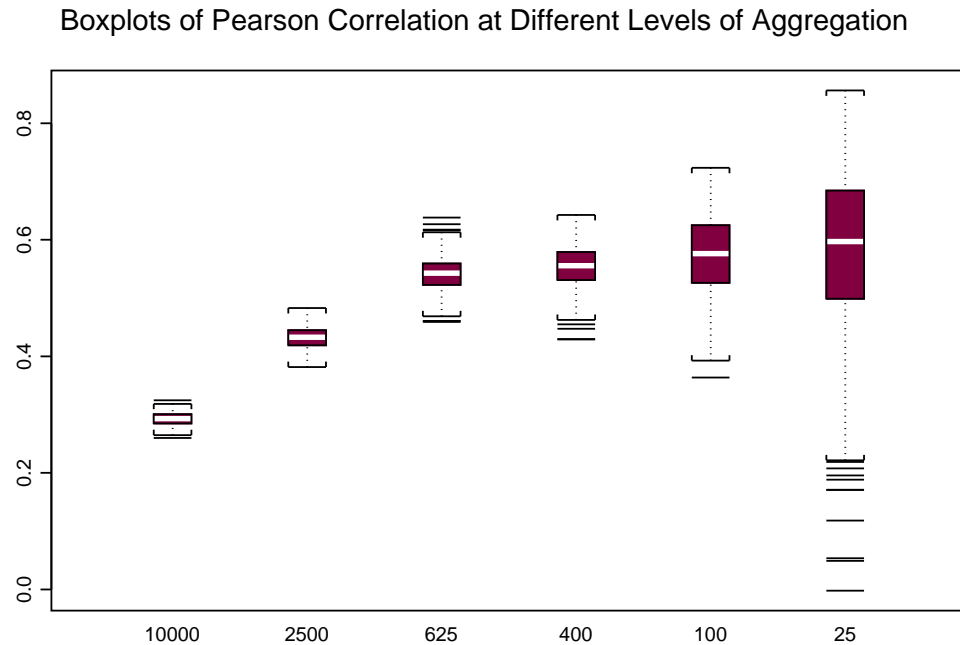




**Figure 4.2: Weighted Variance of X and Covariance(X,Y), X and Y both have low autocorrelation**

The effect of aggregation of the variance and covariance influence the effect on the correlation. Figure 4.3 shows the distribution of the correlation coefficient when the data are aggregated into smaller number of zones. The figures shows that the mean and median of the correlation coefficient increases with the level of aggregation, which is supported by Table 4.10, although the increase is small once we reach 625

zones. It can be noted also that the standard deviation of the values of the Pearson correlation increases with aggregation, due to the reduction in the number of groups in the analysis. The range of values when the data are aggregated to 4 zones almost have the range of possible values of the Pearson correlation, that is, -1 to +1.



**Figure 4.3: Pearson Correlation, X and Y both have low autocorrelation**

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.2929	0.2930	0.2599	0.3247	0.0107
2500	0.4322	0.4328	0.3815	0.4830	0.0185
625	0.5413	0.5427	0.4590	0.6382	0.0288
400	0.5544	0.5555	0.4290	0.6428	0.0343
100	0.5746	0.5761	0.3637	0.7237	0.0682
25	0.5790	0.5972	-0.0022	0.8563	0.1426
4	0.5279	0.7056	-0.9882	0.9998	0.4756

**Table 4.10: Description of the Pearson Correlation, X and Y both have low autocorrelation**

To examine this behavior theoretically, it is known that when the distribution is bivariate normal, then the sample correlation calculated from  $N$  independent

observations has the following properties (Steel and Holt, 1996):

$$E(r_{YX}) = \rho \left( 1 + \frac{1 - \rho^2}{2N} \right) + O(N^{-2}) \quad (4.2)$$

and

$$V(r_{YX}) = \frac{(1 - \rho^2)^2}{N - 1} \left( 1 + \frac{11\rho^2}{2N} \right) + O(N^{-3}). \quad (4.3)$$

Steel and Holt (1996) show that a weighted aggregated correlation calculated from  $M$  randomly formed groups behaves the same as that calculated from  $M$  points. Hence (4.2) and (4.3) apply with  $N$  replaced by  $M$ . Equations 4.2 and 4.3 are used to look at the expected behavior of the correlation coefficient. For different levels of aggregation, Equations 4.2 and 4.3 are used to estimate the expected values of the correlations and variances by using  $M$  instead of  $N$ , where  $M$  is the corresponding number of groups and the results are shown in Table 4.11. Notice that even for randomly formed groups the expected value of the correlation increases when the number of groups is quite small. For the individual level and 2500 zones the standard deviation of the correlation coefficient is approximately equal to the theoretical value. For smaller number of zones the standard deviation is smaller than the theoretical values, although it decreases at a similar rate. However, the average of the correlation does not behave as in the random aggregation, because of the autocorrelation. Theoretically, the expected value of the correlations seems to be approximately constant at different levels of aggregations with some increase as  $M$  becomes small. However, this behavior is not observed in this experiment.

	$E(r_{YX})$	$V(r_{YX})$	$\sqrt{V(r_{YX})}$
Individual	0.3000	0.00008	0.0091
Z2500	0.3000	0.00033	0.0182
Z625	0.3018	0.00133	0.0364
Z400	0.3028	0.00208	0.0456
Z100	0.3113	0.00838	0.0915
Z25	0.3454	0.03477	0.1865
Z4	0.5841	0.30191	0.5492

**Table 4.11: Expected correlation and the variance and SD at different levels of aggregation assuming no autocorrelation**

The percentage loss of variance (plv) of the variance of variables  $X$  and  $Y$  and

so with the percentage loss of covariance (plc) of the covariance at different levels of aggregation has something to do with this behaviour.

First let us define percentage loss of variance (plv). For a variable X, the plv of X is

$$\text{plv}_{XX} = \frac{(S_{XX}^{(2)} - S_{XX}^{(1)})}{S_{XX}^{(1)}} \quad (4.4)$$

where  $S_{XX}^{(1)}$  is the variance at individual level and  $S_{XX}^{(2)}$  is the weighted aggregated variance.

Similarly, for variable Y, the plv is

$$\text{plv}_{YY} = \frac{(S_{YY}^{(2)} - S_{YY}^{(1)})}{S_{YY}^{(1)}}. \quad (4.5)$$

The percentage loss of covariance (plc) is

$$\text{plc}_{YX} = \frac{(S_{YX}^{(2)} - S_{YX}^{(1)})}{S_{YX}^{(1)}}. \quad (4.6)$$

where  $S_{YX}^{(1)}$  is the covariance at individual level and  $S_{YX}^{(2)}$  is the weighted aggregated covariance.

The percentage loss of variances of the variables X and Y with respect to the levels of aggregation are almost the same. Although the corresponding percentage loss of the covariance with respect to the levels of aggregation have similar trend, they are smaller compared with the percentage loss of variances of both X and Y. This results in the tendency for the correlation to increase with aggregation.

Figure 4.4 shows the distribution of the estimated regression coefficient. The standard deviation increases with aggregation. The distribution when the data are aggregated into 4 zones is not included to allow a clearer picture of the distribution for the other levels of aggregation. However, the necessary information for this level of aggregation can be seen in Table 4.12. A trend similar to that observed for the distribution of the correlation coefficient in terms of the increasing mean and increasing standard deviation of the estimates is observed.

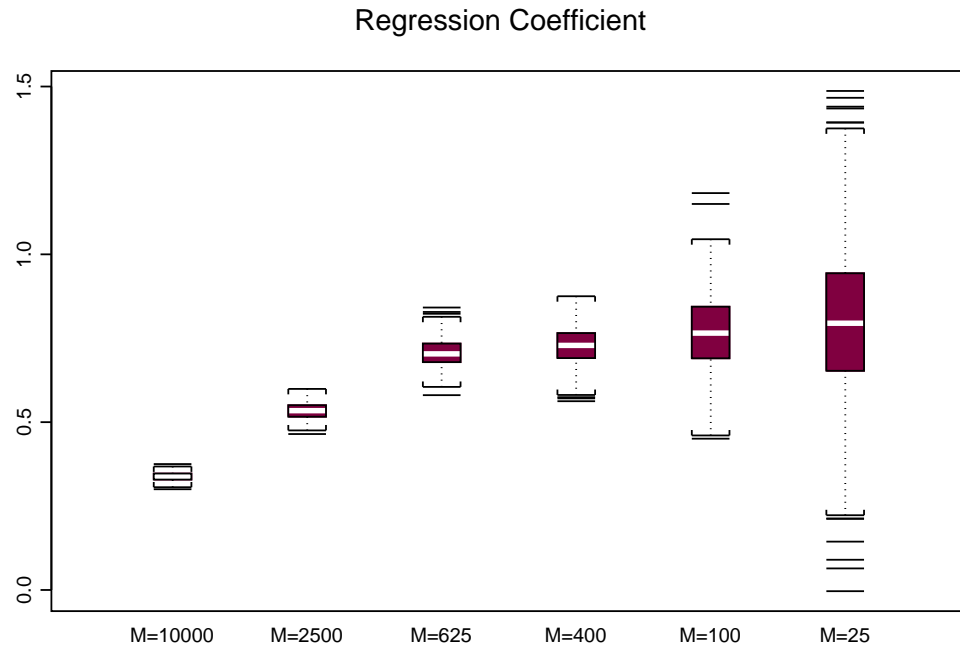


Figure 4.4: Regression Coefficient, X and Y both have low autocorrelation

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.3382	0.3383	0.3001	0.3749	0.0123
2500	0.5344	0.5343	0.4644	0.5989	0.0239
625	0.7055	0.7037	0.5803	0.8418	0.0422
400	0.7288	0.7290	0.5627	0.8750	0.0535
100	0.7654	0.7653	0.4507	1.1826	0.1148
25	0.7904	0.7950	-0.0037	1.4869	0.2354
4	0.7750	0.8681	-6.8700	5.2190	1.0292

Table 4.12: Description of Regression Coefficient, X and Y both have low autocorrelation

### Statistics associated with the multilevel model

We now examine some of the statistics associated with the multilevel model presented above (4.1). Some of the derived statistics display interesting patterns, which may be affected by the level 1 and level 2 variance components. Figure 4.5 shows the estimated level 2 and level 1 variance components of variable X.

Note that the level 1 plus level 2 variance components equal the initial variance of X. Looking at the figure, the mean and median of the level 2 variance components

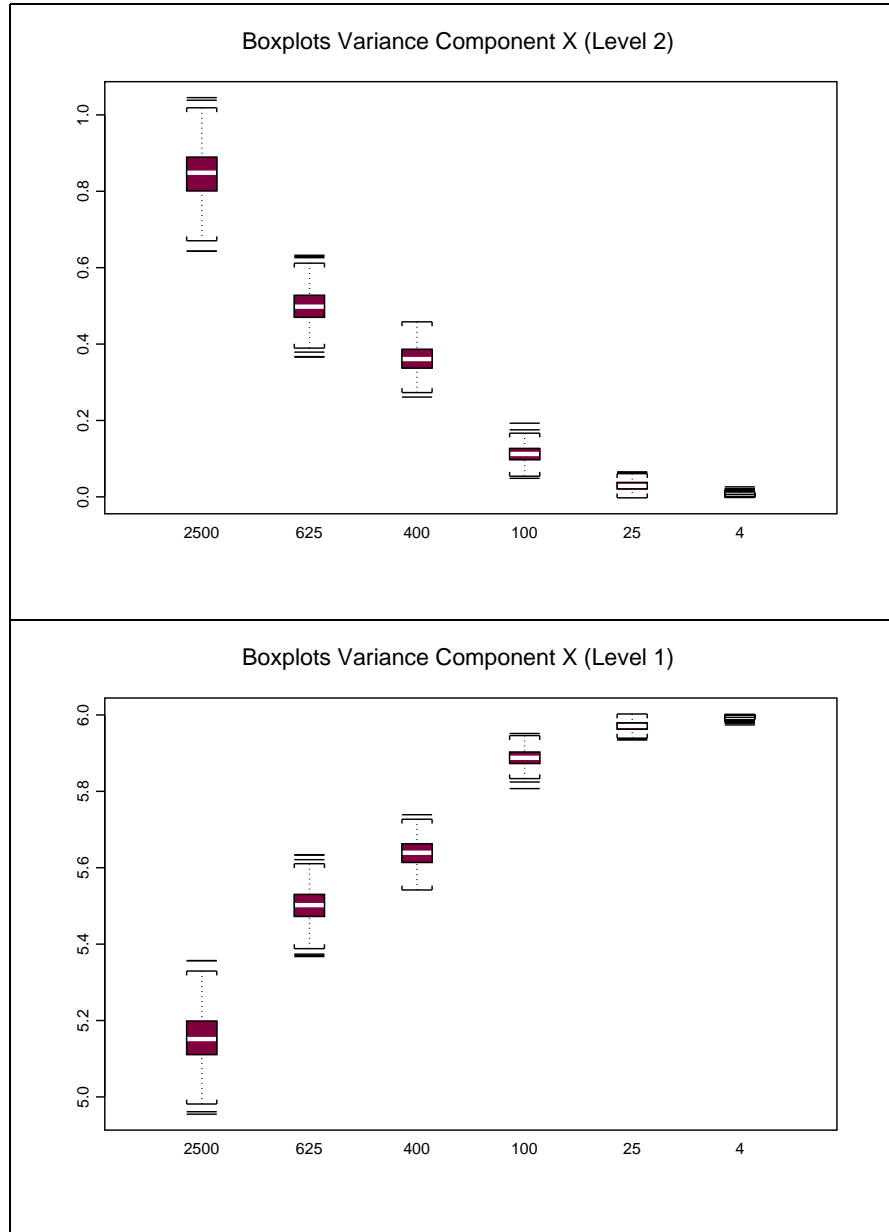


Figure 4.5: Variance Components of X, X have low autocorrelation

decrease with aggregation. Recall that the estimated level 2 variance components have numerator equal to  $S_{XX}^{(2)} - S_{XX}^{(1)}$  and denominator equal to  $\bar{N}^* - 1$  where  $\bar{N}^*$  is equal to  $N/M = n_{units}$  with values 4, 16, 25, 100, 400, and 2500. Both the numerator and denominator increase with aggregation. The variability of the numerator increase with aggregation but is divided by an increasing  $n_{units}$  thus resulting in a decreasing standard deviation of the level 2 variance component. The mean and me-

dian of the level 1 variance component increases with aggregation. However, in both cases, the standard deviation decreases with aggregation for the level 2 component and increases for the level 1 component.

Table (4.13) summarizes the distribution of the level 2 and level 1 variance components. The mean and the median have similar values and decrease with scale.

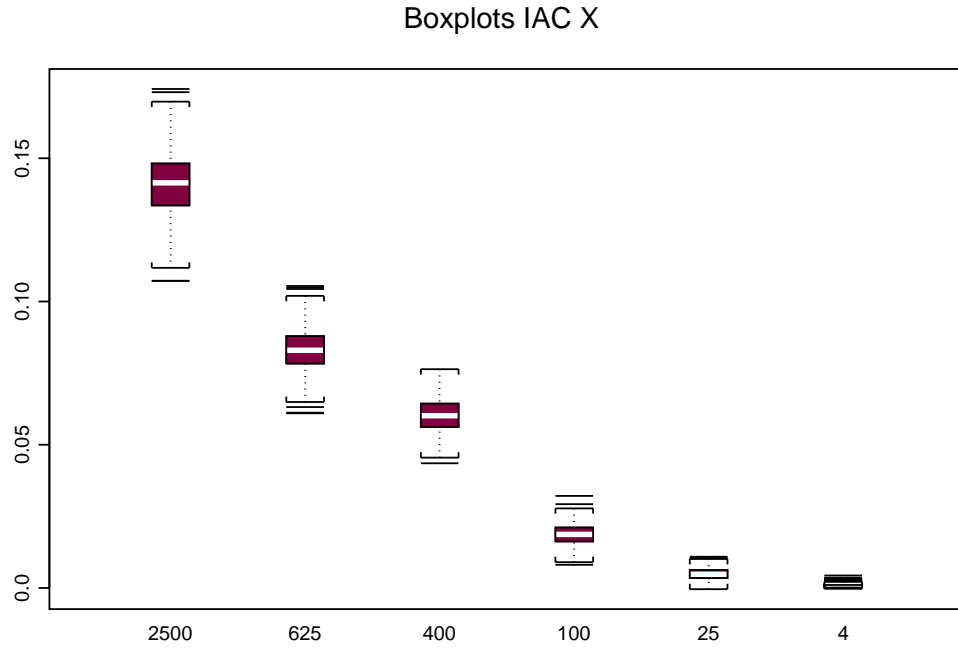
Level 2 Variance Component(X)	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.8467	0.8487	0.6429	1.0452	0.0668
625	0.4987	0.4978	0.3657	0.6324	0.0456
400	0.3607	0.3608	0.2611	0.4581	0.0359
100	0.1130	0.1121	0.0486	0.1928	0.0222
25	0.0296	0.0287	-0.0026	0.0654	0.0121
4	0.0041	0.0031	-0.0018	0.0260	0.0044
Level 1 Variance Component(X)	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	5.1533	5.1513	4.9548	5.3571	0.0668
625	5.5013	5.5022	5.3676	5.6343	0.0456
400	5.6393	5.6392	5.5420	5.7389	0.0359
100	5.8870	5.8879	5.8072	5.9514	0.0222
25	5.9704	5.9713	5.9346	6.0026	0.0121
4	5.9959	5.9969	5.9740	6.0018	0.0044

**Table 4.13: Description of the Level 2 and Level 1 Variance Components, X and Y both have low autocorrelation**

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.1411	0.1414	0.1071	0.1742	0.0111
625	0.0831	0.0830	0.0609	0.1054	0.0076
400	0.0601	0.0601	0.0435	0.0763	0.0060
100	0.0188	0.0187	0.0081	0.0321	0.0037
25	0.0049	0.0048	-0.0004	0.0109	0.0020
4	0.0007	0.0005	-0.0003	0.0043	0.0007

**Table 4.14: Description of the Intra-Area Correlation, X have low autocorrelation**

Figure 4.6 shows the distribution of the IAC at different levels of aggregation. The mean and median and the standard deviation decrease with aggregation. Notice that the behavior of the standard deviation of the IAC is similar to that of the level 2 variance component. This is because the numerator of the estimate of intra-area



**Figure 4.6: Intra-area correlation X, X have low autocorrelation**

correlation is the level 2 variance component while the denominator is the variance of unit level data, which remains constant for the different levels aggregation.

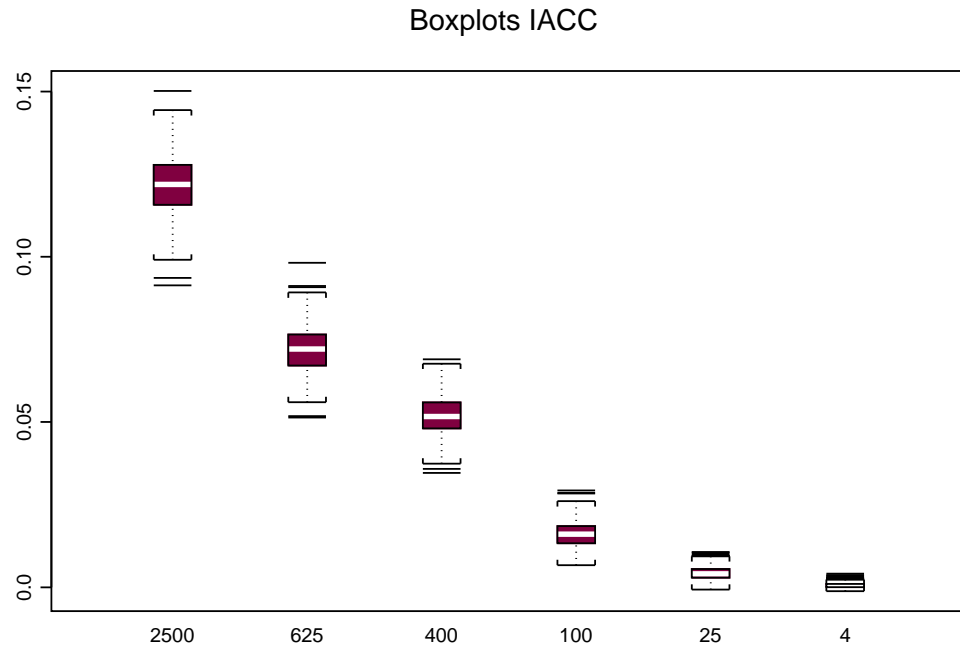
Figure 4.7 shows the behavior of the intra-area cross-correlation (IACC). The mean and the standard deviation of the mean decrease with the level of aggregation.

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.1220	0.1219	0.0914	0.1502	0.0091
625	0.0719	0.0721	0.0514	0.0982	0.0067
400	0.0520	0.0518	0.0346	0.0689	0.0059
100	0.0162	0.0161	0.0067	0.0293	0.0037
25	0.0043	0.0041	-0.0007	0.0107	0.0020
4	0.0006	0.0005	-0.0012	0.0041	0.0007

**Table 4.15: Description of the Intra-Area Cross-Correlation, X and Y both have low autocorrelation**

Figure 4.8 shows both the estimated level 1 and level 2 *pure correlations*. Looking at the figure the mean and median of the level 2 *pure correlation* is not affected by aggregation but the value is much higher than the initial Pearson correlation that



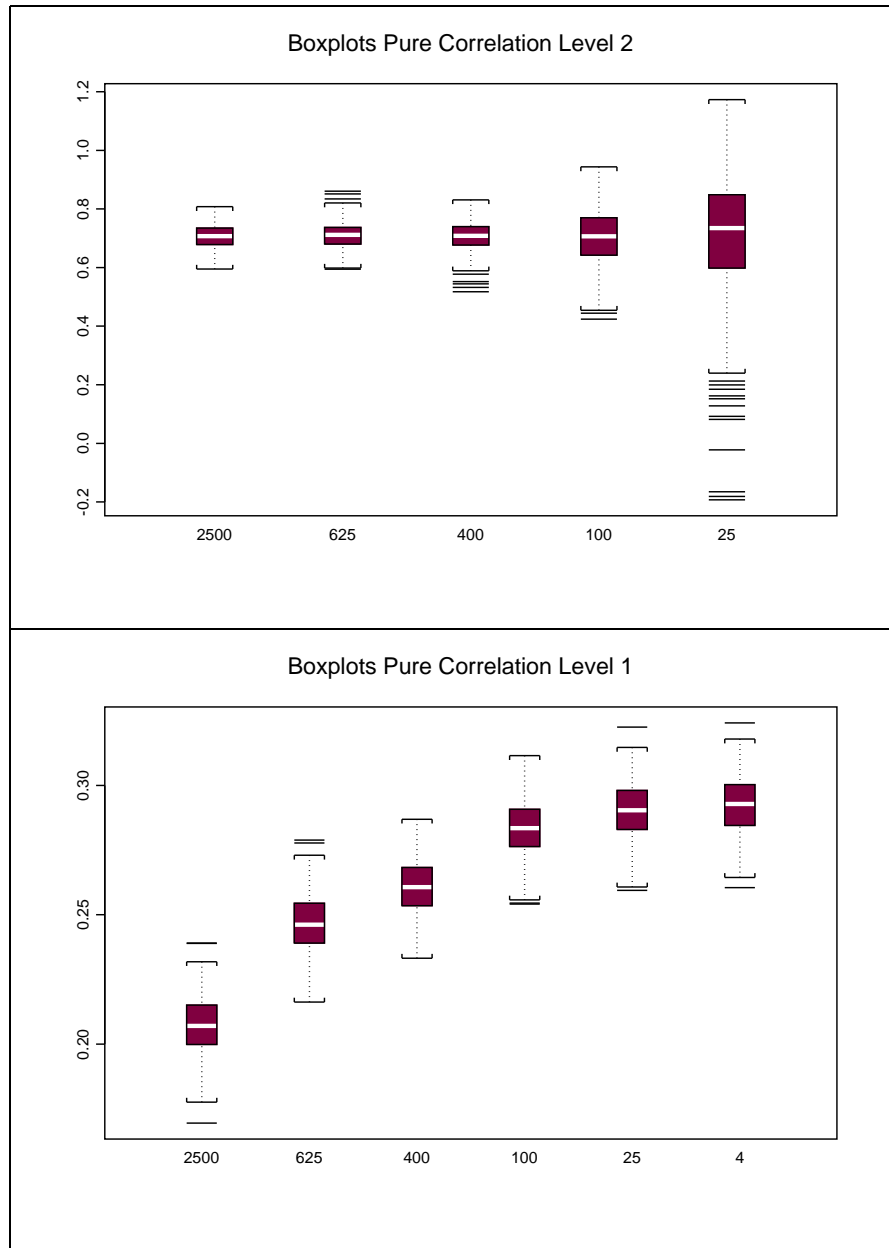


**Figure 4.7: Intra-area cross-correlation of X and Y, X and Y both have low autocorrelation**

ranges from 0.2599 to 0.3247 with mean 0.2929 (Table 4.10). Note that  $z_4(4 \text{ groups})$  is excluded because values ranges from -8.7860 to 7.3132 and inclusion would make examining the distribution for other groupings difficult, see Table 4.16. Note also that some level 2 *pure coefficients* when the data are aggregated into 25 zones are more than 1 or less than -1 which is not a characteristic of a correlation coefficient. This phenomenon will be examined latter in the chapter. Level 1 *pure correlations* seems to have a predictable pattern, with the mean and median increasing with aggregation and approaching the initial Pearson correlation of the generated data.

Table 4.16 shows the descriptive statistics of the pure coefficients. The standard deviation of the level 1 *pure correlations* seems to be constant and from Table 4.16 it can be seen that the standard deviation at different levels of aggregation is the same.

The behavior of the distributions of the *pure regression* coefficient is shown in Figure 4.9 and is similar to that of the distributions of the *pure correlation*. Again



**Figure 4.8: Pure Correlation, X and Y both have low autocorrelation**

z4 is not included because the values ranges from -66.3356 to 67.9241. The mean and median of level 2 *pure regression*, except when the number of groups is 4, are not affected by different levels of aggregation but are a little higher than the mean of the initial regression coefficient. The standard deviation increases with aggregation, the increase is slow at the first three levels of aggregation but gets larger as the number of groups decrease. Similar to the level 1 *pure correlation*, the level 1 *pure regression*

Level 2 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.7080	0.7071	0.5948	0.8079	0.03907
625	0.7093	0.7115	0.5942	0.8609	0.04145
400	0.7071	0.7086	0.5177	0.8312	0.04682
100	0.7036	0.7065	0.4241	0.9439	0.09026
25	0.7084	0.7345	-0.1928	1.1732	0.20077
4	0.7871	0.9861	-8.7860	7.3132	1.2305
Level 1 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2076	0.2070	0.1695	0.2391	0.0110
625	0.2465	0.2461	0.2163	0.2789	0.0103
400	0.2605	0.2607	0.2332	0.2869	0.0102
100	0.2834	0.2835	0.2542	0.3115	0.0105
25	0.2904	0.2904	0.2594	0.3225	0.0105
4	0.2925	0.2928	0.2605	0.3241	0.0106

**Table 4.16: Description of the Level 2 and Level 1 Pure Correlation, X and Y both have low autocorrelation**

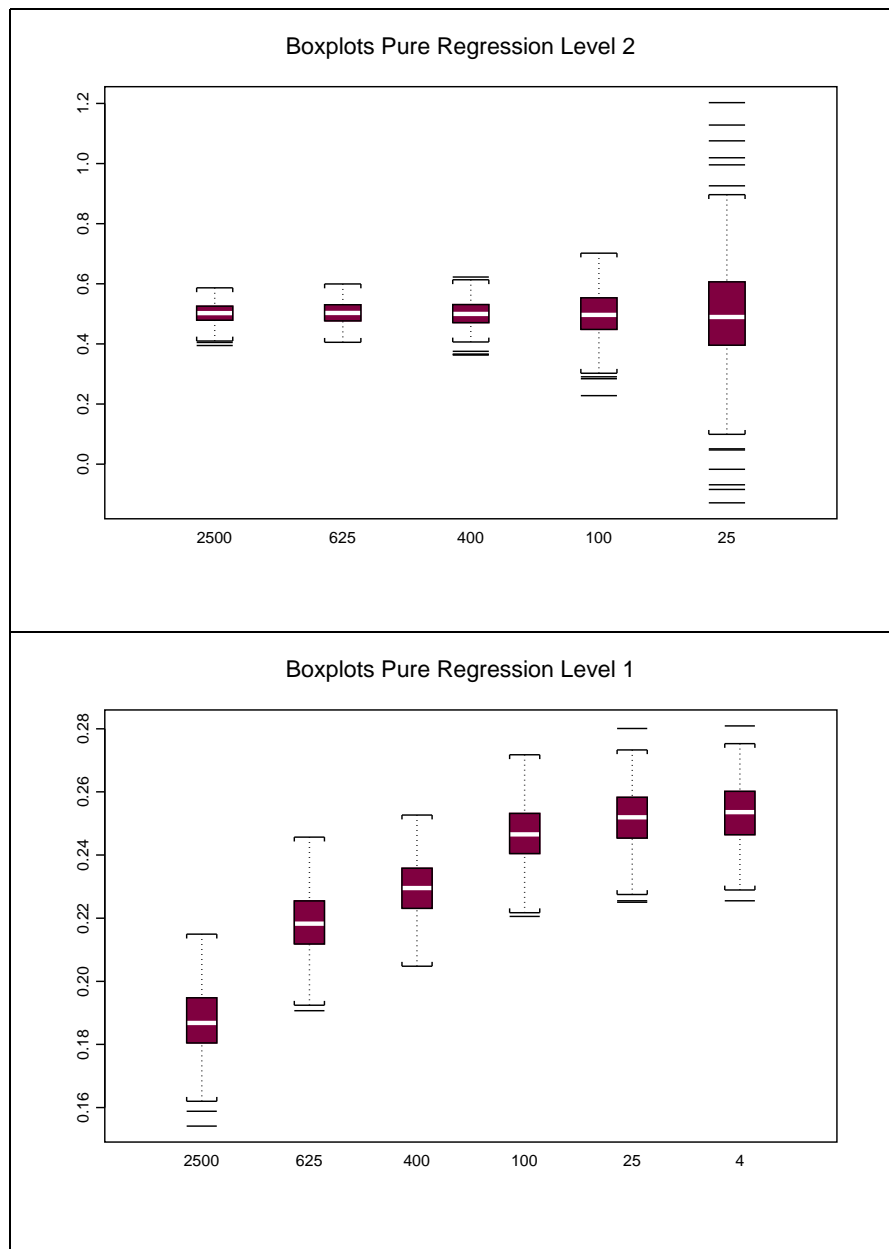
increases as the number of groups decreases. The standard deviation seems to be constant as depicted by Figure 4.9.

Table 4.17 shows descriptive statistics of level 2 and level 1 pure regression coefficients.

## 4.2.2 Data Set 2: Both variables have medium autocorrelation

To look into the effects of aggregation on variables with higher autocorrelation the simulations are repeated but with higher levels of autocorrelations for both variable. To generate a new set of data with a higher autocorrelation, the smoothing process used to generate data set 1 is repeated for the two variables. This is done by taking the average of the neighbors of each of the data points for the previously generated data. We initially examine results for one realization.

The resulting values of Moran's I are shown in Table 4.18. The contiguity matrix used is the queen's case. Note that this time the initial value, that is, the Moran's I at individual level is higher than that of the first data set. The values decrease with



**Figure 4.9: Pure Regression, X and Y both have low autocorrelation**

aggregation.

The unweighted and weighted variance are affected by scale as shown in Table 4.19 and Table 4.20, respectively. Looking at the decrease of the unweighted variance on both variables, it can be seen that the change is somewhat slow compared with the data set 1 (Tables 4.8 and 4.9). The covariance of the new set of data is smaller than the first one. The decrease of the covariance is slower compared with the

Level 2 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.5018	0.5020	0.3943	0.5863	0.0336
625	0.5033	0.5028	0.4052	0.5992	0.0379
400	0.5009	0.4996	0.3631	0.6221	0.0428
100	0.5006	0.4964	0.2280	0.7019	0.0789
25	0.5010	0.4891	-0.1289	1.2025	0.1772
4	0.5662	0.5002	-66.3356	67.9241	4.9090
Level 1 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.1876	0.1868	0.1541	0.2149	0.0101
625	0.2184	0.2182	0.1907	0.2457	0.0092
400	0.2293	0.2295	0.2048	0.2526	0.0090
100	0.2466	0.2465	0.2206	0.2718	0.0091
25	0.2518	0.2520	0.2251	0.2801	0.0092
4	0.2534	0.2536	0.2255	0.2809	0.0092

Table 4.17: Description of the Level 2 and Level 1 Pure Regression, X and Y both have low autocorrelation

Level	$I_{XX}^l$	$I_{YY}^l$	$I_{YX}^l$
Individual	0.4549	0.6223	0.1847
Z2500	0.3740	0.4602	0.1582
Z625	0.1278	0.1797	0.0881
Z400	0.1169	0.1326	0.0745
Z100	0.0259	0.1400	0.0383
Z25	-0.0676	0.1202	0.0160
Z4	-	-	-

Table 4.18: Moran's I

decrease of the covariance of the first data set.

	$\bar{X}$	$\bar{Y}$	$\tilde{S}_{XX}^{(l)}$	$\tilde{S}_{YY}^{(l)}$	$\tilde{S}_{XY}^{(l)}$
Individual	0.005	10.000	6.0000	8.0000	1.9382
Z2500	0.005	10.000	3.7314	5.9828	1.4873
Z625	0.005	10.000	2.0727	3.7572	0.9722
Z400	0.005	10.000	1.5626	2.9484	0.8254
Z100	0.005	10.000	0.5814	1.0004	0.3306
Z25	0.005	10.000	0.1600	0.3550	0.1336
Z4	0.005	10.000	0.0214	0.2051	0.0247

Table 4.19: Unweighted variance and covariance, X and Y both have medium autocorrelation

Table 4.21 shows that the correlation and regression coefficients tend to increase

	$\bar{X}$	$\bar{Y}$	$S_{XX}^{(l)}$	$S_{YY}^{(l)}$	$S_{XY}^{(l)}$
Individual	0.005	10.000	6.0000	8.0000	1.9382
Z2500	0.005	10.000	14.9256	23.9312	5.9493
Z625	0.005	10.000	33.1628	60.1147	15.5557
Z400	0.005	10.000	39.0645	73.7104	20.6348
Z100	0.005	10.000	58.1737	100.0409	33.0574
Z25	0.005	10.000	64.0128	141.9836	53.4285
Z4	0.005	10.000	53.4122	512.7969	61.6289

**Table 4.20: Weighted variance and covariance, X and Y both have medium autocorrelation**

with scale, with the exception of the correlation coefficient for 4 zones. Comparing the initial correlation and the correlation at different levels of aggregation with that of the first data (Table 4.3) it can be noted that the increase from initial correlation up to the different levels of aggregation seems to be slower. There is even a decrease when there are only 4 zones. A similar trend is observed with the regression coefficient.

	Correlation Coefficient	Regression Coefficient
Individual Data	0.2798	0.3230
Number of Zones		
2500	0.3148	0.3986
625	0.3484	0.4691
400	0.3845	0.5282
100	0.4356	0.5685
25	0.5604	0.8347
4	0.3724	1.1538

**Table 4.21: Correlation and regression coefficients at different scales, X and Y both have medium autocorrelation**

The estimated variance components and the intra-area correlations are shown in Table 4.22 and Table 4.23. The values decrease as the level of aggregation is increased and approach zero as the number of zones is decreased. This is because when there are few zones, the population within them will be almost as heterogeneous as the whole population. The level 2 variance component is larger than the level 1 variance component for both variables initially. The level 1 variance component approaches the individual level variance as the number of zones decreases. This

time the Level 2 variance component is larger at each aggregation level compared with the previous data set (Tables 4.4 and 4.23), probably due to the higher level of autocorrelation.

Level 1	Level 2	$\hat{\Lambda}_{XX}^{(2)}$	$\hat{\Lambda}_{XX}^{(1)}$	$\hat{\delta}_{XX}$
Individual	Z2500	2.9736	3.0264	0.4956
	Z625	1.8078	4.1922	0.3013
	Z400	1.3741	4.6259	0.2290
	Z100	0.5214	5.4786	0.0869
	Z25	0.1398	5.8602	0.0233
	Z4	0.0152	5.9848	0.0025

**Table 4.22:** Intra-Area correlations and variance components of X, X have medium autocorrelation

Level 1	Level 2	$\hat{\Lambda}_{YY}^{(2)}$	$\hat{\Lambda}_{YY}^{(1)}$	$\hat{\delta}_{YY}$
Individual	Z2500	5.3076	2.6924	0.6634
	Z625	3.4684	4.5316	0.4335
	Z400	2.7308	5.2692	0.3414
	Z100	0.9204	7.0796	0.1151
	Z25	0.3229	7.6771	0.0404
	Z4	0.1616	7.8384	0.0202

**Table 4.23:** Intra-Area correlations and variance components of Y, Y have medium autocorrelation

Looking at the results shown in Table 4.24 the intra-area cross-correlation have values greater than the previous data set at each level of aggregation. This time the level 2 covariance component when the data are aggregated into 2500 zones is larger than the level 1 covariance component and decreases as the number of zones decreases.

Level 1	Level 2	$\hat{\Lambda}_{YX}^{(2)}$	$\hat{\Lambda}_{YX}^{(1)}$	$\hat{\delta}_{YX}$
Individual	Z2500	1.3363	0.6019	0.1929
	Z625	0.9063	1.0320	0.1308
	Z400	0.7770	1.1612	0.1122
	Z100	0.3112	1.1637	0.0450
	Z25	0.1241	1.8142	0.0179
	Z4	0.0191	1.9191	0.0028

**Table 4.24:** Intra-Area cross-correlations at two levels, X and Y both have medium autocorrelation

*Pure correlations* for the second data set are shown in Table 4.25. The level 2 *pure correlation* increases with the level of aggregation and begins with a value not far from the individual level Pearson correlation. In comparison with the first data set, the corresponding values of the correlation are smaller. The level 1 *pure correlation* increases with the level of aggregation where the values are smaller than the initial Pearson correlation and seems to approach the individual(or initial) level Pearson correlation. Level 2 pure regression increases with aggregation except when the number of zones is 4 where it suddenly decreases. The level 2 pure regression seems to be not affected by aggregation.

Level 1	Level 2	$\hat{\rho}_{YX}^{(2)}$	$\hat{\rho}_{YX}^{(1)}$	$\hat{b}_{YX}^{(2)}$	$\hat{b}_{YX}^{(1)}$
Individual	Z2500	0.3364	0.2108	0.2518	0.2236
	Z625	0.3619	0.2368	0.2613	0.2277
	Z400	0.4011	0.2352	0.2845	0.2204
	Z100	0.4492	0.2612	0.3381	0.2298
	Z25	0.5840	0.2705	0.3942	0.2363
	Z4	0.3858	0.2802	0.1182	0.2448

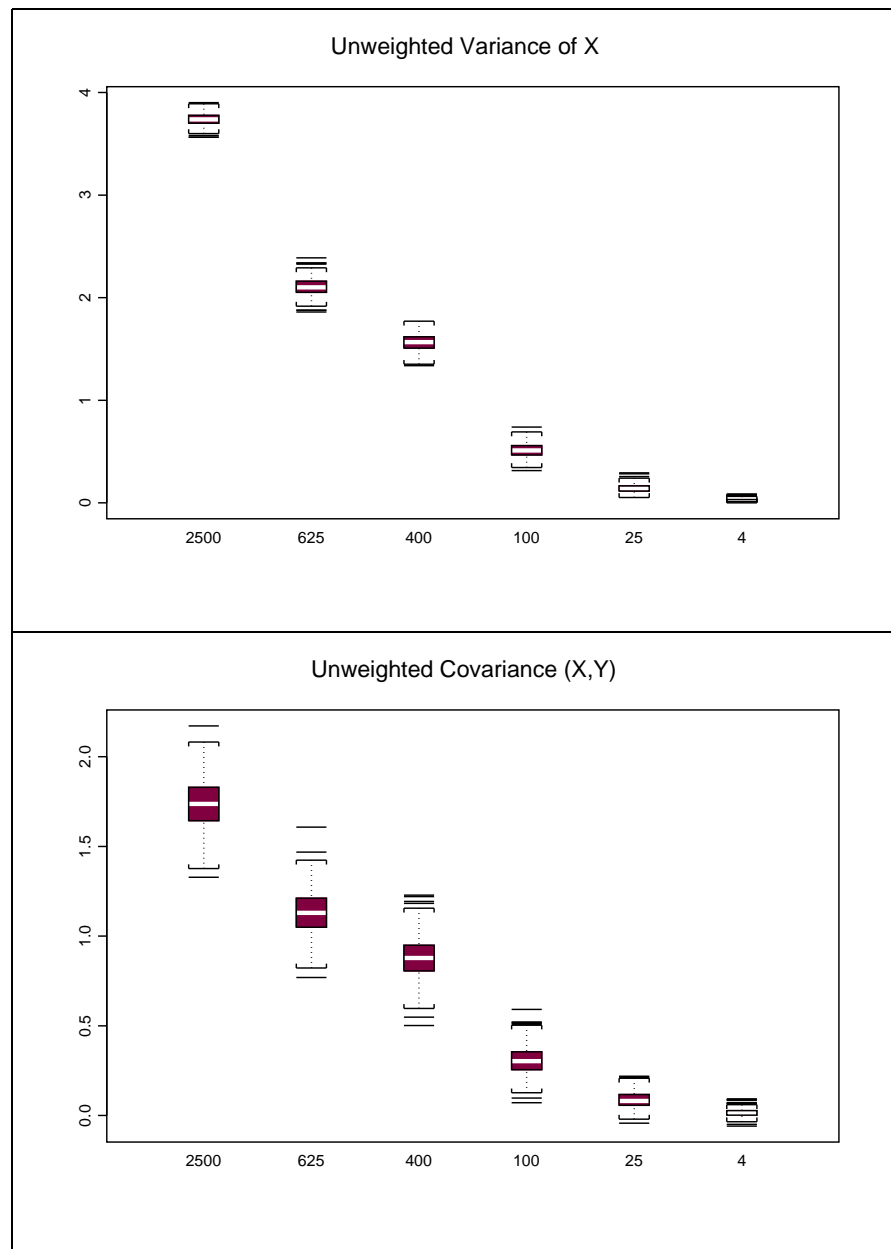
**Table 4.25: Pure correlations and pure regression at two levels, X and Y both have medium autocorrelation**

### Analysis of Distribution of Statistics When both Variables have Medium Autocorrelation

As with data set 1 the simulation is repeated 500 times to investigate the distributions of pertinent statistics.

Figure 4.10 shows the distributions of the unweighted variance of X and unweighted covariance of X and Y. The decreasing mean and median of the variance of X is similar to the results of the previous data set but the values are greater than the corresponding mean and median of data set 1. Aside from the first aggregation level (2500 groups) the standard deviation of the variance of X at different levels of aggregation decreases with aggregation. The decreasing trend of the mean of the covariance is observed but this time the change from unit level to the different levels of aggregation is slower than the corresponding values from data set 1. The



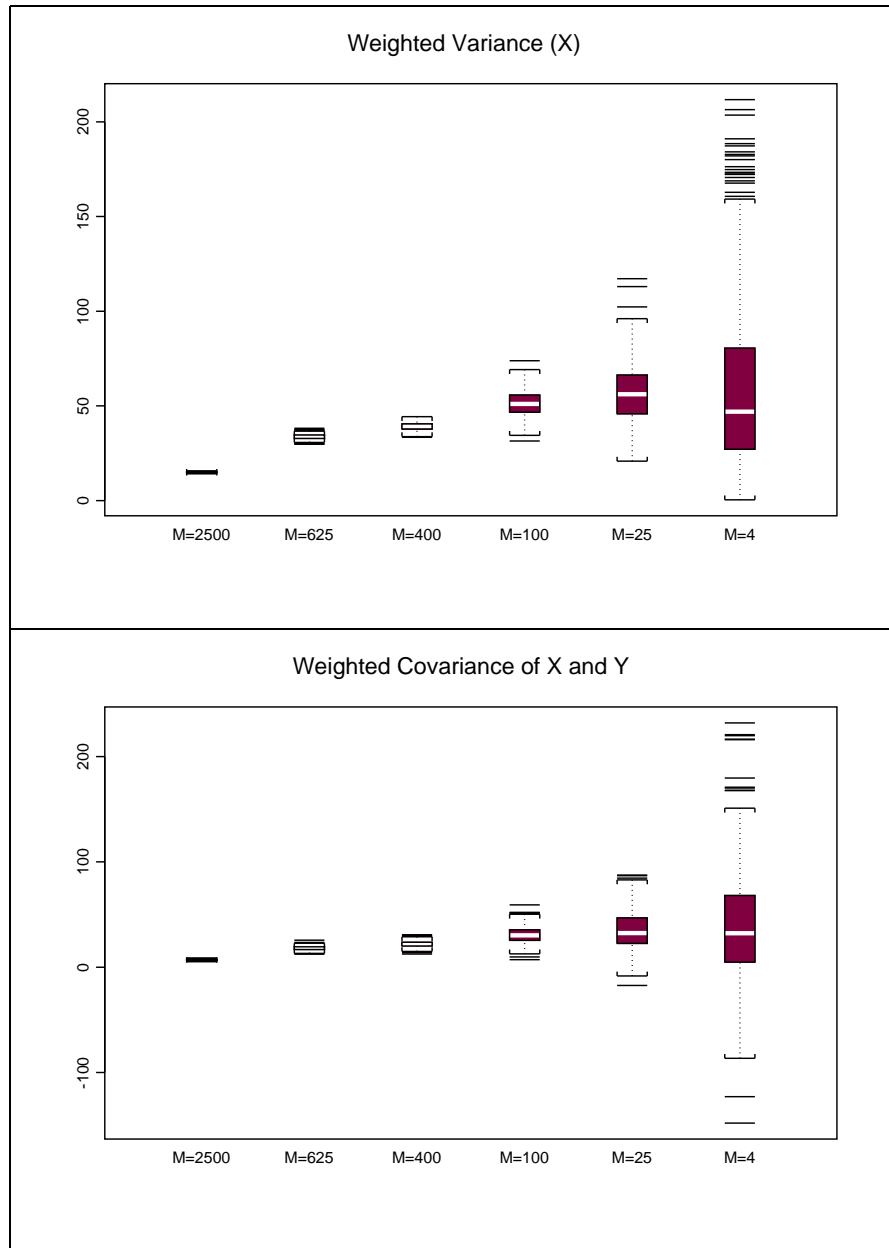


**Figure 4.10: Unweighted Variance of X and Covariance(X,Y), X and Y both have medium autocorrelation**

standard deviation also decreases with aggregation but is larger compared with the corresponding standard deviation of data set 1.

Table 4.26 summarizes the distribution of the estimated unweighted variance and covariance.

Figure 4.11 shows the distribution of the weighted variance of X. The mean of



**Figure 4.11: Weighted Variance of X and Covariance(X,Y), X and Y both have medium autocorrelation**

the weighted variances is larger than the corresponding values for the results from data set 1 at different levels of aggregation. The standard deviation increases with aggregation and is larger in magnitude than the corresponding standard deviation for different levels of aggregation in data set 1. The mean of the covariance also increases with aggregation. Looking at Table 4.27, there are values of the weighted covariance

Unweighted Variance of X	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	6.0000	6.0000	6.0000	6.0000	0.0000
2500	3.7375	3.7380	3.5637	3.9013	0.0590
625	2.1053	2.1007	1.8597	2.3878	0.0812
400	1.5645	1.5673	1.3380	1.7714	0.0795
100	0.5110	0.5105	0.3150	0.7384	0.0650
25	0.1415	0.1403	0.0521	0.2929	0.0381
4	0.0234	0.0188	0.0002	0.0847	0.0171
Unweighted Covariance of X and Y	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	2.1852	2.1844	1.8108	2.6272	0.1440
2500	1.7367	1.7370	1.3278	2.1716	0.1337
625	1.1323	1.1291	0.7690	1.6072	0.1177
400	0.8797	0.8777	0.5012	1.2284	0.1085
100	0.3064	0.3033	0.0715	0.5913	0.0765
25	0.0880	0.0811	-0.0432	0.2186	0.0445
4	0.0151	0.0129	-0.0592	0.0928	0.0199

**Table 4.26: Description of Unweighted Variance of X and Covariance of X and Y, X and Y both have medium autocorrelation**

that are less than the mean weighted covariance at unit level when the data are aggregated to 25 and 4 zones. The standard deviation increase with aggregation.

Figure 4.12 shows the correlation at different levels of aggregation. The figure shows that the mean and median of the correlation increases with the level of aggregation but the increase is slower than the corresponding increase of the correlation of data set 1. The increase is not as much as in data set 1 because of the decrease of both the variances X and Y and the covariance of X and Y is lesser when the data are aggregated into smaller number of groups, thus resulting in a lesser aggregation effect on the correlation. The standard deviation in each level of aggregation increases with aggregation and is similar to the first data set but with slightly larger values. In comparison with the standard deviation of data set 1, data set 2 has slightly higher standard deviations at each level but the trend is similar-they are increasing. The results show increasing correlation with aggregation but the increase is not as much as the increase of the correlation in data set 1. Similar behavior of the standard deviation is observed but slightly larger in magnitude compared with that of data set 1.

Weighted Variance of X	Mean	Median	Minimum	Maximum	Std. Dev.
2500	14.9501	14.9518	14.2550	15.6053	0.2359
625	33.6851	33.6105	29.7553	38.2041	1.2985
400	39.1126	39.1834	33.4487	44.2848	1.9880
100	51.0970	51.0469	31.4948	73.8388	6.4979
25	56.6032	56.0990	20.8290	117.1530	15.2350
4	58.4922	46.9210	0.3910	211.6970	42.8460
Weighted Covariance of X and Y	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	2.1852	2.1844	1.8108	2.6272	0.1439
2500	6.9469	6.9481	5.3110	8.6865	0.5348
625	18.1165	18.0657	12.3041	25.7149	1.8827
400	21.9924	21.9431	12.5299	30.7100	2.7131
100	30.6420	30.3340	7.1480	59.1280	7.6460
25	35.1790	32.4560	-17.2970	87.4560	17.7870
4	37.7311	32.2700	-147.9700	231.9100	49.8300

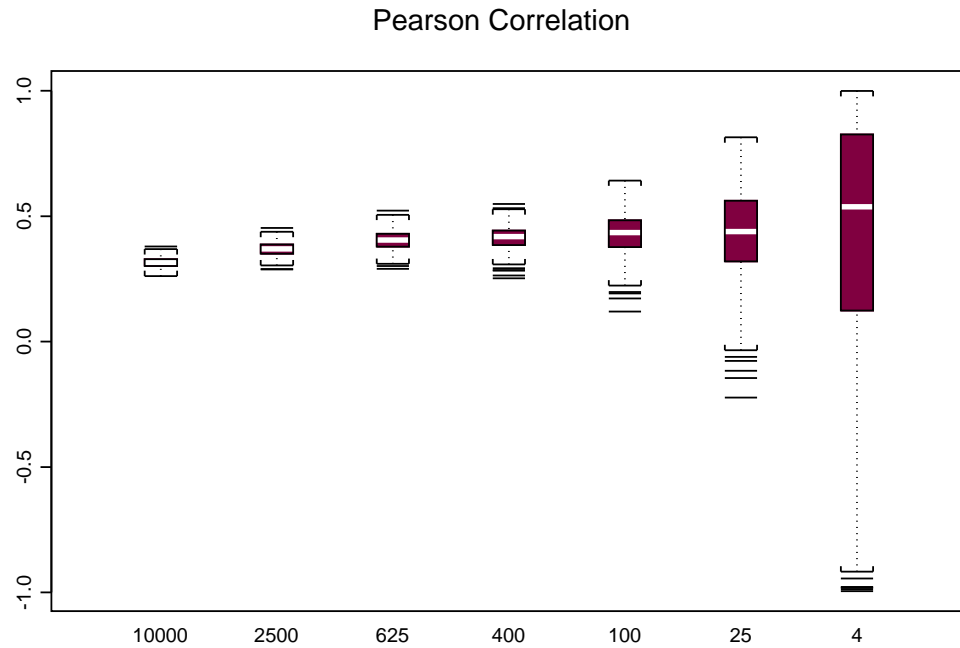
**Table 4.27: Description of Weighted Variance of X and Covariance of X and Y, X and Y both have medium autocorrelation**

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.3154	0.3153	0.2614	0.3792	0.0208
2500	0.3688	0.3696	0.2879	0.4529	0.0268
625	0.4040	0.4052	0.2900	0.5217	0.0369
400	0.4148	0.4183	0.2522	0.5486	0.0439
100	0.4291	0.4343	0.1196	0.6421	0.0834
25	0.4304	0.4385	-0.2234	0.8142	0.1684
4	0.4013	0.5377	-0.9950	0.9991	0.5207

**Table 4.28: Description of Pearson Correlation, X and Y both have medium autocorrelation**

Figure 4.13 shows the distribution of the regression coefficient. Note that results for aggregation to 4 zones are not included so that the distribution of regression coefficients for the other levels of aggregation can be displayed more clearly. A summary of the distribution of the regression coefficient when the data are aggregated into 4 groups is shown in Table 4.29.

Figure 4.14 shows the estimated variance components of variable X. The mean and median of the level 2 variance component decrease with aggregation. In com-

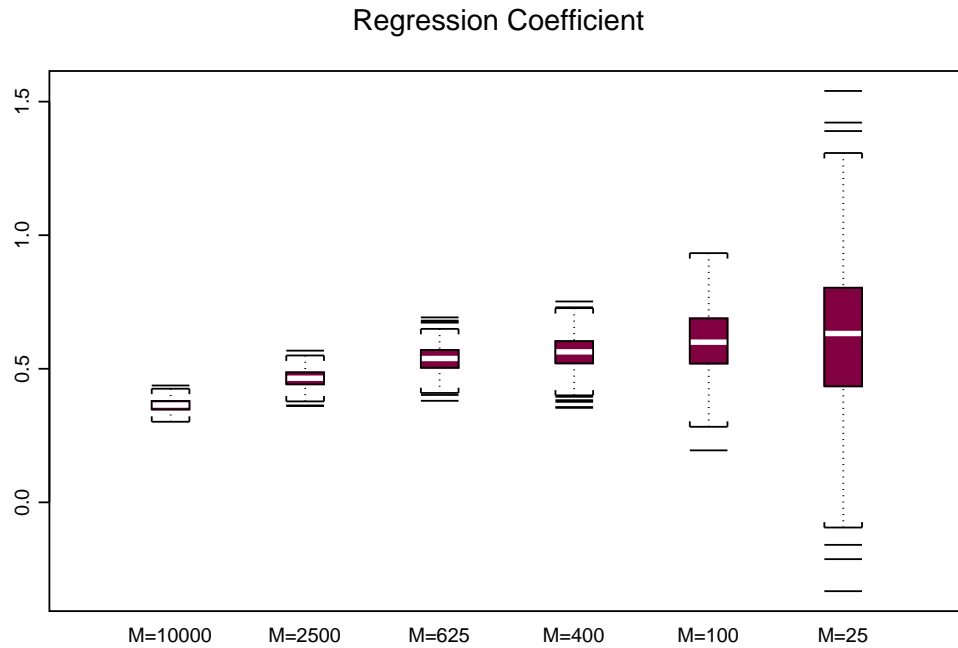


**Figure 4.12: Pearson Correlation** (The horizontal axis denotes number of groups), X and Y both have medium autocorrelation

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.3642	0.3641	0.3018	0.4379	0.0240
2500	0.4646	0.4650	0.3610	0.5681	0.0338
625	0.5377	0.5382	0.3802	0.6922	0.0502
400	0.5623	0.5635	0.3537	0.7523	0.0626
100	0.5997	0.5999	0.1947	0.9330	0.1288
25	0.6238	0.6321	-0.3324	1.5398	0.2712
4	0.6640	0.7130	-6.1942	14.3151	1.2784

**Table 4.29: Description of Regression Coefficient, X and Y both have medium autocorrelation**

parison with data set 1, the level 2 variance component of X is larger in all levels of aggregation. Looking at Table 4.22, the means of the level 2 variance component of X at different levels of aggregation are as follows; 2.9818, 1.8425, 1.3761, 0.4510, 0.1219, and 0.0168, respectively. In data set 1, the means of the level 2 variance component of X at different levels of aggregation are as follows; 0.8467, 0.4987, 0.3607, 0.1130, 0.0296, and 0.0041, respectively. Thus, as the degree of autocorrelation increases, the level 2 variance component at each level of aggregation

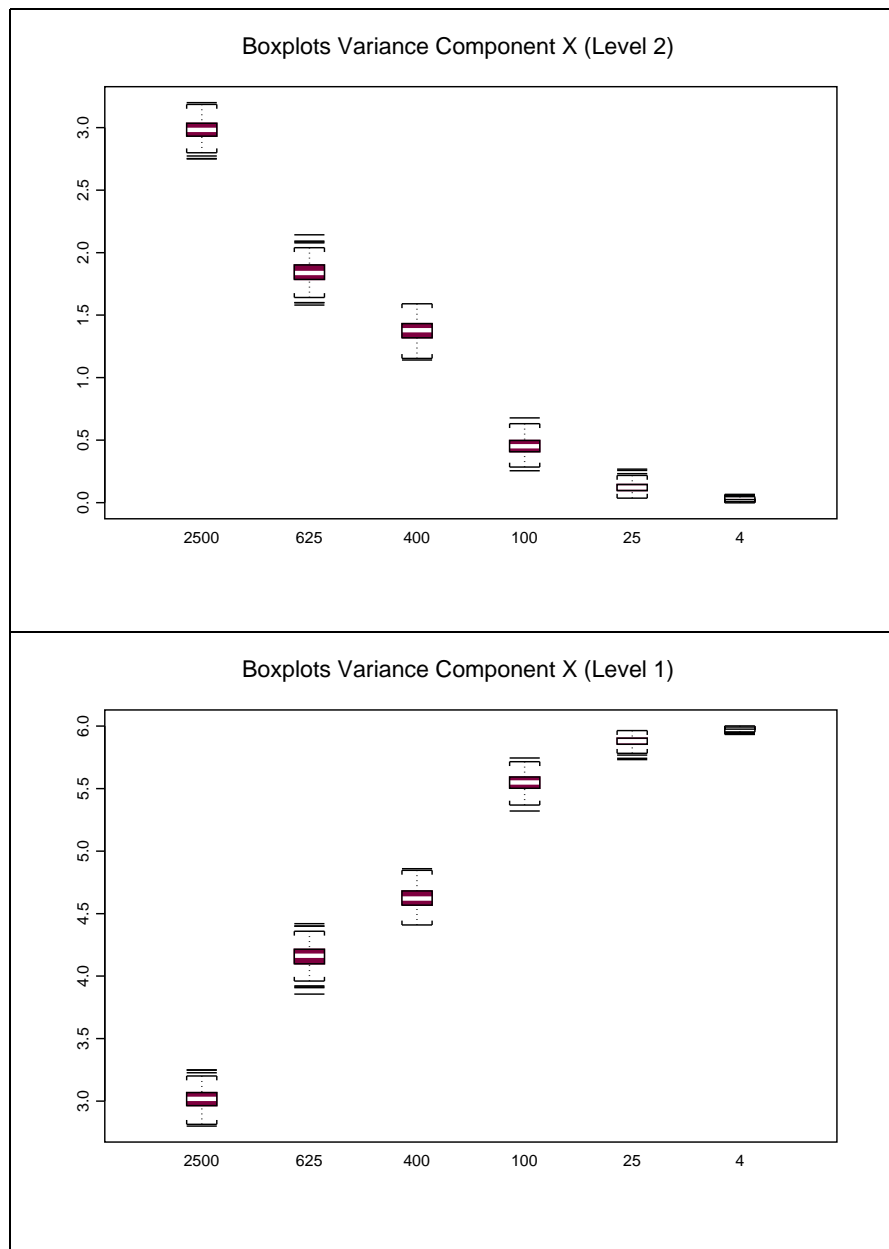


**Figure 4.13: Regression Coefficient, X and Y both have medium autocorrelation**

also increases. Like the unweighted variance, the standard deviation of the level 2 variance component decreases, starting from the second level of aggregation to the last. The ranges of the values from unit level to the last level of aggregation are a little larger than the corresponding ranges of data set 1. The values of the level 1 variance component are also affected by these results. The standard deviation of the values seems to decrease with aggregation but not as much as the decrease of the previous data set.

Figure 4.15 shows the distribution of the estimated intra-area correlation of variable X. The mean decreases with aggregation. The values are larger at all levels of aggregation compared with the results of data set 1. This is because in all levels of aggregation the level 2 variance component of data set 1 is smaller than the level variance component of data set 1. The standard deviation of the intra-area correlation decreases with aggregation. The magnitude of the standard deviations are a little larger than the corresponding standard deviations of data set 1.

Figure 4.16 shows both the level 2 and level 1 *pure correlation* coefficients. The



**Figure 4.14: Variance Components of X, X have medium autocorrelation**

mean and median of the level 2 *pure correlation* coefficients seem to change very slowly except when the number of groups is 4 for the median. The mean is greater than the initial correlation (3.154) and the standard deviation of these values increases with aggregation but these values are much lower than the corresponding values from data set 1 and are nearer to the initial correlation. Looking at the upper part of Table 4.32, the standard deviation of the level 2 *pure correlation* increases

Level 2 Variance Component(X)	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	2.9818	2.9824	2.7502	3.2001	0.0786
625	1.8425	1.8376	1.5810	2.1433	0.0864
400	1.3761	1.3791	1.1407	1.5911	0.0826
100	0.4510	0.4505	0.2550	0.6784	0.0650
25	0.1219	0.1207	0.0357	0.2678	0.0367
4	0.0168	0.0131	-0.0018	0.0658	0.0137
Level 1 Variance Component(X)	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	3.0182	3.0176	2.7999	3.2498	0.0786
625	4.1575	4.1624	3.8567	4.4190	0.0864
400	4.6239	4.6210	4.4089	4.8593	0.0826
100	5.5490	5.5495	5.3216	5.7451	0.0650
25	5.8781	5.8793	5.7322	5.9643	0.0367
4	5.9832	5.9870	5.9342	6.0018	0.0137

Table 4.30: Description of the Level 2 and Level 1 Variance Components, X and Y both have medium autocorrelation

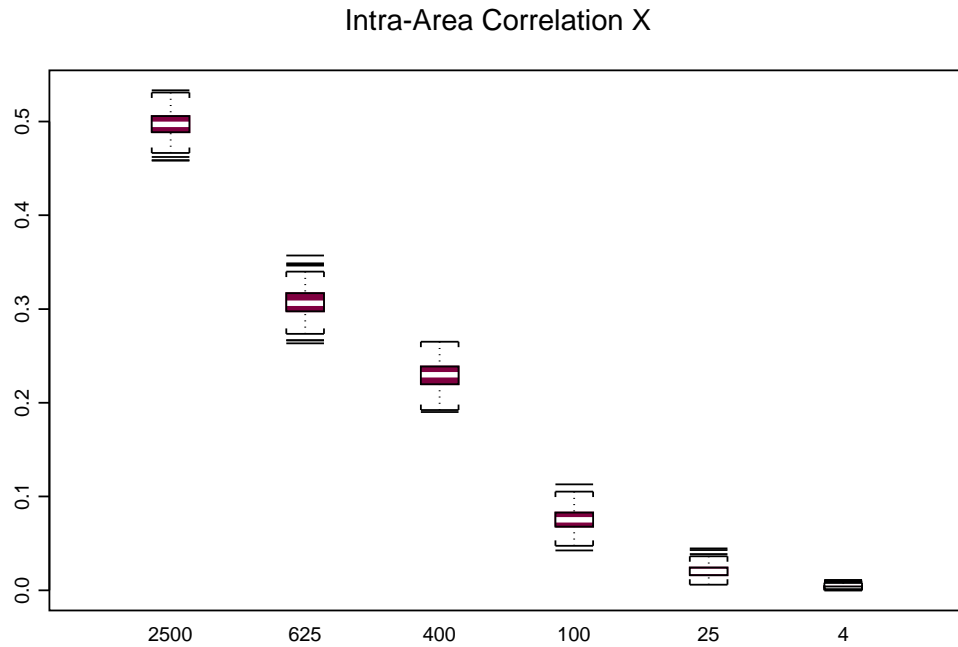


Figure 4.15: Intra-Area Correlation X, X have medium autocorrelation



Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.4970	0.4971	0.4584	0.5333	0.0131
625	0.3071	0.3063	0.2635	0.3572	0.0144
400	0.2296	0.2298	0.1901	0.2652	0.0138
100	0.0752	0.0751	0.0425	0.1131	0.0108
25	0.0203	0.0201	0.0060	0.0446	0.0061
4	0.0028	0.0022	-0.0003	0.0110	0.0023

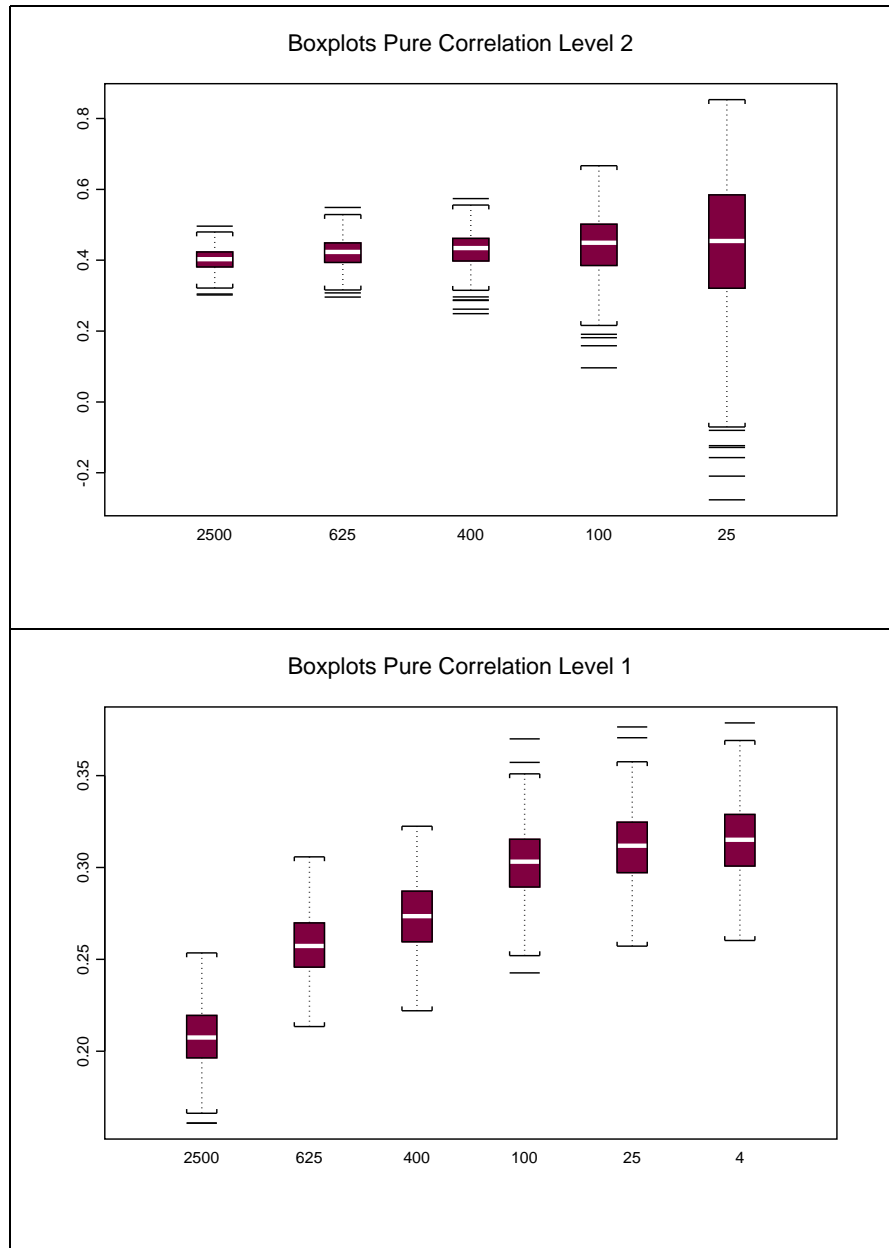
**Table 4.31: Description of Intra-Area Correlation (X), X both have medium autocorrelation**

with aggregation as depicted by Figure 4.16. These values are very similar to the corresponding values of data set 1 except for the last level of aggregation. The mean and median of the level 1 *pure correlation*, start with a smaller value and approaches the initial correlation as the number of groups decrease. These values are very similar to the corresponding values from data set 1. The standard deviation of these values seems to be approximately constant as depicted by Figure 4.16 and Table 4.32 but a little larger in magnitude compared with the results from data set 1.

Level 2 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.4012	0.4027	0.3019	0.4960	0.0308
625	0.4212	0.4234	0.2962	0.5487	0.0404
400	0.4308	0.4342	0.2489	0.5737	0.0480
100	0.4423	0.4491	0.0961	0.6665	0.0912
25	0.4425	0.4541	-0.2762	0.8531	0.1853
4	0.4270	0.6170	-2.6191	2.6062	0.6770
Level 1 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2073	0.2075	0.1609	0.2535	0.0163
625	0.2573	0.2573	0.2134	0.3058	0.0180
400	0.2735	0.2735	0.2221	0.3225	0.0189
100	0.3032	0.3032	0.2427	0.3700	0.0204
25	0.3119	0.3119	0.2572	0.3765	0.0206
4	0.3151	0.3151	0.2603	0.3787	0.0207

**Table 4.32: Description of the Level 2 and Level 1 Pure Correlation, X and Y both have medium autocorrelation**

Figure 4.17 shows the distribution of level 2 and level 1 *pure regression* coeffi-



**Figure 4.16: Pure Correlation, X and Y both have medium autocorrelation**

cients. The mean of the level 2 *pure regression* coefficient seems to be not affected by aggregation but the standard deviation of the values increases with aggregation as depicted by Figure 4.17 and the last column of the upper part of Table 4.33. These values are smaller than the corresponding values from data set 1 and a bit lower than the mean initial regression. The mean of the level 1 *pure regression* started with a value less than the initial regression coefficient at the individual level

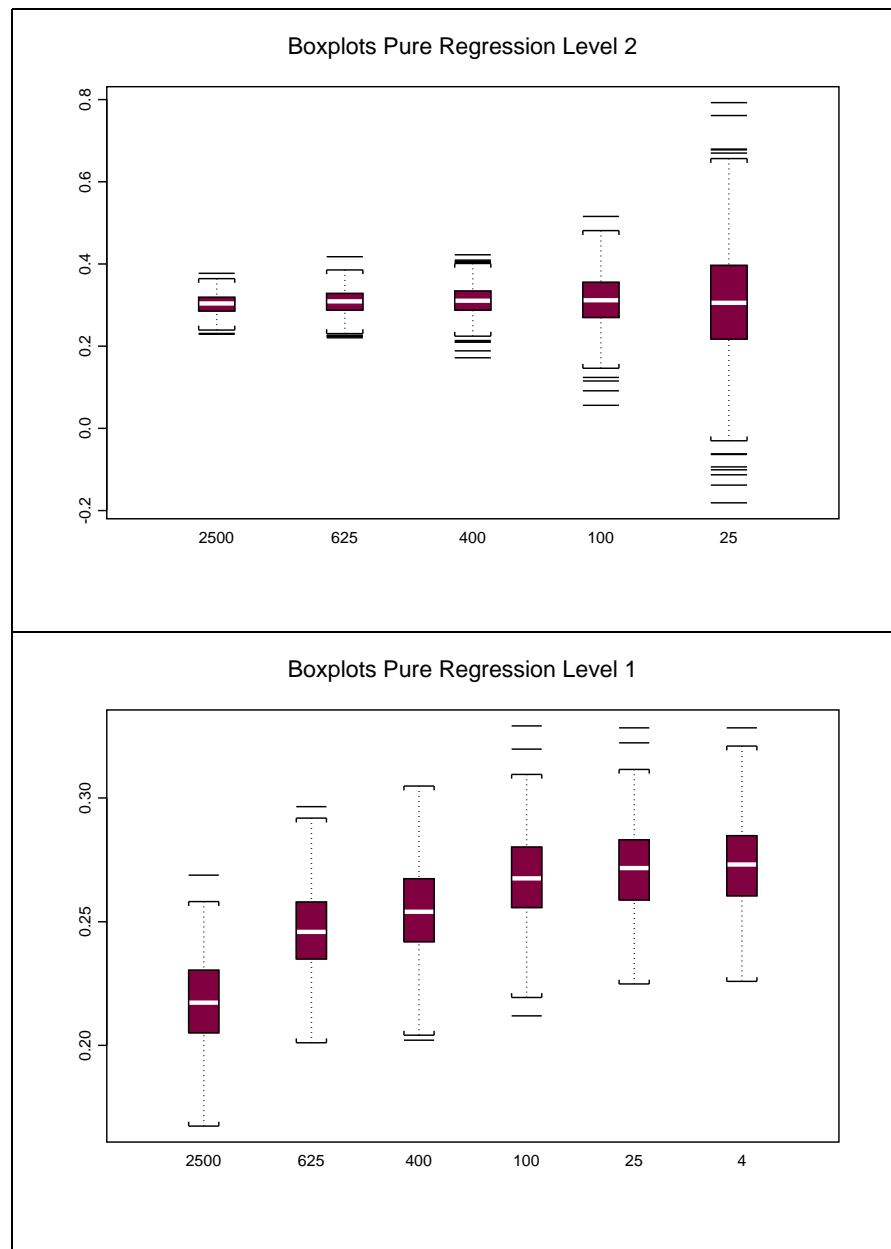


Figure 4.17: Pure Regression, X and Y both have medium autocorrelation

and slowly approaches that value as the number of groups decrease. The standard deviation in each level of aggregation seems to be constant just as the results from data set 1 but a bit larger.

Level 2 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.3027	0.3041	0.2291	0.3772	0.0239
625	0.3086	0.3089	0.2204	0.4176	0.0310
400	0.3106	0.3108	0.1719	0.4224	0.0363
100	0.3122	0.3118	0.0564	0.5155	0.0686
25	0.3082	0.3055	-0.1810	0.7925	0.1419
4	0.2827	0.3308	-20.5394	12.2534	1.3818
Level 1 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2172	0.2173	0.1674	0.2688	0.0175
625	0.2465	0.2459	0.2010	0.2965	0.0177
400	0.2546	0.2540	0.2021	0.3048	0.0180
100	0.2681	0.2675	0.2119	0.3291	0.0184
25	0.2719	0.2717	0.2248	0.3284	0.0181
4	0.2730	0.2731	0.2258	0.3283	0.0179

**Table 4.33: Description of the Level 2 and Level 1 Pure Regression, X and Y both have medium autocorrelation**

### 4.2.3 Data Set 3: Both variables have high autocorrelation

To be able to observe the behavior of pertinent statistics when two variables both have high autocorrelation, another set of data is generated. Data set 2 is again subjected to the smoothing process. The average of the neighbors of each of the data points of data Set 2 is recorded and becomes the new data set. The process was able to generate a data set that is more autocorrelated than the data set 2. As before, the following initial results are from one realization of the data generation. Table 4.34 shows the Moran's I at different levels of aggregation, which are higher than in data set 2.

Level	$I_{XX}^l$	$I_{YY}^l$	$I_{YX}^l$
Individual	0.7138	0.8033	0.2370
Z2500	0.5037	0.5757	0.1709
Z625	0.1621	0.2207	0.0754
Z400	0.1040	0.1361	0.0608
Z100	-0.0403	0.0378	0.0266
Z25	0.0155	0.2076	0.1155
Z4	-	-	-

**Table 4.34: Moran's I**

Table 4.35 shows the variances and covariances of the variables at different levels. The unweighted variance decreases as the number of zones decreases and the values are a little larger than the corresponding value in each level from the results from data set 2. The decrease is slower compared with data set 2 and data set 1. Except for the first aggregation, that is when the number of groups is 2500, the covariance at all levels of aggregation is larger than the corresponding covariances of data set 2 and data set 1.

	$\bar{X}$	$\bar{Y}$	$\tilde{S}_{XX}^{(l)}$	$\tilde{S}_{YY}^{(l)}$	$\tilde{S}_{XY}^{(l)}$
Individual	0.005	10.000	6.0000	8.0000	2.0259
Z2500	0.005	10.000	4.8461	6.9693	1.7602
Z625	0.005	10.000	3.1691	4.8816	1.1971
Z400	0.005	10.000	2.5333	4.1442	1.0289
Z100	0.005	10.000	0.9373	1.7078	0.4645
Z25	0.005	10.000	0.2598	0.5333	0.1836
Z4	0.005	10.000	0.0360	0.1399	0.0684

**Table 4.35: Unweighted variance and covariance, X and Y both have high autocorrelation**

Table 4.36 shows the weighted variance and covariance of the two variables. Both the variances of X and Y and the covariance increase with aggregation except for the variance of X when  $m=4$ . The values of these statistics are larger than the corresponding statistics computed from data set 2 in all levels of aggregation, reflecting the higher level of autocorrelation.

	$\bar{X}$	$\bar{Y}$	$S_{XX}^{(l)}$	$S_{YY}^{(l)}$	$S_{XY}^{(l)}$
Individual	0.005	10.000	6.0000	8.0000	2.0259
Z2500	0.005	10.000	19.3844	27.8773	7.0409
Z625	0.005	10.000	50.7071	78.1063	19.1530
Z400	0.005	10.000	63.3335	103.6046	25.7221
Z100	0.005	10.000	93.7345	170.7823	46.4457
Z25	0.005	10.000	103.9118	213.3024	73.4217
Z4	0.005	10.000	90.0712	349.8199	171.0138

**Table 4.36: Weighted variance and covariance, X and Y both have high autocorrelation**

Table 4.37 shows the correlation and regression coefficients at different levels. Again, the pattern of increasing correlation and regression coefficients as the number of zones is decreased is evident. The increase of the correlation with aggregation

except when the aggregated to 25 and 4 groups seems to be slow in comparison with the increase of the correlation of data set 1 and data set 2. The same pattern is observed with the regression coefficient, that is the increase is slow as the data are aggregated into smaller number of groups.

	Correlation Coefficient	Regression Coefficient
Individual Data	0.2924	0.3376
Number of Zones		
2500	0.3029	0.3642
625	0.3043	0.3777
400	0.3175	0.4061
100	0.3671	0.4955
25	0.4932	0.7067
4	0.9634	1.8987

**Table 4.37: Correlation and regression coefficients at different scales, X and Y both have high autocorrelation**

The intra area correlations of variable X is shown in Table 4.38. The values are larger than the previous data set because the data generation will generate data with higher autocorrelation. The value of the intra-area correlation is almost equal to Moran's I at the individual level when the data are aggregated into 2500 groups. The level 2 variance component is initially much larger than the level 1 variance component, and decreases and seems to approach zero as the number of zones is decreased. The level 1 variance component approaches the individual level variance as the number of zones is decreased.

Level 1	Level 2	$\hat{\Lambda}_{XX}^{(2)}$	$\hat{\Lambda}_{XX}^{(1)}$	$\hat{\delta}_{XX}$
Individual	Z2500	4.4591	1.5409	0.7432
	Z625	2.9754	3.0246	0.4959
	Z400	2.3827	3.6173	0.3971
	Z100	0.8873	5.1227	0.1462
	Z25	0.2359	5.7641	0.0393
	Z4	0.0269	5.9731	0.0045

**Table 4.38: Intra-Area correlations and variance components of X, X have high autocorrelation**

Table 4.39 shows the intra-area correlation of variable Y. The behavior of the scale effect is similar to the intra-area correlation of variable X but larger in value

at each level of aggregation, this is because of the way the data are generated. The level 2 and level 1 variance components have behavior similar to that of the variable X.

Level 1	Level 2	$\hat{\Lambda}_{YY}^{(2)}$	$\hat{\Lambda}_{YY}^{(1)}$	$\hat{\delta}_{YY}$
Individual	Z2500	6.6222	1.3778	0.8278
	Z625	4.6658	3.3342	0.5832
	Z400	3.9732	4.0268	0.4946
	Z100	1.6278	6.3722	0.2035
	Z25	0.4947	7.5053	0.0618
	Z4	0.0269	7.8906	0.0137

**Table 4.39: Intra-Area correlations and variance components of Y, Y have high auto-correlation**

Table 4.40 displays the covariance components that are used in the estimation of the intra-area cross-correlation. The intra-area cross correlation decreases with aggregation.

Level 1	Level 2	$\hat{\Lambda}_{YX}^{(2)}$	$\hat{\Lambda}_{YX}^{(1)}$	$\hat{\delta}_{YX}$
Individual	Z2500	1.6708	0.3551	0.2412
	Z625	1.1398	0.8860	0.1645
	Z400	0.9848	1.0411	0.1421
	Z100	0.4442	1.5817	0.0641
	Z25	0.1721	1.8538	0.0248
	Z4	0.0541	1.9718	0.0078

**Table 4.40: Intra-Area cross-correlations at two levels, X and Y both have high auto-correlation**

Table 4.41 shows level 2 and level 1 *pure correlations*. Both the level 2 and level 1 *pure correlation* increases with aggregation. The level 1 pure correlation approaches the initial correlation and the increase is slow in comparison with the increase in the previous two data sets. The level 2 pure correlation increase very slowly for the first three levels of aggregation.

The level 1 *pure regression* seems to have only a slight increase when the number of zones is 625 and decreases slightly at 400 zones and becomes stable for the rest of the levels of aggregation including 4 zones. The level 2 *pure regression* increases with aggregation except when the data are aggregated into 4 zones, in which case there is a sudden increase.

Level 1	Level 2	$\hat{\rho}_{YX}^{(2)}$	$\hat{\rho}_{YX}^{(1)}$	$\hat{b}_{YX}^{(2)}$	$\hat{b}_{YX}^{(1)}$
Individual	Z2500	0.3075	0.2437	0.2523	0.2578
	Z625	0.3059	0.2790	0.2443	0.2657
	Z400	0.3200	0.2728	0.2479	0.2585
	Z100	0.3717	0.2768	0.2729	0.2482
	Z25	0.5036	0.2818	0.3478	0.2470
	Z4	0.9968	0.2876	0.4944	0.2499

Table 4.41: Pure coefficients at two levels, X and Y both have high autocorrelation

### Analysis of Distribution of Statistics when both Variables have High Autocorrelation

As in the previous sub-sections, the data generation is repeated 500 times to investigate the distributions of some pertinent statistics. Figure 4.18 shows the distribution of the variance of variable X and the covariance of X and Y.

Unweighted Variance (X)	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	6.0000	6.0000	6.0000	6.0000	0.0000
2500	4.8941	4.8943	4.7439	5.00635	0.0428
625	3.2719	3.2674	2.9694	3.58812	0.0915
400	2.5808	2.5828	2.2233	2.90657	0.1088
100	0.9305	0.9342	0.5834	1.28437	0.1125
25	0.2684	0.2625	0.0928	0.55646	0.0720
4	0.0454	0.0374	0.0005	0.171253	0.0336
Unweighted Covariance of X and Y	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	2.1081	2.1098	1.4813	2.7771	0.2124
2500	1.8728	1.8819	1.2525	2.5238	0.2030
625	1.3644	1.3713	0.8132	2.0952	0.1809
400	1.1116	1.1077	0.5210	1.7280	0.1693
100	0.4280	0.4227	-0.0088	0.8833	0.1256
25	0.1277	0.1180	-0.1019	0.3436	0.0750
4	0.0224	0.0185	-0.1218	0.1651	0.0352

Table 4.42: Description of the Unweighted Variance(X) and Covariance(X,Y), X and Y both have high autocorrelation

Only the distributions of the variance of X is shown because the distributions of the variance of Y have similar pattern. The mean of the unweighted variance of X decreases with aggregation and the change of the mean of unweighted variance of variable X is slow compared with that of data set 2 and 1. The standard deviation



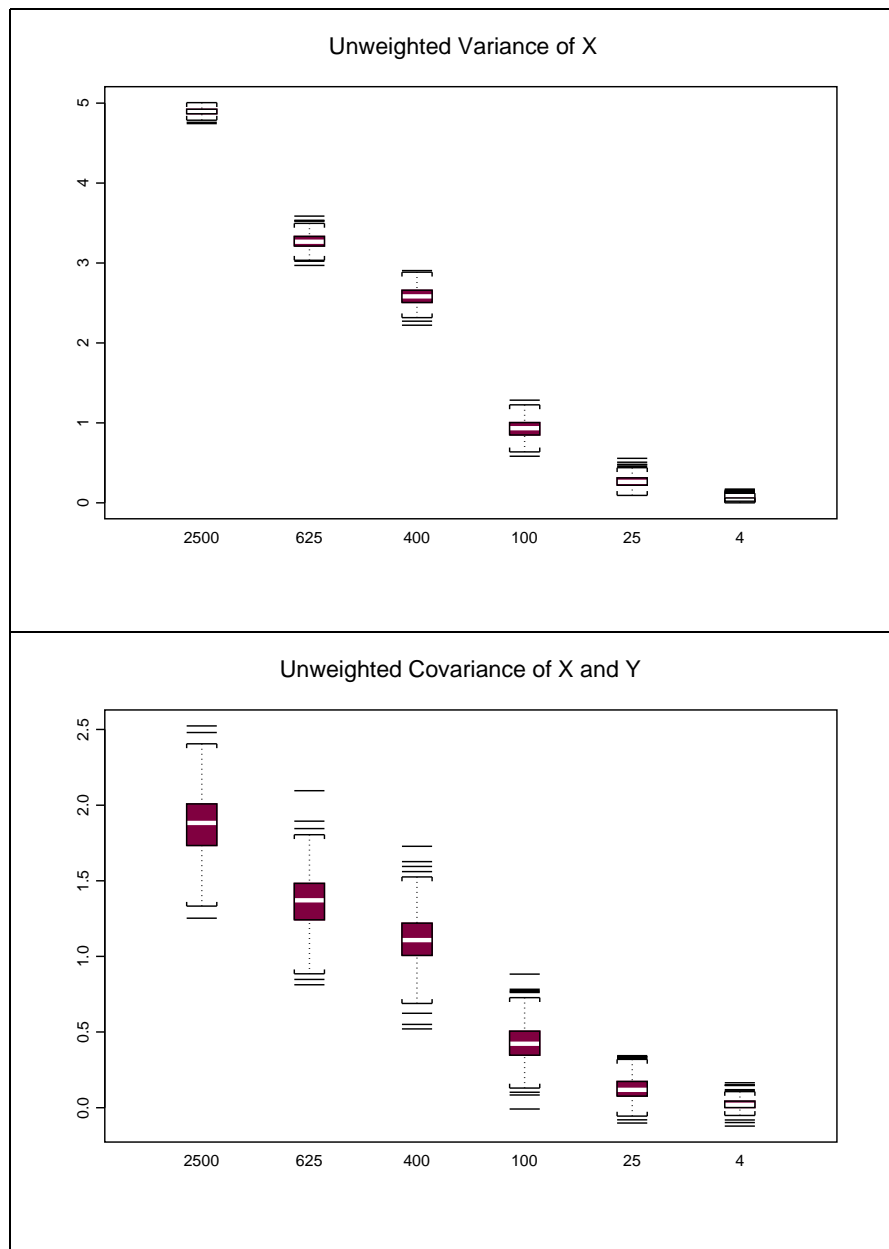


Figure 4.18: Unweighted Variance of X and Covariance (X,Y)

of the unweighted variance seems to be in agreement with the claim of Reynolds that "When significantly positive autocorrelated variables are aggregated, ... expect the variability of possible aggregate variance values to increase with a decrease in the number of cells." (page 23, Reynolds, 1988). Looking at Figure 4.18 and Table 4.42 the standard deviation of the unweighted variance displayed the same pattern. The result is very different when the variable have *low autocorrelation* (data set 1) and

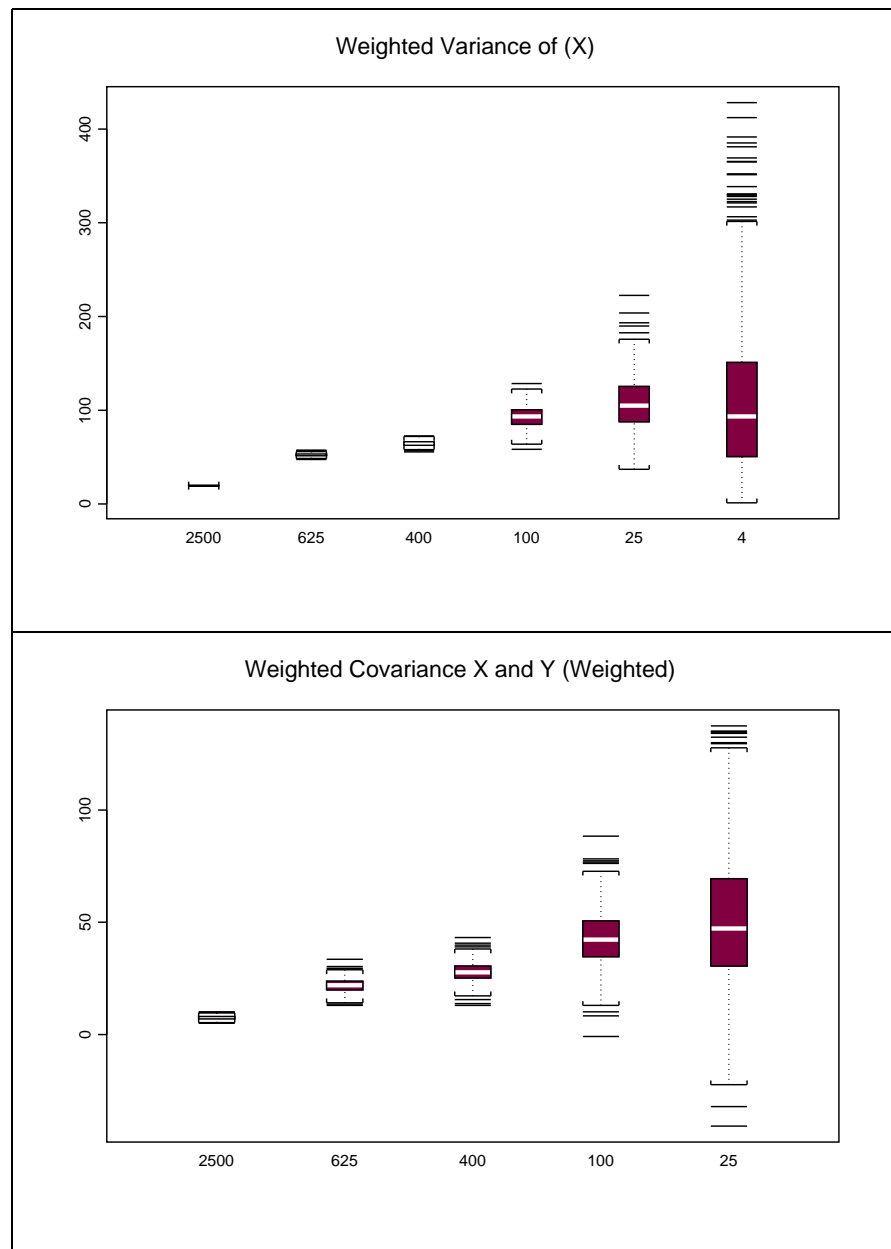
similar to when the variable have *medium autocorrelation*(data set 2). The mean of the covariance at different levels of aggregation decreases with aggregation and the values are larger than the results from data set 2. The standard deviation decrease with aggregation and the magnitude are larger than the corresponding value from data set 2.

Figure 4.19 shows the details of the weighted variance of X and weighted covariance of X and Y. The weighted variances increase with aggregation and have values greater than the corresponding values from data set 2. Table 4.43 shows the description of the distribution. Looking at the results of Figure 4.19 and Table 4.43, the standard deviation increases with aggregation. The magnitude of the standard deviations are larger than the corresponding results from data set 1 at different levels of aggregation. The mean weighted covariance increases with aggregation.

Weighted Variance (X)	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	6.0000	6.0000	6.0000	6.0000	0.0000
2500	27.9407	27.9489	27.4546	28.4352	0.1707
625	79.9964	80.1087	75.1792	84.9025	1.8018
400	101.6260	101.6700	91.1970	111.2201	3.8030
100	155.5266	154.9506	107.2980	208.5658	18.1176
25	186.2580	184.1820	71.3480	351.8320	49.0450
4	199.8920	164.8220	0.8660	1155.2280	154.0120
Weighted Covariance of X and Y	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	2.1081	2.1098	1.4813	2.7771	0.2124
2500	7.4914	7.5277	5.0098	10.0951	0.8121
625	21.8300	21.9410	13.0100	33.5230	2.8940
400	27.7894	27.6920	13.0243	43.2011	4.2328
100	42.8040	42.2730	-0.8770	88.3270	12.5600
25	51.0650	47.2130	-40.7540	137.4190	29.9950
4	56.0200	46.1580	-304.6190	412.7350	87.9640

**Table 4.43: Description of the Weighted Variance(X) and Covariance(X,Y), X and Y both have high autocorrelation**

Figure 4.20 shows the distribution of correlations at different levels of aggregation. The mean and median of the correlation increase in a very slow manner and decrease slightly when the data are aggregated to 4 zones. The slow decrease of the variances of variable X and Y and the covariance cause the slow increase of the mean of the correlations. The standard deviation increases with aggregation and



**Figure 4.19: Weighted Variance of X and Covariance(X,Y), X and Y both have high autocorrelation**

the magnitudes of the standard deviations are a little higher than the corresponding standard deviations from data set 2. The results show a very slow increase of the mean correlation as the data are aggregated into smaller number of groups.

Figure 4.21 shows the distribution of regression coefficients when the variables X and Y both have high autocorrelation. The trend is similar to that of the behavior

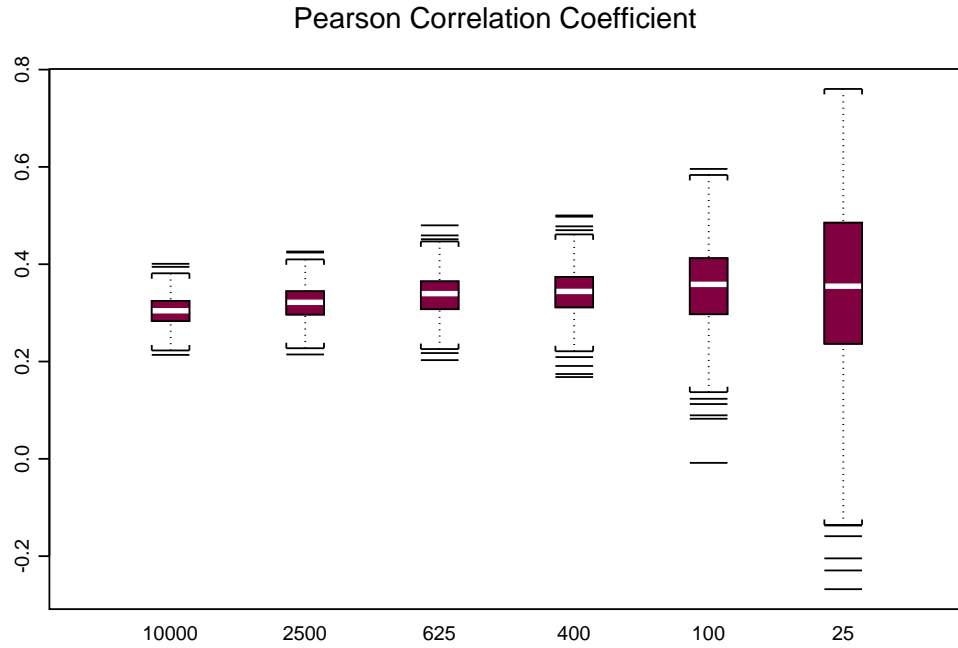


Figure 4.20: Pearson Correlation, X and Y both have high autocorrelation

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.3043	0.3045	0.2138	0.4009	0.0307
2500	0.3203	0.3218	0.2145	0.4260	0.0342
625	0.3372	0.3395	0.2031	0.4802	0.0425
400	0.3429	0.3442	0.1682	0.5003	0.0486
100	0.3542	0.3585	-0.0082	0.5958	0.0893
25	0.3553	0.3549	-0.2679	0.7604	0.1762
4	0.3306	0.4556	-0.9971	0.9994	0.5392

Table 4.44: Description of Pearson Correlation, X and Y both have high autocorrelation

of the correlation coefficient, with a low increase of the mean regression coefficient as the data are aggregated to smaller number of groups. The standard deviation increases with the level of aggregation.

Figure 4.22 shows the distribution of the estimated variance components of variable X. The mean of the level 2 variance component of X decrease with aggregation and the values are larger than the corresponding results from data set 1. The standard deviation increases up to when data are aggregated to 400 groups and decrease

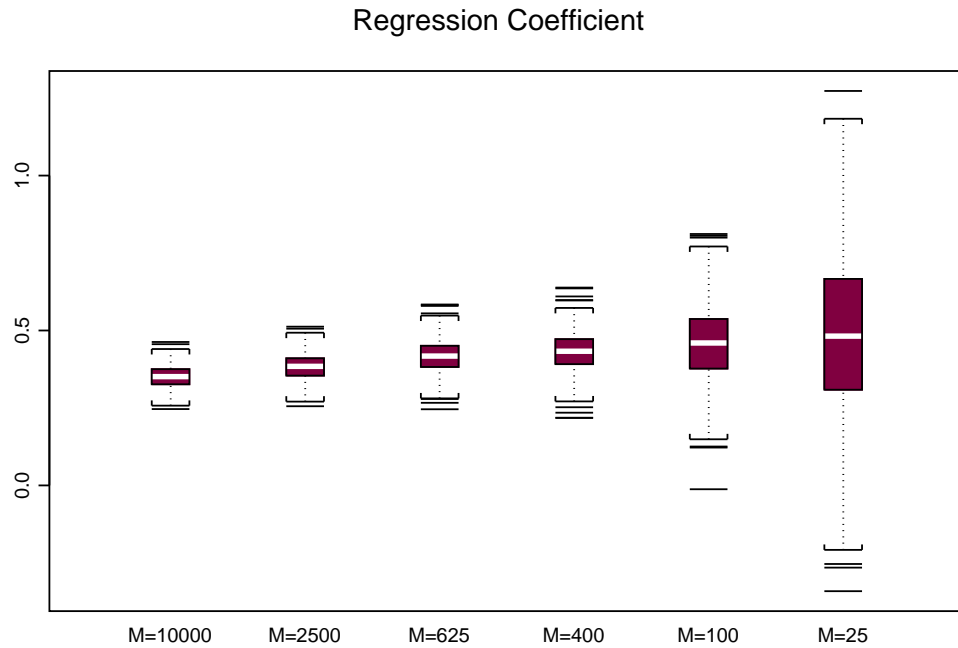


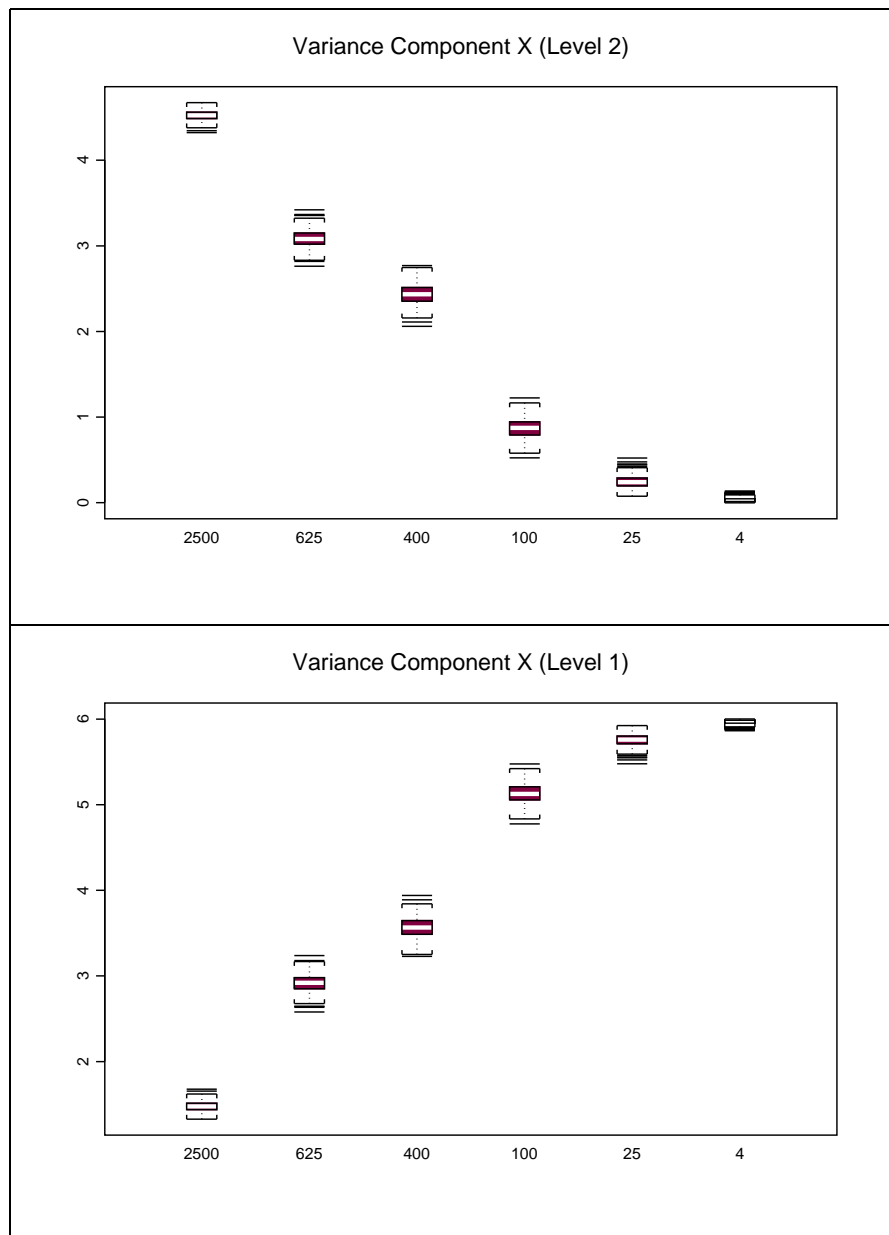
Figure 4.21: Regression Coefficient, X and Y both have high autocorrelation

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.3514	0.3516	0.2469	0.4629	0.0354
2500	0.3826	0.3846	0.2559	0.5124	0.0409
625	0.4169	0.4179	0.2461	0.5839	0.0532
400	0.4307	0.4333	0.2184	0.6387	0.0626
100	0.4600	0.4598	-0.0119	0.8120	0.1241
25	0.4795	0.4823	-0.3411	1.2732	0.2538
4	0.4880	0.5456	-6.5081	7.1885	1.1234

Table 4.45: Description of Regression Coefficient, X and Y both have high autocorrelation

then on. The mean of the level 1 variance component increases with aggregation. The distribution of the variance components of Y is not shown since the behavior is similar to the behavior of the distributions of variable X.

Figure 4.23 displays the distribution of the intra-area correlation of variable X. The mean (or median) of level 2 variance component decrease with aggregation. The mean intra-area correlation at different levels of aggregation are larger than the corresponding results from data set 2.



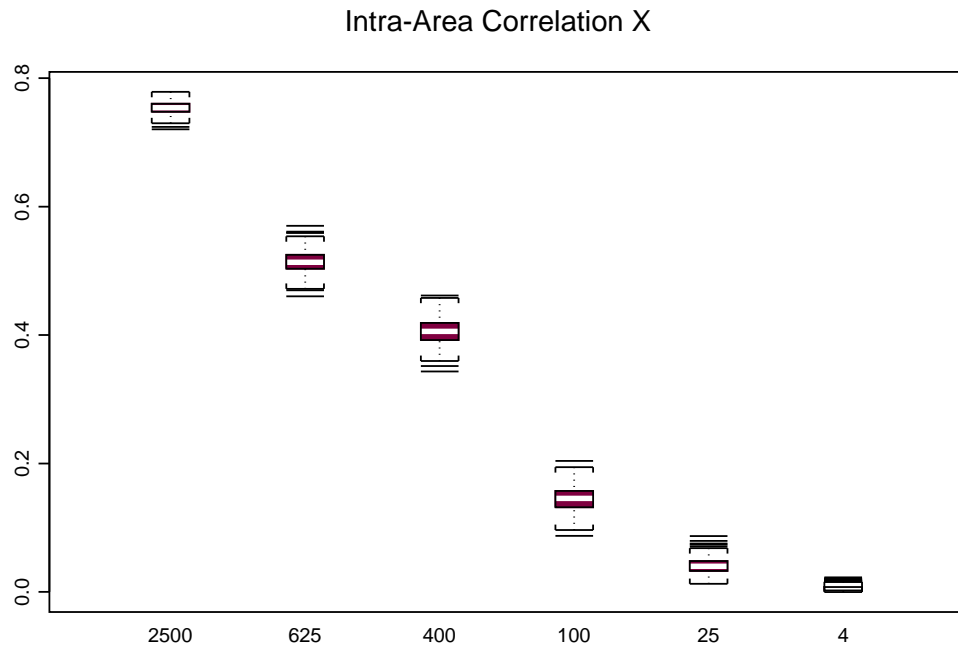
**Figure 4.22: Variance Components of X, X have high autocorrelation**

Figure 4.24 shows the distribution of intra-area cross-correlation. Similar to the results of the previous data set, the mean decreases with aggregation, although the values are greater than the previous two data sets.

Figure 4.25 shows the distribution of the estimated level 2 and level 1 *pure correlations*. Note that the figure shows only aggregation up to 25 zones, this is because when the data are aggregated to 4 zones, there are some values that are not a characteristic

Level 2 Variance Component X	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	4.5230	4.5233	4.3229	4.6726	0.0571
625	3.0847	3.0799	2.7627	3.4215	0.0974
400	2.4320	2.4341	2.0605	2.7705	0.1130
100	0.8705	0.8742	0.5234	1.2244	0.1125
25	0.2442	0.2386	0.0749	0.5219	0.0694
4	0.0344	0.0280	-0.0015	0.1351	0.0269
Level 1 Variance Component X	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	1.4770	1.4767	1.3274	1.6771	0.0571
625	2.9153	2.9201	2.5785	3.2373	0.0974
400	3.5681	3.5660	3.2295	3.9395	0.1130
100	5.1295	5.1258	4.7756	5.4767	0.1125
25	5.7558	5.7614	5.4781	5.9251	0.0694
4	5.9656	5.9720	5.8649	6.0015	0.0269

**Table 4.46: Description of the Level 2 and Level 1 Variance Component X, X have high autocorrelation**



**Figure 4.23: Intra-Area correlation of X, X have high autocorrelation**

of a correlation coefficient. In this particular case there are values that exceed 1 and less than -1. This phenomenon will be investigated later. The increase of the

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.7538	0.7539	0.7205	0.7788	0.0095
625	0.5141	0.5133	0.4604	0.5702	0.0162
400	0.4053	0.4057	0.3434	0.4617	0.0188
100	0.1451	0.1457	0.0872	0.2041	0.0188
25	0.0407	0.0398	0.0125	0.0870	0.0116
4	0.0057	0.0047	-0.0003	0.0225	0.0045

Table 4.47: Description of Intra-Area Correlation of X, X have high autocorrelation

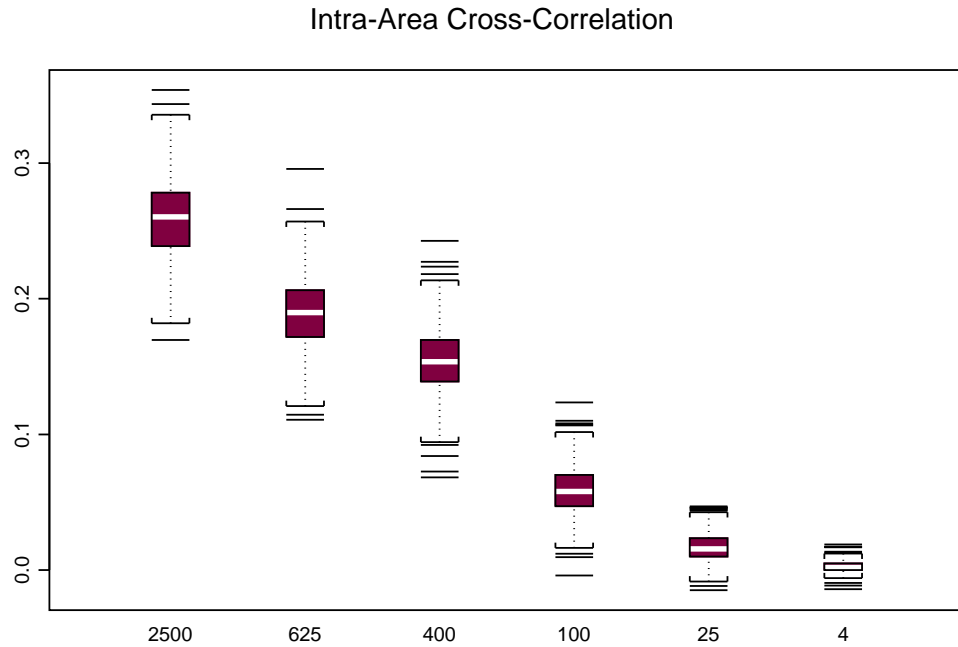
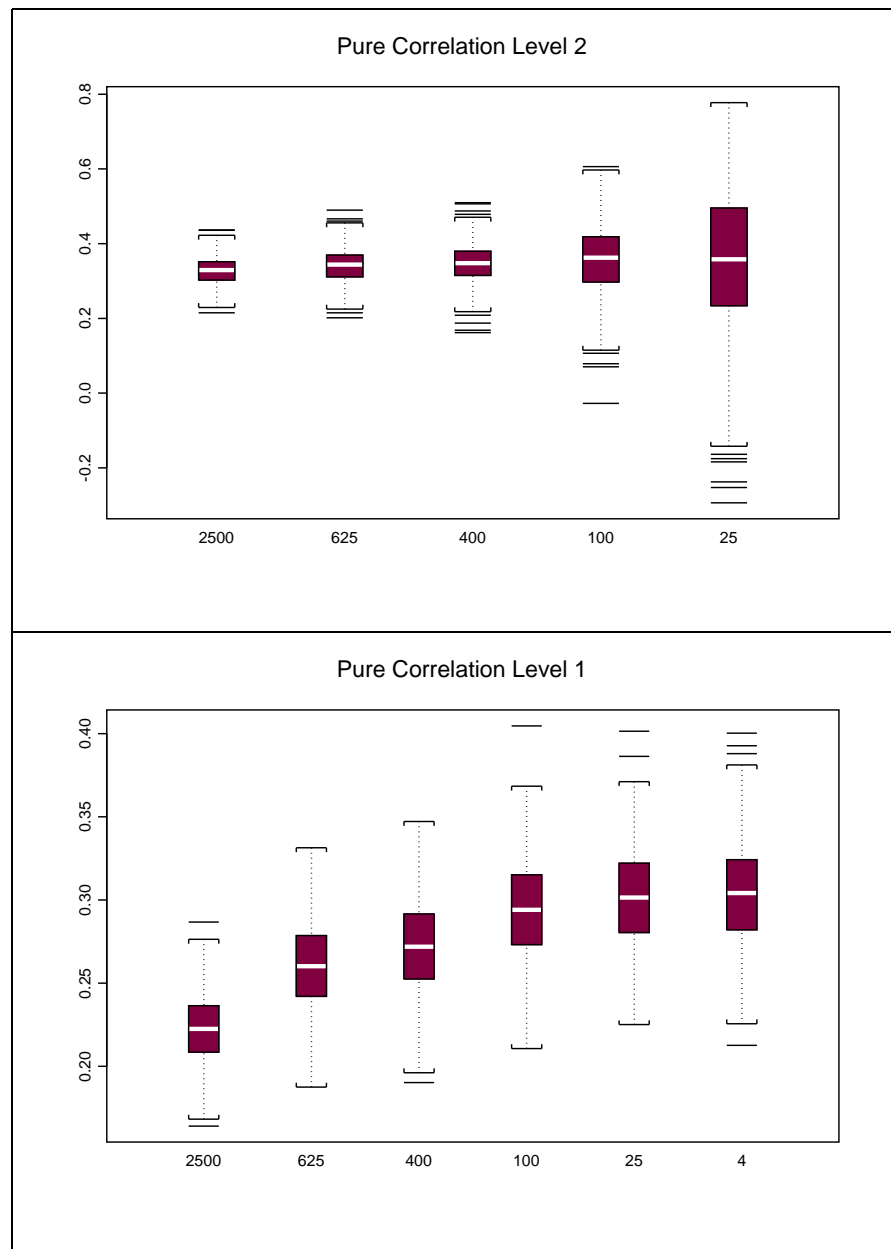


Figure 4.24: Intra-Area Cross-Correlation, X and Y both have high autocorrelation

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2589	0.2604	0.1697	0.3539	0.0289
625	0.1895	0.1898	0.1108	0.2958	0.0259
400	0.1541	0.1536	0.0684	0.2427	0.0243
100	0.0587	0.0580	-0.0040	0.1235	0.0180
25	0.0170	0.0157	-0.0148	0.0470	0.0104
4	0.0025	0.0020	-0.0141	0.0190	0.0041

Table 4.48: Description of Intra-Area Cross-Correlation, X and Y both have high autocorrelation





**Figure 4.25: Pure Correlation, X and Y both have high autocorrelation**

mean level 2 *pure correlation* is very slow and the values are near to the initial correlation coefficient of 0.3. The variability of the level 1 *pure correlation* for each level of aggregation is very similar to the results from data set 2 but slightly larger. Level 1 pure correlations also have results similar to the results from data set 2 and again are a bit larger. The variability of the level 1 *pure correlation* start with a smaller value and stabilizes when the level of aggregation reaches 400 groups and

stay constant up to the last level of aggregation.

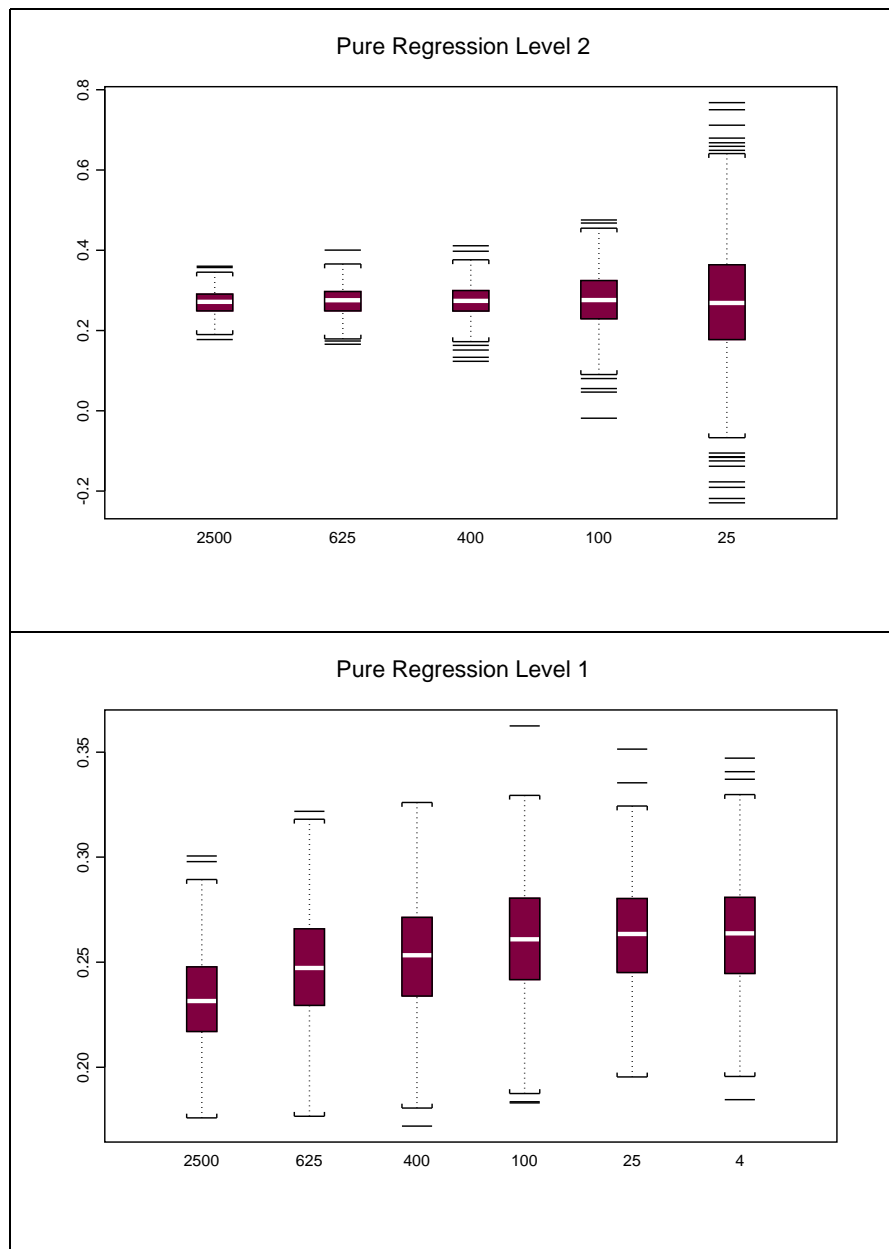
Level 2 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.3271	0.3295	0.2148	0.4369	0.0357
625	0.3412	0.3439	0.2018	0.4897	0.0441
400	0.3467	0.3478	0.1618	0.5098	0.0507
100	0.3574	0.3624	-0.0275	0.6061	0.0937
25	0.3580	0.3583	-0.2936	0.7775	0.1855
4	0.3421	0.5040	-1.4352	2.4131	0.6084
Level 1 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2224	0.2225	0.1641	0.2867	0.0206
625	0.2602	0.2602	0.1876	0.3314	0.0258
400	0.2719	0.2719	0.1902	0.3472	0.0280
100	0.2941	0.2941	0.2107	0.4046	0.0304
25	0.3015	0.3014	0.2251	0.4014	0.0306
4	0.3038	0.3042	0.2126	0.4002	0.0306

**Table 4.49: Description of the Level 2 and Level 1 Pure Correlation, X and Y both have high autocorrelation**

Figure 4.26 displays the distributions of level 2 and level 1 *pure regressions*. The mean of level 2 *pure regression* seems to be not affected by aggregation and the values are very near but smaller than the initial regression coefficient. The standard deviation increase with aggregation. The level 1 *pure regression* increase with aggregation but this time in a very slow manner, the standard deviation of the values seems to be constant.

#### 4.2.4 Discussion of Experiment 1

The mean is not affected by aggregation in both the weighted and unweighted analysis. The unweighted variance always decreases with aggregation regardless of the level of autocorrelation. However, the decrease depends on the level of autocorrelation as measured using the Moran's I statistic. The decrease for the variables with high positive autocorrelation seems to be small compared with variables with lower positive autocorrelation. The reason is that the variable with high positive autocorrelation have neighboring values that are likely to be similar and when they are



**Figure 4.26: Pure Regression, X and Y both have high autocorrelation**

aggregated, a relatively smaller variation is lost compared with a lower positively autocorrelated variable. Reynolds (1998) stated that when a variable is spatially located, the variance can be partitioned into sums of variances within various subregions and the variance of the average values of all subregions. He further stated that the process of aggregation removes the sum of the variances within subregions so that a variable with positive autocorrelation will have on the average smaller

Level 2 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2699	0.2717	0.1776	0.36016	0.0295
625	0.2738	0.2756	0.1656	0.4003	0.0356
400	0.2741	0.2740	0.1232	0.4111	0.0406
100	0.2750	0.2757	-0.0187	0.4756	0.0739
25	0.2718	0.2685	-0.2295	0.7680	0.1528
4	0.2525	0.2962	-14.841	13.8764	1.2087
Level 1 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2322	0.2315	0.1759	0.3006	0.0225
625	0.2482	0.2472	0.1767	0.3218	0.0257
400	0.2535	0.2533	0.1720	0.3261	0.0269
100	0.2608	0.2609	0.1831	0.3625	0.0277
25	0.2629	0.2634	0.1954	0.3514	0.0269
4	0.2634	0.2638	0.1846	0.3471	0.0265

**Table 4.50: Description of the Level 2 and Level 1 Pure Regression, X and Y both have high autocorrelation**

variance within each subregion so, less variance is lost. Looking at a particular result from the experiments above, Table 4.51 shows three variables with different levels of autocorrelation as measured using the Moran's I with queen's connectivity matrix and their corresponding individual level variances. It also shows the variance when the individual level data are aggregated into smaller number of zones. Looking at the values, we can see that the more autocorrelated the variable, the lesser the change in the variance when the data are aggregated compared with the variables with lower autocorrelation.

Variable	X1	X2	X3
Moran's I (Individual)	0.1222	0.4549	0.7138
<b>Variance (Individual)</b>	<b>6.000</b>	<b>6.000</b>	<b>6.000</b>
Variance (2500 Zones)	2.1361	3.7314	4.8461
Variance (625 Zones)	0.8065	2.0727	3.1691
Variance (400 Zones)	0.6126	1.5626	2.5333
Variance (100 Zones)	0.1516	0.5814	0.9373
Variance (25 Zones)	0.0462	0.1600	0.2598
Variance (4 Zones)	0.0067	0.0214	0.0360

**Table 4.51: Moran's I and Unweighted Variance**

Note that from the simple multi-level model given by 5.1 we have,  $V(\bar{X}_g) = \frac{\sigma_{XX}}{N_g} \left[ 1 + (N_g - 1)\Lambda_{XX}^{(2)} \right]$  and the unweighted variance reflects the variances of these means. In subsection 3.1.8 we saw that larger Moran's I will result in larger  $\Lambda^{(2)}$  leading to the larger variances. Also, as zones get bigger  $N_g$  increases and  $\Lambda^{(2)}$  decreases, so the variance goes down and thus explains the behavior of the decrease of the variance.

A connection between Moran's I and  $\Lambda^{(2)}$  with appropriate choice of connectivity was derived in section 3.1.8 and will be examined more closely in Chapter 6.

The correlation coefficient is affected by the level of aggregation and tends to increase as the number of zones decrease regardless of the level of autocorrelation. The mean correlation increases with aggregation. The increase however, depends on the level of autocorrelation. Data set 1 consist of variables X and Y that are both *low autocorrelated* although variable Y has a little larger autocorrelation. For data set 2, variables X and Y have medium autocorrelation and the increasing trend of the correlation is observed but this time the increase is not as fast as the increase that is observed from data set 1. Data set 3, the variables have *high autocorrelation* displayed patterns similar to data set 2 but the increase is slower. Table 4.52 shows the theoretical values of the mean of the correlation and the corresponding standard deviation computed using Equations (4.2) and (4.3) assuming no autocorrelation. For the different levels of aggregation, the number of groups were substituted into the equations. The results shows that the expected mean correlation depends on the initial correlation at the individual level. Looking at the results, even the reduction of the number of groups does not have much effect except when the data are aggregated to 4 groups, in which case there is a small but sudden increase of the mean correlation.

Table 4.53 shows the results computed from the simulated data. Compared with the corresponding theoretical results, the mean of the correlations is affected by the level of autocorrelation of the variables. The percentage loss of variance of the variables and the corresponding percentage loss of covariance cause the correlation to increase with aggregation. The increase depends on the difference between the per-

	<b>Data Set 1</b>	<b>Data Set 2</b>	<b>Data Set 3</b>
Correlation	Mean ( <i>Std. Dev.</i> )	Mean ( <i>Std. Dev.</i> )	Mean( <i>Std. Dev.</i> )
Individual	0.2989 ( <i>0.0091</i> )	0.3154 ( <i>0.0090</i> )	0.3043 ( <i>0.0090</i> )
2500	0.2990 ( <i>0.0182</i> )	0.3155 ( <i>0.0180</i> )	0.3044 ( <i>0.0182</i> )
625	0.2991 ( <i>0.0365</i> )	0.3156 ( <i>0.0361</i> )	0.3045 ( <i>0.0363</i> )
400	0.2992 ( <i>0.0456</i> )	0.3158 ( <i>0.0451</i> )	0.3046 ( <i>0.0455</i> )
100	0.3004 ( <i>0.0918</i> )	0.3169 ( <i>0.0908</i> )	0.3057 ( <i>0.0914</i> )
25	0.3059 ( <i>0.1879</i> )	0.3227 ( <i>0.1860</i> )	0.3098 ( <i>0.1873</i> )
4	0.3954 ( <i>0.5710</i> )	0.4134 ( <i>0.5683</i> )	0.3388 ( <i>0.5701</i> )

**Table 4.52: Theoretical Expected Value and Standard Deviation of Correlation of Data sets 1, 2, and 3**

centage loss of variance and the percentage loss of covariance. When both variables have *low autocorrelation*, the difference between percentage loss of variance and the covariance is larger compare with the other cases, that is, when the variables have *medium* and *high autocorrelation*. These will contribute to the increase (or decrease) in the correlation.

	<b>Data Set 1</b>	<b>Data Set 2</b>	<b>Data Set 3</b>
Correlation	Mean ( <i>Stan Dev.</i> )	Mean ( <i>Stan Dev.</i> )	Mean( <i>Stan Dev.</i> )
Individual	0.2929 ( <i>0.0107</i> )	0.3154 ( <i>0.0208</i> )	0.3043 ( <i>0.0307</i> )
2500	0.4322 ( <i>0.0185</i> )	0.3686 ( <i>0.0268</i> )	0.3203 ( <i>0.0342</i> )
625	0.5413 ( <i>0.0288</i> )	0.4040 ( <i>0.0369</i> )	0.3372 ( <i>0.0425</i> )
400	0.5544 ( <i>0.0344</i> )	0.4148 ( <i>0.0439</i> )	0.3429 ( <i>0.0486</i> )
100	0.5746 ( <i>0.0683</i> )	0.4291 ( <i>0.0834</i> )	0.3542 ( <i>0.0893</i> )
25	0.5790 ( <i>0.1427</i> )	0.4304 ( <i>0.1684</i> )	0.3553 ( <i>0.1762</i> )
4	0.5279 ( <i>0.4756</i> )	0.4013 ( <i>0.5207</i> )	0.3306 ( <i>0.5392</i> )

**Table 4.53: Summary of Correlation of Data sets 1, 2, and 3**

Except for the cases of the individual level and 2500 groups with medium or high autocorrelation, the theoretical standard deviation provides a reasonable indication of the actual standard deviation.

The mean of the correlations and their standard deviation increase as the scale increases, except when there are 4 zones in the case of the mean. The rate of increase of the mean correlation is reduced as the autocorrelation increases.

The regression coefficient is also affected by aggregation. The effects are similar to the correlation. Regression will be dealt with later in the chapter.

Table 4.54 is the summary of the level 2 and level 1 pure correlations for the three data sets. The mean correlation of data set 1 (both variables have low autocorrelation) is not affected by aggregation except when the data are aggregated into 4 zones, in which case there is a big jump of the average correlation but these values are much larger than the initial average correlation of around 0.3. When both variables have *medium autocorrelation*, the level 2 *pure correlation* increases slowly with aggregation but starts with a little larger correlation when the data are aggregated into 2500 zones. When both variables have *high autocorrelation*, the level 2 pure correlation increases slowly with aggregation and the mean correlation when the data are aggregated into 2500 groups is near the initial individual level correlation. Level 1 pure correlations, in all cases considered, display a similar pattern; as the number of groups decreases, the mean correlation approaches the initial correlation.

Level 2 Pure Correlation	Data Set 1 Mean ( <i>Stan Dev.</i> )	Data Set 2 Mean ( <i>Stan Dev.</i> )	Data Set 3 Mean( <i>Stan Dev.</i> )
2500	0.7080 ( <i>0.0391</i> )	0.4012 ( <i>0.0308</i> )	0.3271 ( <i>0.0357</i> )
625	0.7093 ( <i>0.0415</i> )	0.4212 ( <i>0.0404</i> )	0.3412 ( <i>0.0441</i> )
400	0.7071 ( <i>0.0468</i> )	0.4308 ( <i>0.0480</i> )	0.3467 ( <i>0.0507</i> )
100	0.7036 ( <i>0.0903</i> )	0.4423 ( <i>0.0912</i> )	0.3574 ( <i>0.0937</i> )
25	0.7084 ( <i>0.2008</i> )	0.4425 ( <i>0.1853</i> )	0.3580 ( <i>0.1855</i> )
4	0.7871 ( <i>1.2305</i> )	0.4270 ( <i>0.6770</i> )	0.3421 ( <i>0.6084</i> )
Level 1 Pure Correlation	Data Set 1 Mean ( <i>Stan Dev.</i> )	Data Set 2 Mean ( <i>Stan Dev.</i> )	Data Set 3 Mean( <i>Stan Dev.</i> )
2500	0.2076 ( <i>0.0110</i> )	0.2079 ( <i>0.0163</i> )	0.2224 ( <i>0.0206</i> )
625	0.2465 ( <i>0.0103</i> )	0.2573 ( <i>0.0180</i> )	0.2602 ( <i>0.0258</i> )
400	0.2605 ( <i>0.0102</i> )	0.2735 ( <i>0.0189</i> )	0.2719 ( <i>0.0280</i> )
100	0.2834 ( <i>0.0105</i> )	0.3032 ( <i>0.0204</i> )	0.2941 ( <i>0.0304</i> )
25	0.2904 ( <i>0.0105</i> )	0.3119 ( <i>0.0206</i> )	0.3015 ( <i>0.0306</i> )
4	0.2925 ( <i>0.0106</i> )	0.3151 ( <i>0.0207</i> )	0.3038 ( <i>0.0306</i> )

Table 4.54: Summary of Pure Correlation of Data sets 1, 2, and 3

Table 4.55 shows the summary for pure regression coefficient.

Level 2 Pure Regression	Data Set 1 Mean ( <i>Stan Dev.</i> )	Data Set 2 Mean ( <i>Stan Dev.</i> )	Data Set 3 Mean( <i>Stan Dev.</i> )
2500	0.5018 ( <i>0.0336</i> )	0.3027 ( <i>0.0239</i> )	0.2699 ( <i>0.0295</i> )
625	0.5033 ( <i>0.0379</i> )	0.3086 ( <i>0.0310</i> )	0.2738 ( <i>0.0356</i> )
400	0.5009 ( <i>0.0428</i> )	0.3106 ( <i>0.0363</i> )	0.2741 ( <i>0.0406</i> )
100	0.5006 ( <i>0.0789</i> )	0.3122 ( <i>0.0686</i> )	0.2750 ( <i>0.0739</i> )
25	0.5010 ( <i>0.1772</i> )	0.3082 ( <i>0.1419</i> )	0.2718 ( <i>0.1528</i> )
4	0.5662 ( <i>4.9090</i> )	0.2827 ( <i>1.3818</i> )	0.2525 ( <i>1.2087</i> )
Level 1 Pure Regression	Data Set 1 Mean ( <i>Stan Dev.</i> )	Data Set 2 Mean ( <i>Stan Dev.</i> )	Data Set 3 Mean( <i>Stan Dev.</i> )
2500	0.1876 ( <i>0.0101</i> )	0.2172 ( <i>0.0175</i> )	0.2322 ( <i>0.0225</i> )
625	0.2184 ( <i>0.0092</i> )	0.2465 ( <i>0.0177</i> )	0.2482 ( <i>0.0257</i> )
400	0.2293 ( <i>0.0090</i> )	0.2546 ( <i>0.0180</i> )	0.2535 ( <i>0.0269</i> )
100	0.2466 ( <i>0.0091</i> )	0.2681 ( <i>0.0184</i> )	0.2608 ( <i>0.0277</i> )
25	0.2518 ( <i>0.0092</i> )	0.2719 ( <i>0.0181</i> )	0.2629 ( <i>0.0269</i> )
4	0.2534 ( <i>0.0092</i> )	0.2730 ( <i>0.0179</i> )	0.2634 ( <i>0.0265</i> )

Table 4.55: Summary of Pure Regression of Data sets 1, 2, and 3

Since level 2 pure correlation is estimated using

$$\hat{\rho}_{YX}^{(2)} = \frac{\hat{\Lambda}_{YX}^{(2)}}{\sqrt{\hat{\Lambda}_{YY}^{(2)} \hat{\Lambda}_{XX}^{(2)}}}$$

and the level 2 pure regression is estimated using

$$\hat{\beta}_{YX}^{(2)} = \frac{\hat{\Lambda}_{YX}^{(2)}}{\hat{\Lambda}_{XX}^{(2)}}$$

the way the correlations and cross correlations change relative to each other is a factor in determining how the pure coefficients change as we aggregate.

### 4.3 Experiment 2: Scale effects when both variables are not autocorrelated

This experiment is done to investigate the aggregation effect when the data are randomly spatially distributed, that is, the measure of spatial autocorrelation is



zero. In comparison with the other experiments, the data generation is done in such a way that both the variables have no spatial autocorrelation.

Two variables (X and Y) are generated using R, using the multivariate normal function. This case was considered by Steel et. al. (1996). To help compare it with the other experiments, the variables are generated such that the means are 0.005 and 10.0 for X and Y, respectively. The variance are 6 and 8, respectively for X and Y and the individual level correlation is 0.3.

Table 4.56 shows the Moran's I using RookCase, an add-on software for Excel, the proximity matrix used is "queen's move".

Level	$I_{XX}^l$	$I_{YY}^l$	$I_{YX}^l$
Individual	0.00135	0.0010	-0.0011
Z2500	-0.0159	-0.0025	-0.0084
Z625	-0.0221	-0.0191	0.0144
Z400	0.0098	0.0626	0.0360
Z100	-0.0124	0.0370	0.0279
Z25	-0.4136	0.0563	0.2397
Z4	-	-	-

**Table 4.56: Moran's I**

As before, one realization of the data generation is examined initially. Table 4.57 shows the unweighted variance and covariance for the two variables X and Y. The rate of decrease from the initial variance of the two variables is large compared with the case when the variables are autocorrelated.

	$\bar{X}$	$\bar{Y}$	$\tilde{S}_{XX}^{(l)}$	$\tilde{S}_{YY}^{(l)}$	$\tilde{S}_{YX}^{(l)}$
Individual	0.005	10.000	6.0000	8.0000	2.0648
Z2500	0.005	10.000	1.6009	2.0422	0.5616
Z625	0.005	10.000	0.3872	0.5091	0.1231
Z400	0.005	10.000	0.2210	0.3387	0.0918
Z100	0.005	10.000	0.0559	0.0895	0.0319
Z25	0.005	10.000	0.0180	0.0247	0.0115
Z4	0.005	10.000	0.0031	0.0077	0.0042

**Table 4.57: Unweighted Variance and Covariance, X and Y both are not autocorrelated**

Table 4.58 shows the weighted variance and covariance for the two variables X and Y. The values appear to be approximately constant until the number of groups

is 25 and 4.

	$\bar{X}$	$\bar{Y}$	$S_{XX}^{(l)}$	$S_{YY}^{(l)}$	$S_{XY}^{(l)}$
Individual	0.005	10.000	6.0000	8.0000	2.0259
Z2500	0.005	10.000	6.4037	8.1688	2.2464
Z625	0.005	10.000	6.1945	8.1449	1.9696
Z400	0.005	10.000	5.5254	8.4668	2.2951
Z100	0.005	10.000	5.5947	8.9459	3.1864
Z25	0.005	10.000	7.2012	9.8850	4.5833
Z4	0.005	10.000	7.7141	19.3487	10.5313

**Table 4.58: Weighted Variance and Covariance, X and Y both are not autocorrelated**

Table 4.59 shows the correlation and regression coefficients at different scales. An increase of the correlation coefficients is observed. Steel et al (1996) suggested that there should be no aggregation effect on this statistics in this case. The reason for the increase may be that it was just one randomly selected realization of the data set generation. The regression coefficients also displayed the increasing pattern. Results for repeated generation of the population are given later in this section.

	Correlation Coefficient	Regression Coefficient
Individual Data	0.2980	0.3441
Number of Zones		
2500	0.3106	0.3508
625	0.2773	0.3180
400	0.3356	0.4154
100	0.4504	0.5695
25	0.5432	0.6365
4	0.8620	1.3652

**Table 4.59: Correlation and regression coefficients at different scales, X and Y both are not autocorrelated**

Tables 4.60 shows the estimated variance components and intra-area correlations of the variable X at different levels of aggregation. Looking at the level 2 variance components at different levels of aggregation, it can be noticed that some values are negative. Recall that the initial variance of variable X is 6.0, when the estimated level 2 variance component is negative, the level 1 variance component will be more than 6.0 and thus, the resulting intra-area correlation is negative. The moment approach

gives unbiased estimates, but allows negative estimates of variance components. Looking at the values of  $\hat{\delta}_{XX}$  they are all close to zero which is consistent with Moran's I being approximately equal to zero.

Level 1	Level 2	$\hat{\Lambda}_{XX}^{(2)}$	$\hat{\Lambda}_{XX}^{(1)}$	$\hat{\delta}_{XX}$
Individual	Z2500	0.1345	5.8650	0.0224
	Z625	0.0129	5.9871	0.0022
	Z400	-0.0197	6.0197	-0.0033
	Z100	-0.0041	6.0041	-0.0007
	Z25	0.0029	5.9971	0.0005
	Z4	0.0005	5.9995	0.0001

**Table 4.60: Intra-Area correlations and variance components of X, X not autocorrelated**

Table 4.61 displays the estimated variance components of variable Y. In this particular realization of the data set generator, no negative level 2 variance component is observed but this is not the general case. The level 2 variance component is small in relation to the level 1 variance component resulting in small intra-area correlation.

Level 1	Level 2	$\hat{\Lambda}_{YY}^{(2)}$	$\hat{\Lambda}_{YY}^{(1)}$	$\hat{\delta}_{YY}$
Individual	Z2500	0.0562	7.9438	0.0070
	Z625	0.0096	7.9904	0.0012
	Z400	0.0194	7.9806	0.0024
	Z100	0.0095	7.9905	0.0012
	Z25	0.0045	7.9955	0.0006
	Z4	0.0036	7.9964	0.0005

**Table 4.61: Intra-Area correlations and variance components of Y, Y both not autocorrelation**

Tables 4.62 shows the Intra-Area Cross Correlation. It can be noted that level 1 covariance component seems to be not affected by aggregation.

Table 4.63 displays level 2 and level 1 *pure correlation* components. It can be seen that the level 2 *pure correlation* is severely affected by aggregation, from negative value to positive value, values more than 1 and worse, with entries in the table labelled NA. The last case happened because either level 2 variance component of X or Y is negative. This phenomenon will be examined later. The level 1 *pure correlation*

Level 1	Level 2	$\hat{\Lambda}_{YX}^{(2)}$	$\hat{\Lambda}_{YX}^{(1)}$	$\hat{\delta}_{YX}$
Individual	Z2500	0.0605	2.0043	0.0087
	Z625	-0.0063	2.0710	-0.0009
	Z400	0.0096	2.0552	0.0014
	Z100	0.0112	2.0536	0.0016
	Z25	0.0061	2.0587	0.0009
	Z4	0.0027	2.0621	0.0004

**Table 4.62: Intra-Area Cross-Correlation, X and Y both are not autocorrelated**

seems to be not affected by aggregation. Notice that column 3 of Table 4.63 have rows without an entry, this is because the level 2 variance components on these two levels of aggregation is negative.

Recall that pure correlation is computed using

$$\hat{\rho}^{(l)} = \frac{\hat{\Lambda}_{YX}^{(l)}}{\sqrt{\hat{\Lambda}_{YY}^{(l)} \hat{\Lambda}_{XX}^{(l)}}} \quad (4.7)$$

so that level 2 pure correlation is computed using

$$\hat{\rho}^{(2)} = \frac{\hat{\Lambda}_{YX}^{(2)}}{\sqrt{\hat{\Lambda}_{YY}^{(2)} \hat{\Lambda}_{XX}^{(2)}}}. \quad (4.8)$$

The level 2 variance components when aggregated into 400 and 100 zones have negative values. This will result in an operation of taking the square-root of a negative number.

Level 1	Level 2	$\hat{\rho}_{YX}^{(2)}$	$\hat{\rho}_{YX}^{(1)}$
Individual	Z2500	0.6957	0.2936
	Z625	-0.5672	0.2994
	Z400	NA	0.2965
	Z100	NA	0.2965
	Z25	1.6737	0.2973
	Z4	1.9200	0.2977

**Table 4.63: Pure correlations at two levels, X and Y both are not autocorrelated**

Table 4.64 shows the level 2 and level 1 *pure regression*. The level 2 *pure regression* does not display a predictable aggregation effect. Level 1 *pure regression* seems to be not affected by aggregation but the values are lower than the initial regression coefficient.

Level 1	Level 2	$\hat{b}_{YX}^{(2)}$	$\hat{b}_{YX}^{(1)}$
Individual	Z2500	1.0759	0.2523
	Z625	-0.6572	0.2592
	Z400	0.4934	0.2575
	Z100	1.8580	0.2570
	Z25	1.5395	0.2575
	Z4	1.3361	0.2579

Table 4.64: Pure Regressions at two levels, X and Y both are not autocorrelated

### Analysis of distributions of statistics when both variables are not autocorrelated

The data generation is then repeated 500 times to investigate the distributions.

Figure 4.27 shows the distribution of the Pearson correlation at the individual level and the different levels of aggregation. The mean of the correlation coefficients is not affected by aggregation as predicted by Steel and Holt (1996). The standard deviation of the Pearson correlation increase with aggregation.

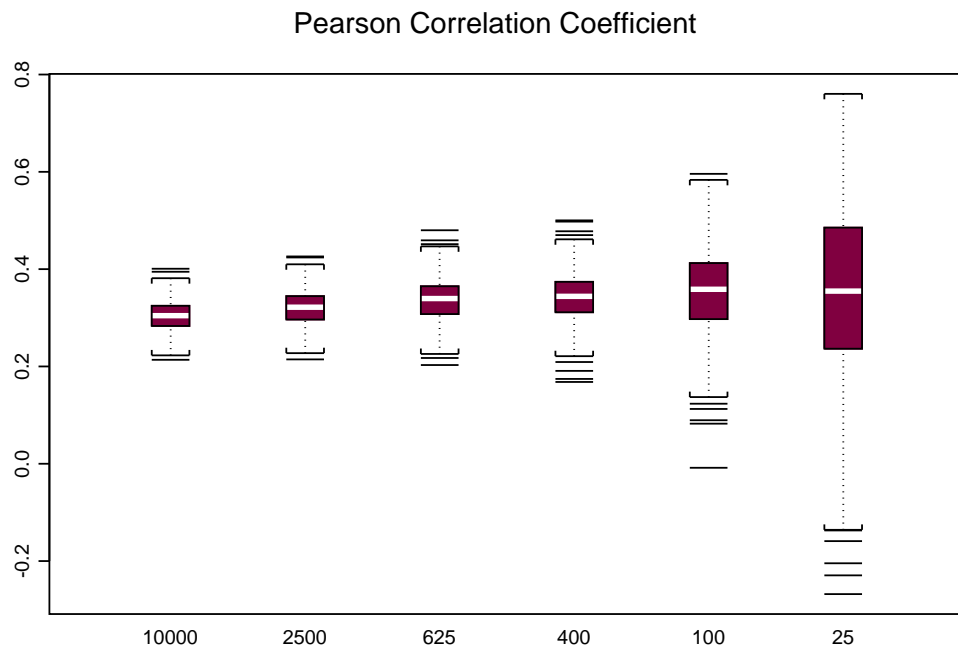


Figure 4.27: Pearson Correlation, X and Y both are not autocorrelated

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	0.2996	0.2993	0.2722	0.3321	0.0095
2500	0.2986	0.2989	0.2360	0.3558	0.0203
625	0.2970	0.2980	0.1750	0.4138	0.0378
400	0.2996	0.2978	0.1648	0.4356	0.0458
100	0.2994	0.3019	0.0669	0.5354	0.0863
25	0.2850	0.2980	-0.2910	0.7090	0.1780
4	0.2600	0.3630	-0.9980	0.9990	0.5351

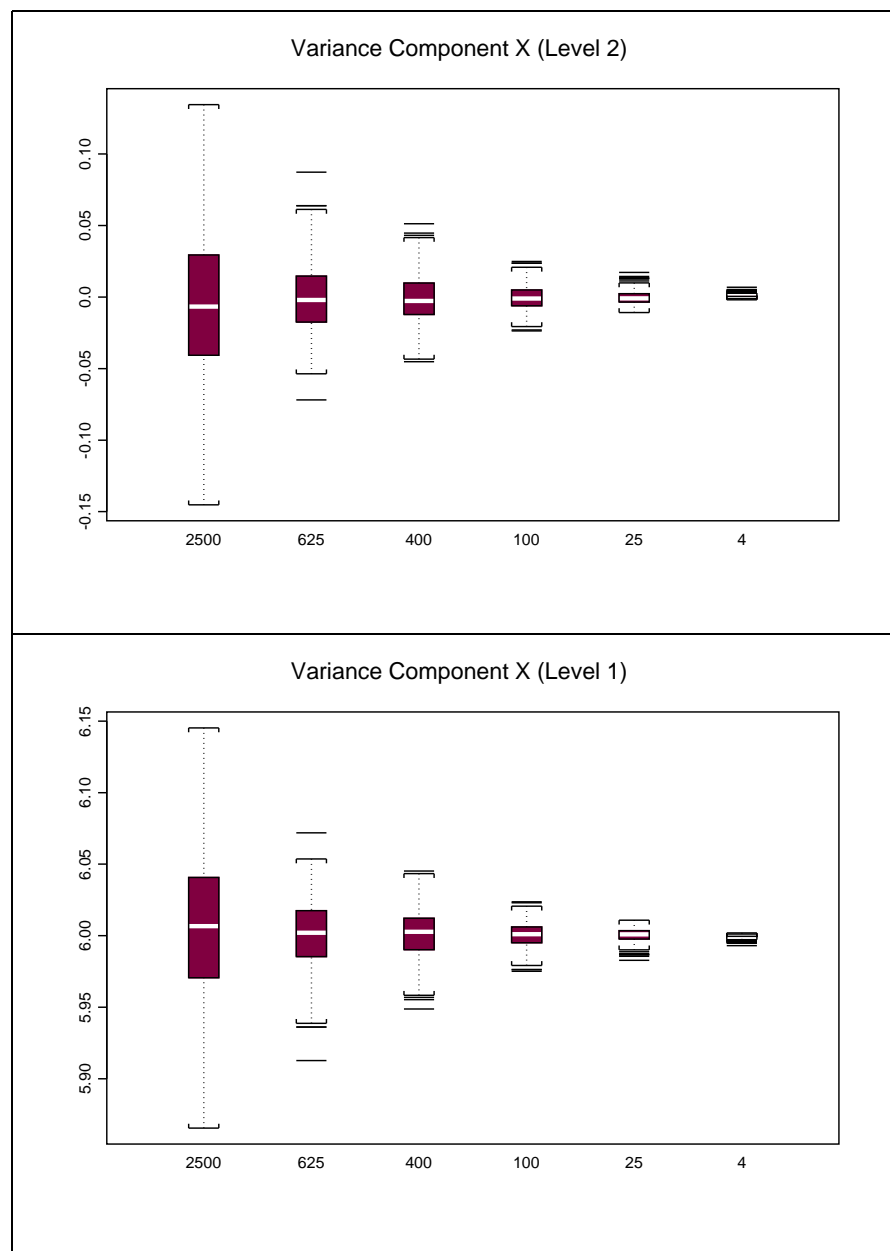
**Table 4.65: Description of Pearson Correlation, X and Y both are not autocorrelated**

Steel and Holt (1996) derived some theoretical results on group-level analysis procedures for random aggregation. One of the results is that the expectation of the group-level correlation is  $\rho \left(1 - \frac{1-\rho^2}{M}\right)$  and the variance is  $\frac{(1-\rho^2)^2}{M-1} \left(1 + \frac{11\rho^2}{2M}\right)$  where M is the number of groups. Table 4.66 shows the theoretical values of the mean, standard deviation, and 95% interval of the group-level correlation for the different values of M. The results are very similar to those observed in table 4.67.

Description	Mean	Lower Limit	Upper Limit	Std. Dev.
Individual	0.2999	0.2899	0.3418	0.0091
2500	0.2999	0.2771	0.3639	0.0181
625	0.2997	0.2349	0.4092	0.0364
400	0.2993	0.2140	0.4326	0.0456
100	0.2973	0.1094	0.5587	0.0917
25	0.2891	-0.1151	0.7241	0.1876

**Table 4.66: Theoretical Mean, Standard deviation, Lower and Upper Limits of group-level correlation**

Figure 4.28 shows the distribution of the estimated variance components of variable X. Looking at the figure, we can notice that the mean or median of the level 2 variance components seems to be not affected by aggregation and they are all slightly less than zero. The moment based estimates that are being used are unbiased and since the level 2 variance is zero, they will give negative estimates. The level 1 variance is also estimated unbiasedly. The standard deviations of these values decrease with aggregation. Looking at the right side of the figure, the distribution of the level 1 variance components displays a similar pattern. The mean (or median) seems to be not affected by aggregation but some of the values are more



**Figure 4.28: Variance Components of X, X not autocorrelated**

than the initial variance because of the negative level 2 variance component.

Table 4.67 summarizes the distribution of the level 2 and level 1 variance component of X. Level 2 variance component have negative values. As mentioned earlier, this is because of the method used in the computation of the variance components.

Figure 4.29 shows the estimated intra-area correlation of X where it can be noted that the standard deviation of the values decreases with aggregation. The mean (or

Level 2 Variance Component X	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	-0.0050	-0.0066	-0.1452	0.1345	0.0505
625	-0.0006	-0.0021	-0.0719	0.0873	0.0225
400	-0.0014	-0.0027	-0.0452	0.0512	0.0169
100	-0.0005	-0.0011	-0.0237	0.0248	0.0083
25	-0.0004	-0.0009	-0.0108	0.0172	0.0042
4	-0.00009	-0.0004	-0.0019	0.0069	0.0015
Level 1 Variance Component X	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	6.0050	6.0066	5.8655	6.1452	0.0505
625	6.0006	6.0021	5.9128	6.0719	0.0225
400	6.0014	6.0027	5.9488	6.0452	0.0169
100	6.0005	6.0011	5.9752	6.0237	0.0083
25	6.0004	6.0009	5.9828	6.0108	0.0042
4	6.0000	6.0004	5.9931	6.0019	0.0015

**Table 4.67: Description of the Level 2 and Level 1 Variance Component of X, X not autocorrelated**

median) of the values seems to be not affected by aggregation and the values are almost equal to zero but negative.

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	-0.00083	-0.00111	-0.02421	0.02242	0.00842
625	-0.00009	-0.00034	-0.01198	0.01454	0.00376
400	-0.00024	-0.00045	-0.00754	0.00854	0.00282
100	-0.00008	-0.00018	-0.00400	0.00410	0.00140
25	-0.00007	-0.00016	-0.00180	0.00290	0.00070
4	-0.00002	-0.00007	-0.00032	0.00120	0.00024

**Table 4.68: Description of Intra-Area Correlation of X, X not autocorrelated**

Figure 4.30 show the estimated intra-area correlation of Y, which is similar to intra-area correlation of X. The standard deviation of the values decrease with aggregation.

Figure 4.31 displays the distributions of the estimated level 2 and level 1 *pure correlations*. The boxplot of the level 2 *pure correlations* ended up with a very different appearance because there are many entries which end up with NA's because the estimated variance component at this level is negative. Actually at every aggregation level,



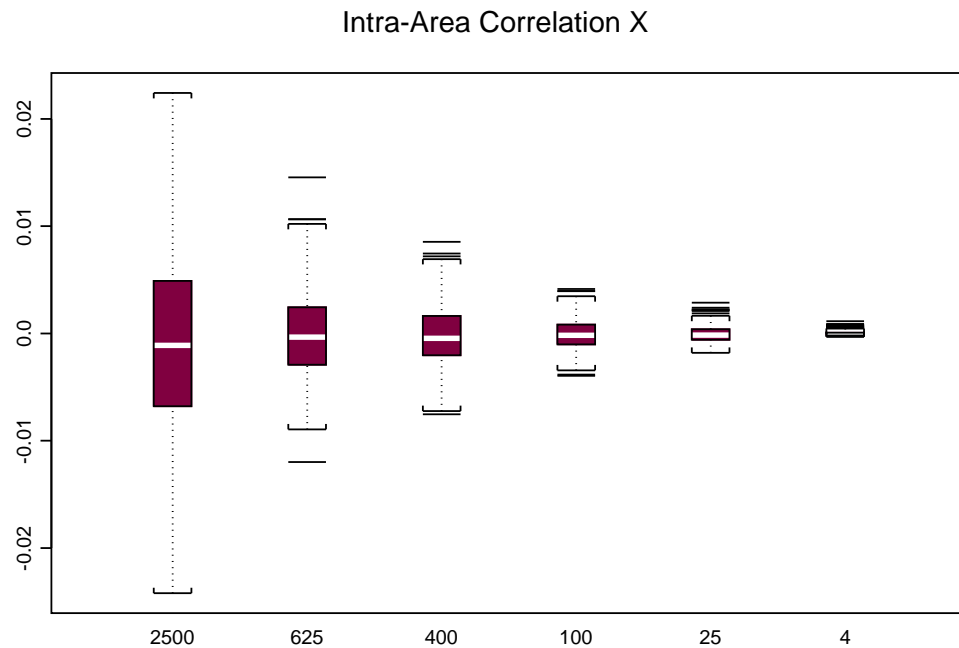
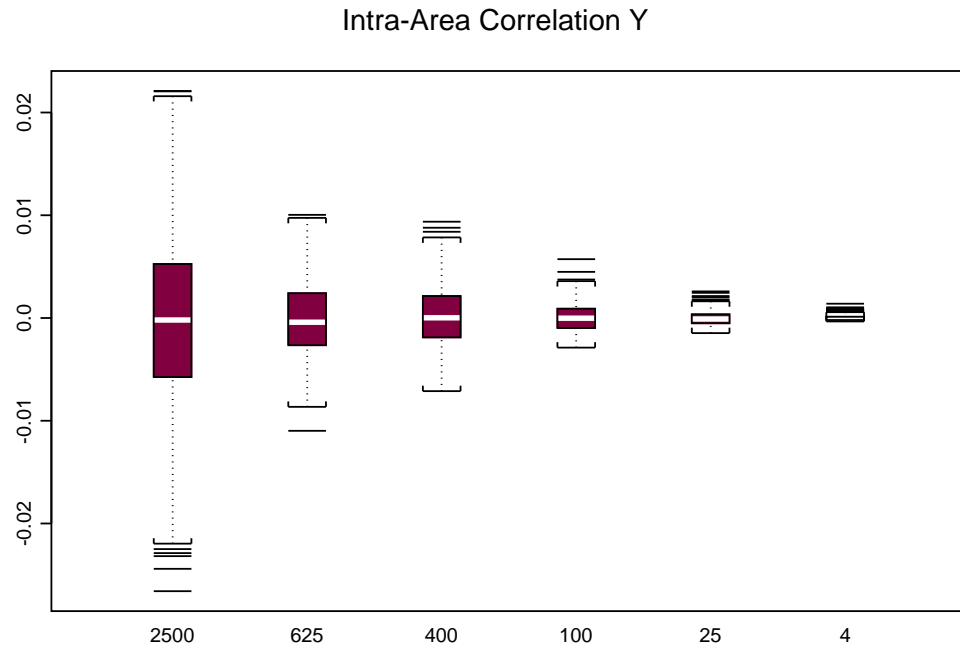


Figure 4.29: Intra-Area Correlation X, X not autocorrelated

Description	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	-0.00024	-0.00020	-0.02658	0.02211	0.00852
625	-0.00016	-0.00040	-0.01010	0.01005	0.00359
400	0.00013	0.00003	-0.00710	0.00094	0.00290
100	0.00004	-0.000001	-0.00290	0.00570	0.00140
25	-0.00003	-0.00010	-0.00150	0.00260	0.00066
4	0.000004	-0.00007	-0.00032	0.00140	0.00026

Table 4.69: Description of Intra-Area Correlation of Y, Y not autocorrelated

the number of NA's is almost 50 percent because of the level 2 variance component is estimated unbiasedly and has mean 0, so approximately half will give negative estimated level 2 variance components. These values can be set to zero. These results suggest that estimated level 2 coefficients will be unstable when there is no autocorrelation present. This is reflected in the relatively high standard deviations in Table 4.70. In practice it would be important to examine confidence intervals on the level 2 variance components before attempting to calculate level 2 pure correlation or regression coefficients. Methods for testing whether variance components



**Figure 4.30: Intra-Area Correlation Y, Y not autocorrelated**

are zero are given in Snijders and Bosker (1999). The mean (or median) of the level 2 *pure correlation* is not affected by aggregation. The estimated level 2 pure correlation coefficients are not well behaved when autocorrelation is zero.

Figure 4.32 shows the distributions of level 2 and level 1 *pure regression* coefficients at different levels of aggregation. The mean of the level 2 *pure regression* is affected by aggregation. The mean of the level 1 *pure regression* seems to be not affected by aggregation and the standard deviation of values seems to be constant except when there are 2500 zones. Although, the mean (or median) is not affected by aggregation, the value is smaller than the mean (or median) of the initial regression coefficient. Similar to the level 2 pure correlation, the estimated level 2 pure regression coefficients are not well behaved when autocorrelation is zero.

When the variables are both not spatially autocorrelated, the level 1 *pure correlations* and the level 1 *pure regression* coefficients seem to be not affected by aggregation.

The level 1 intra-area cross-correlations have values almost equal to the correlation coefficient at the individual level. This is because there is not much change of

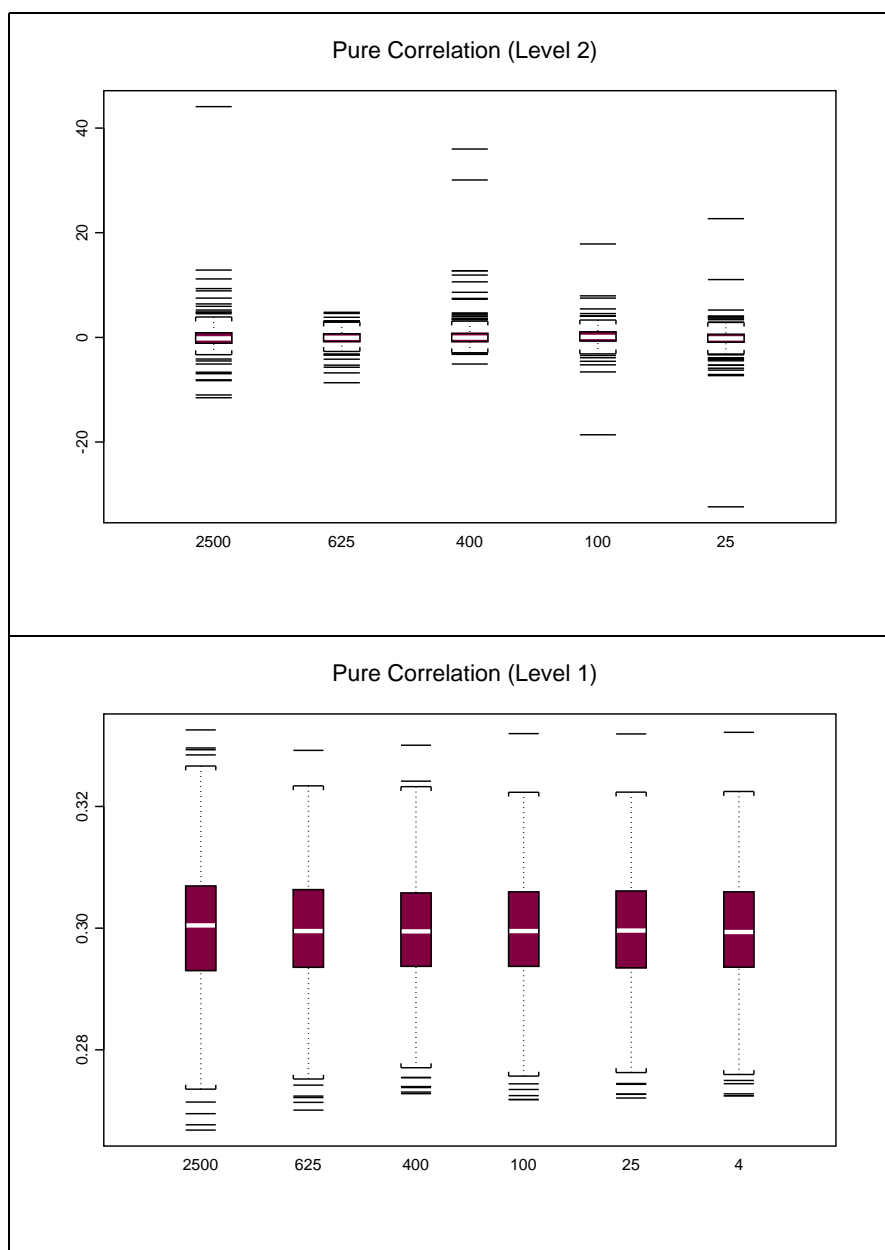


Figure 4.31: Pure Correlation, X and Y both are not autocorrelated

the values of the level 2 covariance components and are much larger than the level 1 covariance components.

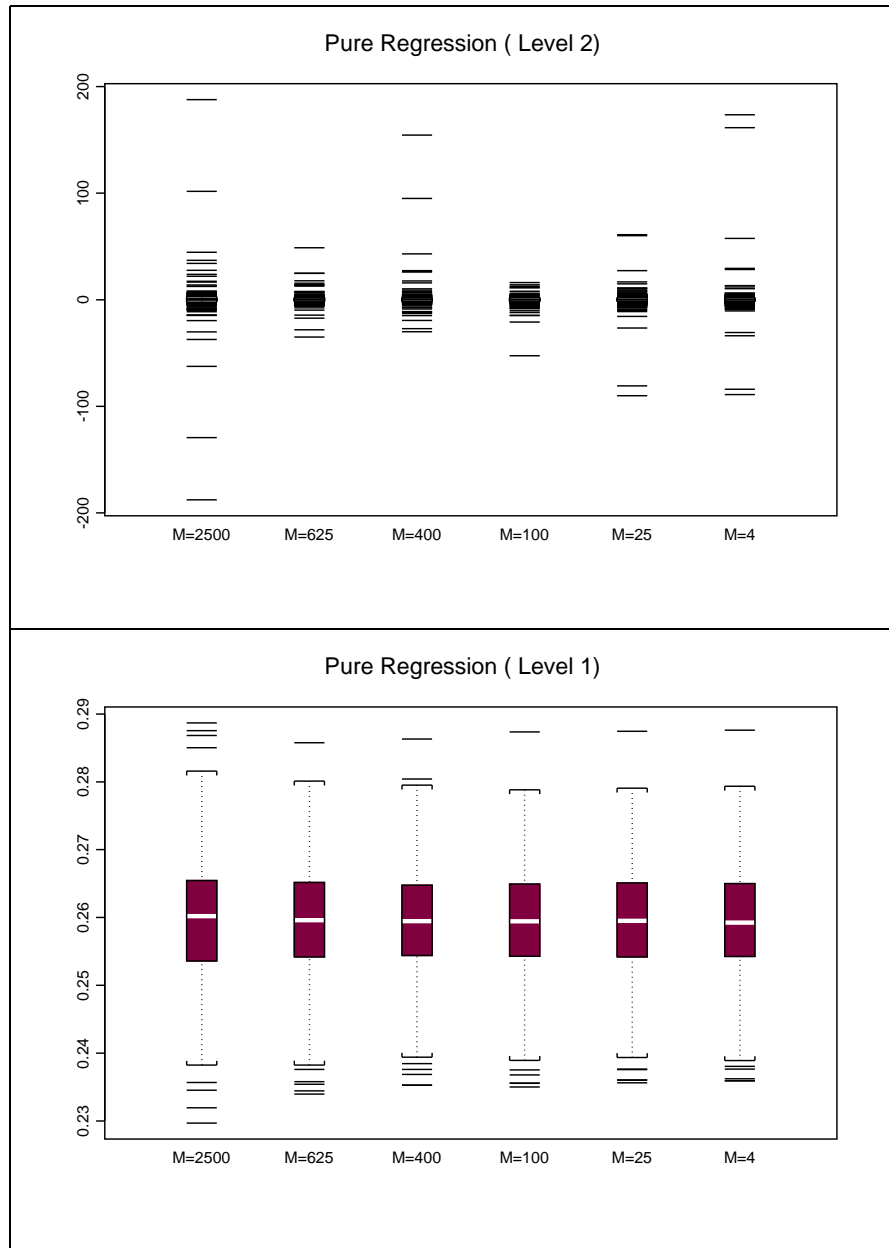


Figure 4.32: Level 2 and Level 1 Pure Regression, X and Y both are not autocorrelated

## 4.4 Comments on Experiments 1 and 2

Figure 4.33 shows the distributions of the correlation coefficients at different levels of aggregation and different degrees of autocorrelation. The data generation process cannot precisely control the specific degree of autocorrelation. The autocorrelation is described categorically as *very low*, *low*, *medium* and *high*. The *Very low* autocor-

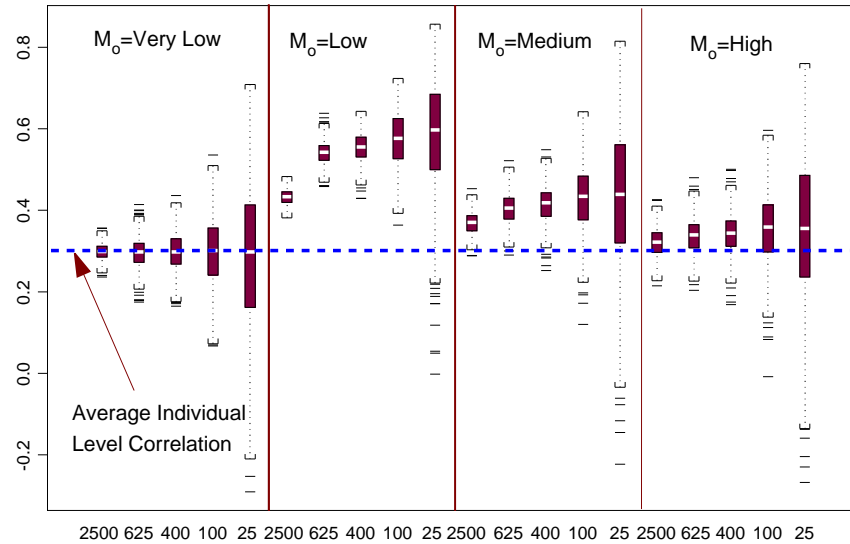
Level 2 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.1020	-0.1950	-11.549	44.1040	3.6360
625	-0.1110	-0.0709	-8.6598	4.7934	1.5719
400	0.5392	-0.0391	-5.0850	36.0027	3.7218
100	0.1710	0.0903	-18.6183	17.8646	2.3040
25	-0.2440	-0.197	-32.3930	22.6810	2.9840
4	0.0307	-0.2538	-16.7514	28.5288	3.3820
Level 1 Pure Correlation	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2999	0.3005	0.2668	0.3326	0.0107
625	0.2998	0.2995	0.2701	0.32926	0.0097
400	0.2996	0.2995	0.2728	0.3301	0.0095
100	0.2996	0.2995	0.2718	0.3320	0.0095
25	0.2996	0.2996	0.2721	0.3319	0.0095
4	0.2996	0.2994	0.2724	0.3322	0.0095

**Table 4.70: Description of the Level 2 and Level 1 Pure Correlation, X and Y both are not autocorrelated**

Level 2 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2440	0.2820	-187.695	187.704	15.149
625	0.3892	0.2512	-35.0007	48.7674	4.2018
400	0.7725	0.2587	-29.9521	154.3832	9.1065
100	0.0191	0.2107	-52.4610	16.1126	3.4556
25	0.0715	0.2072	-89.9981	61.0558	7.3015
4	0.5170	0.1650	-88.937	173.5570	12.686
Level 1 Pure Regression	Mean	Median	Minimum	Maximum	Std. Dev.
Individual	-	-	-	-	-
2500	0.2598	0.2602	0.2297	0.2887	0.0094
625	0.2596	0.2596	0.2340	0.2858	0.0084
400	0.2595	0.2595	0.2353	0.2863	0.0083
100	0.2595	0.2594	0.2350	0.2874	0.0083
25	0.2595	0.2595	0.2356	0.2874	0.0082
4	0.2595	0.2592	0.2359	0.2876	0.0082

**Table 4.71: Description of the Level 2 and Level 1 Pure Regression, X and Y both are not autocorrelated**

relation category has auto correlations almost equal to zero at the *individual* level as measured by Moran's I. The *Low* autocorrelation category has measures approximately equal to 0.2 at the *individual* level. The *Medium* autocorrelation and *High*



**Figure 4.33: Correlations at different levels of aggregation and degrees of autocorrelation**

autocorrelation categories have measures 0.6 and 0.8, respectively at the *individual* level.

Table 4.72 supports figure 4.33. When the autocorrelation of both variables (X and Y) initially are *very low*, the mean and median seems to be not affected by aggregation but the standard deviation increases with aggregation. When both variables have *low* autocorrelation, the mean and median and the standard deviations increase with aggregation. The values of the mean and median are larger than the corresponding Pearson correlation when the variables both have very low autocorrelation. From the same figure, the mean and median when both variables have *medium* degree of autocorrelation increase with aggregation but this time the increase is slower than when the variables have *low* autocorrelation. When both variables have *high* autocorrelation, the mean and median increase with aggregation in a very slow manner and the range and standard deviation increase with aggregation.

For a given scale there is an initial increase in the mean correlation as the autocorrelation goes from zero to a low level. Then, as the autocorrelation increases, the mean correlation decreases and the standard deviation increases.

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )
# of Zones				
2500	0.2989(0.0203)	0.4322(0.0185)	0.3686(0.0268)	0.3203(0.0342)
625	0.2980(0.0378)	0.5413(0.0288)	0.4040(0.0370)	0.3372(0.0425)
400	0.2978(0.0458)	0.5544(0.0343)	0.4148(0.0439)	0.3430(0.0486)
100	0.3019(0.0863)	0.5746(0.0682)	0.4291(0.0834)	0.3542(0.0893)
25	0.2980(0.1780)	0.5790(0.1426)	0.4304(0.1684)	0.3553(0.1762)
4	0.3630(0.5350)	0.5279(0.4756)	0.4013(0.5207)	0.3306(0.5392)

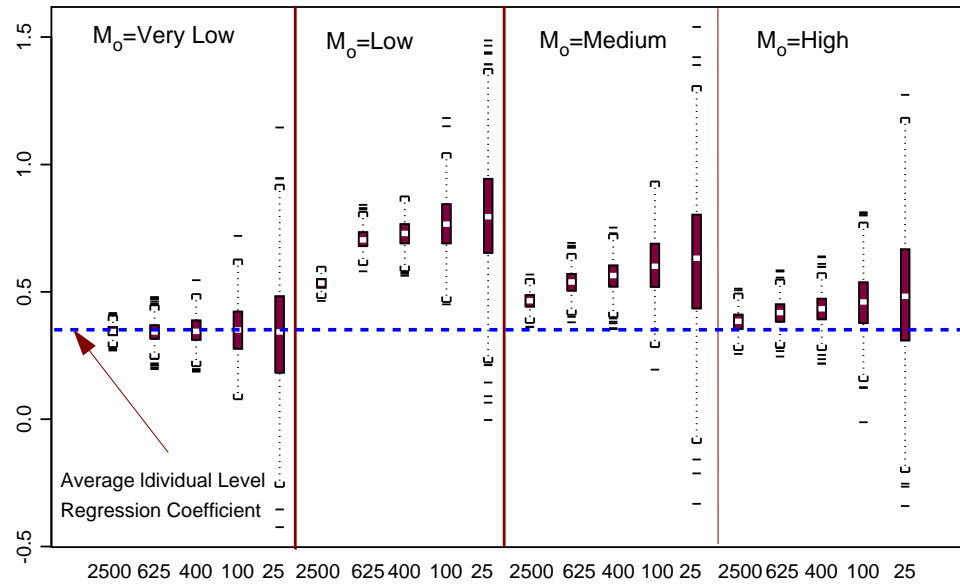
**Table 4.72: Summary of Correlations at Different Degrees of Autocorrelation and Levels of Aggregation**

Figure 4.34 shows the distribution of the regression coefficients at different levels of aggregation and different degrees of spatial autocorrelation. The figure is supported by Table 4.73. Looking at the table it is noted that the mean regression coefficient is not affected by aggregation when the level of autocorrelation is very low. Also it can be noted there is an increasing effect for other levels of autocorrelation but a slower increase as the level of autocorrelation increases.

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )
# of Zones				
2500	0.3452(0.0244)	0.5371(0.0288)	0.4646(0.0338)	0.3826(0.0409)
625	0.3430(0.0450)	0.7090(0.0478)	0.5377(0.0502)	0.4169(0.0532)
400	0.3479(0.0558)	0.7325(0.0583)	0.5623(0.0626)	0.4307(0.0626)
100	0.3495(0.1069)	0.7692(0.1174)	0.5997(0.1288)	0.4600(0.1241)
25	0.3360(0.2230)	0.7944(0.2375)	0.6238(0.2712)	0.4795(0.2538)
4	0.3340(1.1570)	0.7750(1.0290)	0.6640(1.2784)	0.4880(1.1234)

**Table 4.73: Summary of Regression Coefficient at Different Degrees of Autocorrelation and Levels of Aggregation**

Figure 4.35 shows the distributions of unweighted covariance of variables X and Y with different degrees of autocorrelation. Recall that the variance at the *individual* level for the two variables X and Y are 6.0 and 8.0, respectively. The mean



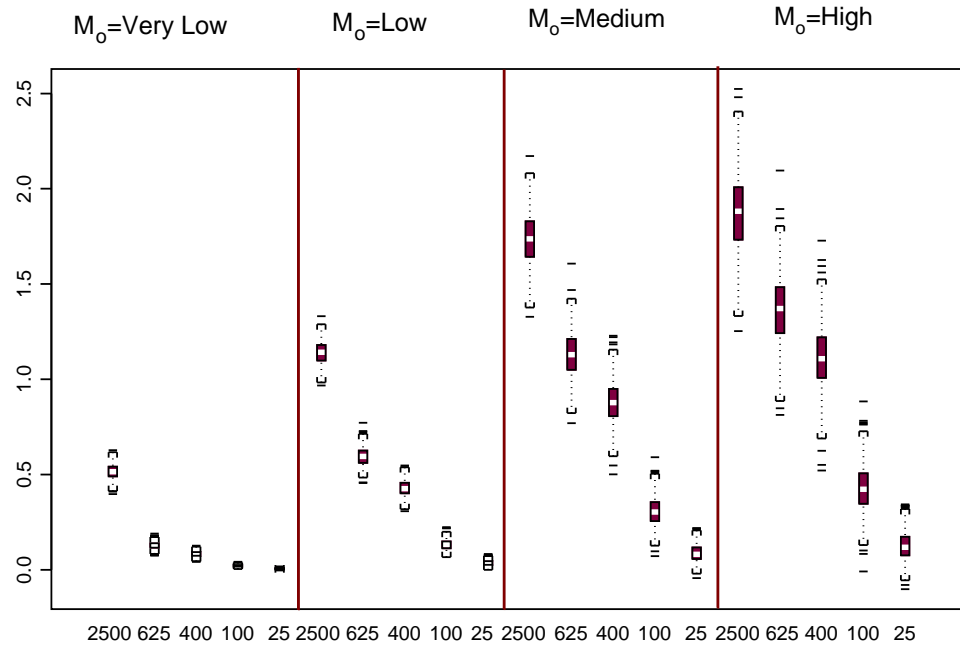
**Figure 4.34: Regression Coefficient at Different Levels of Aggregation and Degrees of Autocorrelation**

covariance for the different degrees of autocorrelation varies: for a *very low* degree of autocorrelation, the mean covariance is 0.5165, for *low*, *medium*, and *high*, the mean covariance are 1.1413, 1.7367, and 1.8728, respectively. The standard deviation of the covariance at each degree of autocorrelation decreases with aggregation. If we look at the standard deviation at each level of aggregation, the standard deviation increase as the degree of autocorrelation increases.

The unweighted variance of X and Y display a similar pattern. Because of these pattern, the standard deviation of the Pearson correlation as reflected in Figure 4.72 decrease with aggregation.

Figure 4.36 shows the distribution of the estimated level 2 *pure correlation* at *low*, *medium* and *high* autocorrelation. The results for data generated with *very low* autocorrelation are omitted because there are values that are not a characteristics of a correlation coefficient. For example, when the grid was aggregated to 2500 groups, the maximum value from the generated data of the level 2 *pure correlation* is 44.100





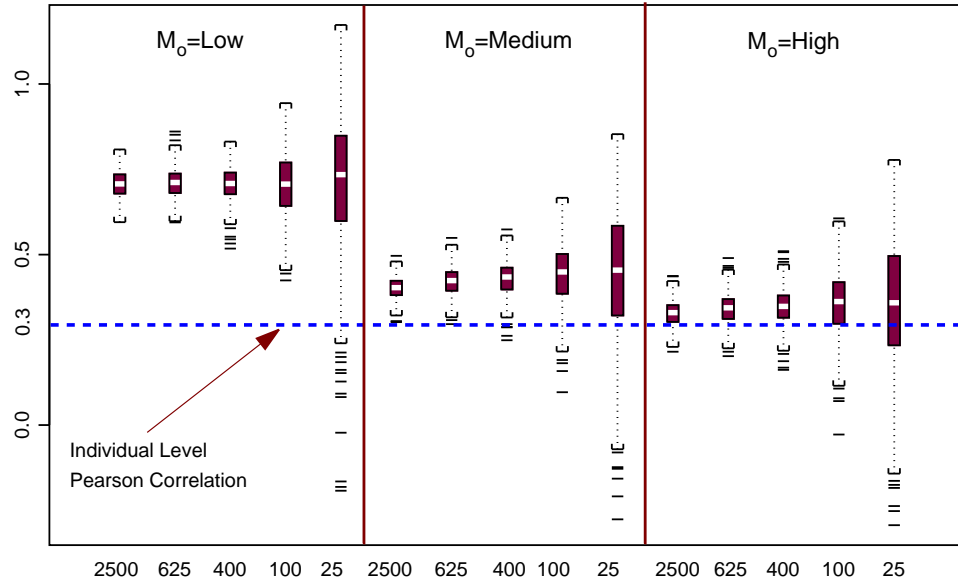
**Figure 4.35: Unweighted Covariance at Different Levels of Aggregation and Autocorrelation**

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
# of Zones				
2500	0.5165(0.0391)	1.1644(0.0800)	1.7367(0.1337)	1.8728(0.2030)
625	0.1284(0.0184)	0.6070(0.0556)	1.1323(0.1177)	1.3644(0.1810)
400	0.0830(0.0144)	0.4368(0.0476)	0.8797(0.1085)	1.1116(0.1693)
100	0.0207(0.0067)	0.1351(0.0274)	0.3064(0.0765)	0.4280(0.1256)
25	0.0049(0.0034)	0.0369(0.0151)	0.0880(0.0445)	0.1277(0.0750)
4	0.0007(0.0015)	0.0061(0.0066)	0.0151(0.0199)	0.0224(0.0352)

**Table 4.74: Summary of Covariance of X and Y at Different Degrees of Autocorrelation and Levels of Aggregation**

while the minimum is -11.550. Thus, in this particular case, that is, when both variables have very low autocorrelation, the level 2 *pure correlation* is not useful, as noted before.

When both variables have *low* autocorrelation, the mean (or median) of the level 2 *pure correlation* is not affected by aggregation but the value is higher than the



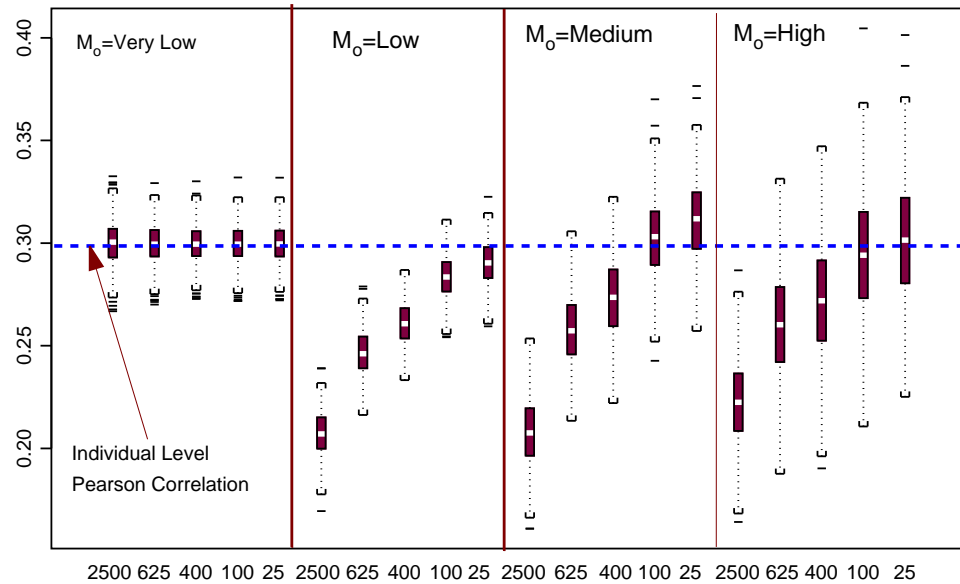
**Figure 4.36: Level 2 Pure Correlation at Different levels of Aggregation and Degrees of Autocorrelation**

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )
# of Zones				
2500	0.1020(3.636)	0.7080(0.0391)	0.4012(0.0308)	0.3271(0.0357)
625	-0.1110(1.5719)	0.7093(0.0415)	0.4212(0.0404)	0.3412(0.0441)
400	0.5392(3.7218)	0.7071(0.0468)	0.4308(0.0480)	0.3467(0.0507)
100	0.1710(2.3040)	0.7036(0.0903)	0.4423(0.0912)	0.3574(0.0937)
25	-0.2440(2.9840)	0.7084(0.2008)	0.4425(0.1853)	0.3580(0.1855)
4	0.0307(3.3820)	0.7871(1.2305)	0.4270(0.6770)	0.3421(0.6084)

**Table 4.75: Summary of Level 2 Pure Correlation at Different Degrees of Autocorrelation and Levels of Aggregation**

initial Pearson correlation and the standard deviation increases with aggregation. When both variables have *medium* autocorrelation, the mean (or median) increase slowly but the values are still higher than the initial Pearson correlation. When both the variables have *high* autocorrelation, the increase of the mean (or median) of the level 2 *pure correlation* is very slow and the values are near to the initial Pearson

correlation.



**Figure 4.37: Level 1 Pure Correlation at Different levels of Aggregation and Degrees of Autocorrelation**

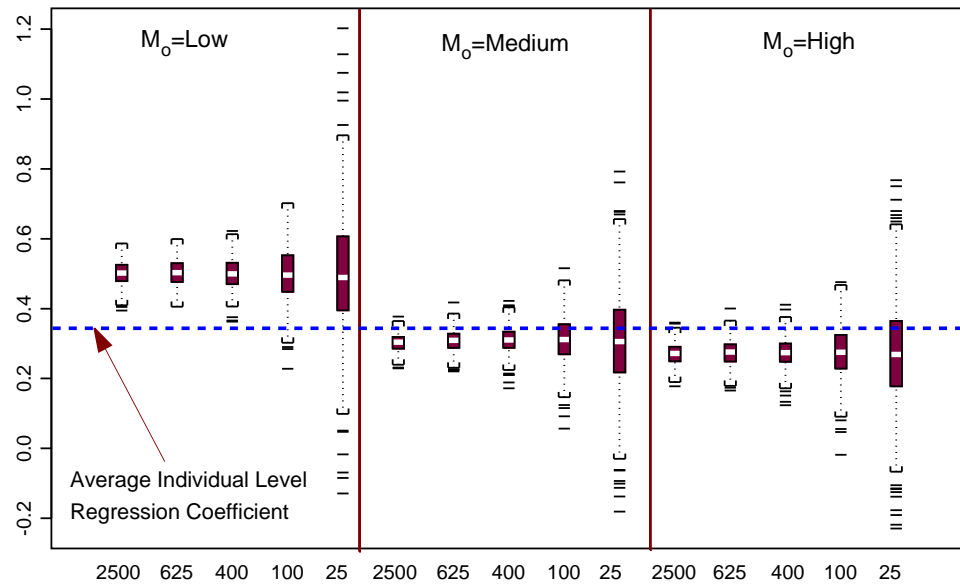
Figure 4.37 shows the distribution of the level 1 *pure correlation*. When both the variables have *very low* autocorrelation, the mean (or median) seems to be not affected by aggregation and the standard deviations decrease slightly with aggregation. When both variables have *low* autocorrelation, the mean (or median) of the level 1 *pure correlation* increases with aggregation and the standard deviation seems to be constant. The initial value of the mean is lower than the initial Pearson correlation and approaches the initial Pearson correlation as the number of zones decreases.

When both the variables have *medium* autocorrelation, the mean (or median) level 1 *pure correlation* displays a similar pattern, starting with a value less than the initial Pearson correlation and approaches the mean of the initial value. However, the standard deviation seems to be constant but is larger than when the both variables have *low* autocorrelation. When both variables have *high* autocorrelation, the same

pattern is displayed, that is, the mean increases with aggregation and approaches the initial mean Pearson correlation. The standard deviation seems to be constant but larger than when both variables have *low* and *medium* autocorrelation.

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
# of Zones				
2500	0.2999(0.0107)	0.2076(0.0111)	0.2076(0.0163)	0.2224(0.0206)
625	0.2998(0.0097)	0.2465(0.0103)	0.2583(0.0180)	0.2602(0.0259)
400	0.2996(0.0095)	0.2605(0.0102)	0.2739(0.0189)	0.2719(0.0281)
100	0.2996(0.0095)	0.2834(0.0105)	0.3030(0.0204)	0.2941(0.0305)
25	0.2996(0.0095)	0.2904(0.0105)	0.3121(0.0206)	0.3015(0.0307)
4	0.2996(0.0095)	0.2925(0.0106)	0.3149(0.0207)	0.3038(0.0306)

**Table 4.76: Summary of Level 1 Pure Correlation at Different Degrees of Autocorrelation and Levels of Aggregation**



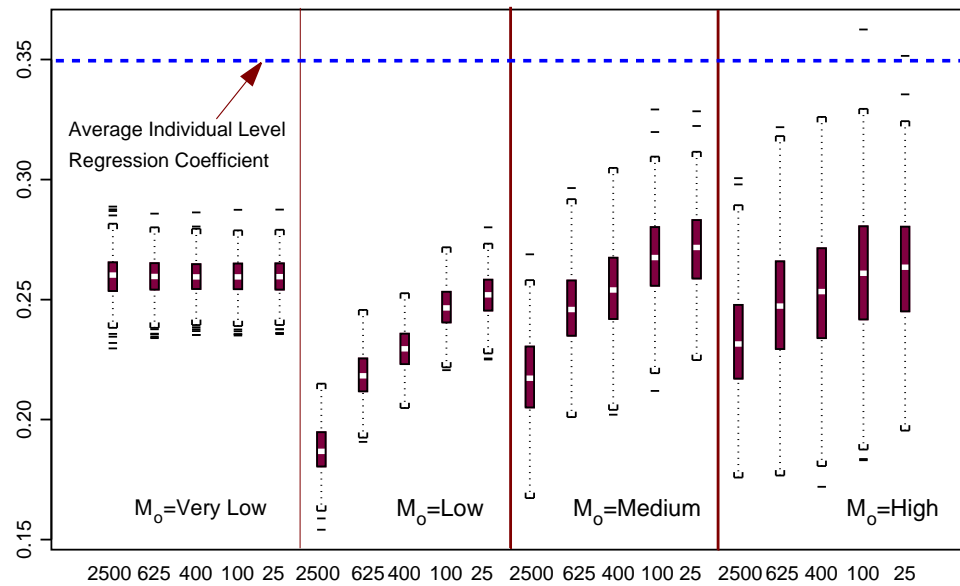
**Figure 4.38: Level 2 Pure Regression at Different levels of Aggregation and Degrees of Autocorrelation**

Figure 4.38 shows the distribution of level 2 pure regression coefficients at different levels of aggregation and different degrees of spatial autocorrelation. The

distribution of the variables with very low spatial autocorrelation was not included because of the issues with this coefficient in this case discussed previously.

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )
# of Zones				
2500	0.2440(15.149)	0.5018(0.0336)	0.3027(0.0239)	0.2699(0.0295)
625	0.3892(4.2018)	0.5033(0.0379)	0.3086(0.0311)	0.2738(0.0356)
400	0.7725(9.1065)	0.5009(0.0428)	0.3106(0.0364)	0.2741(0.0406)
100	0.0191(3.4556)	0.5006(0.0790)	0.3122(0.0687)	0.2750(0.0739)
25	0.0715(7.3015)	0.5010(0.1772)	0.3082(0.1419)	0.2718(0.1528)
4	0.5170(12.686)	0.5662(4.9090)	0.2827(1.3818)	0.2525(1.2087)

**Table 4.77: Summary of Level 2 Pure Regression at Different Degrees of Autocorrelation and Levels of Aggregation**



**Figure 4.39: Level 1 Pure Regression at Different levels of Aggregation and Degrees of Autocorrelation**

Table 4.77 shows the standard deviation at the different levels of aggregation. These values are large, in fact, when the individual level data is grouped into 2500 groups, the minimum level 2 pure regression is -187.6950 while the maximum is

187.704. When the variables both have *low* spatial autocorrelation, the mean of the level 2 pure regression at different levels of aggregation seems to be constant except when the data is grouped into 4 zones, where the value rises. However, the means are larger than the average individual level regression coefficient. When both variables have *medium* autocorrelation, the mean of the level 2 pure regression seems to be constant but this time the value is slightly less than the average individual level regression coefficient. Similar effects were observed when the variables have high autocorrelation, the means at different levels of aggregation seem to be constant but the values are less than when the variables have medium autocorrelation. In all cases, the standard deviation increases with aggregation.

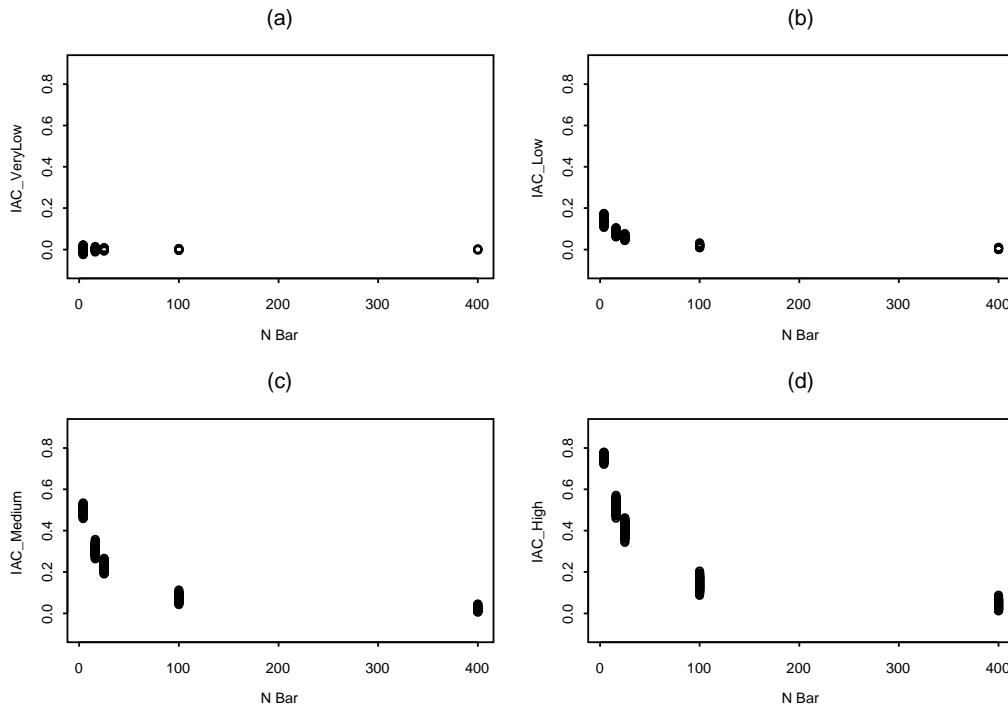
Figure 4.39 shows the distribution of level 1 pure regression at different levels of aggregation and different degrees of autocorrelation. When both variables have *low* autocorrelations, the mean of the level 1 pure regression seems to be constant for all levels of aggregation. However, the values are below the average individual level regression coefficient. When both variables have low autocorrelation, the mean of the level 1 pure regression coefficients increases with aggregation. A similar trend was observed when both variables have medium autocorrelation but this time the values are slightly higher at each level of aggregation compared with the case when both variables have low autocorrelation. When both variables have high autocorrelation, the mean of the level 1 pure regression increases with aggregation except when the data is aggregated into 4 groups where the mean decreases slightly.

	$M_o = \text{Very Low}$	$M_o = \text{Low}$	$M_o = \text{Medium}$	$M_o = \text{High}$
	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
# of Zones				
2500	0.2598(0.0094)	0.1876(0.0101)	0.2172(0.0175)	0.2322(0.0225)
625	0.2596(0.0084)	0.2184(0.0092)	0.2465(0.0177)	0.2482(0.0257)
400	0.2595(0.0083)	0.2293(0.0090)	0.2546(0.0180)	0.2535(0.0269)
100	0.2595(0.0083)	0.2466(0.0091)	0.2681(0.0184)	0.2608(0.0277)
25	0.2595(0.0082)	0.2518(0.0092)	0.2719(0.0181)	0.2629(0.0269)
4	0.2595(0.0082)	0.2534(0.0092)	0.2730(0.0179)	0.2634(0.0265)

**Table 4.78: Summary of Level 1 Pure Regression at Different Degrees of Autocorrelation and Levels of Aggregation**

In all cases, the standard deviation at each level of aggregation and degree of autocorrelation (that is, very low, low, medium, and high) seems to be approximately constant but increases as the degree of autocorrelation increases as shown in the Figure 4.39 and Table 4.78.

### Relationship between Intra-Area Correlation and $\bar{N}$



**Figure 4.40: Relationship between Intra-Area Correlation and N-Bar at different degrees of autocorrelation for the two variable: (a) very low, (b) low, (c) medium, and (d) high**

As we aggregate into larger groups, that is as  $M$  decreases, we would expect the intra-area correlation to decrease as we are including more units and increasing the average distance between units within a group. Figure 4.40 shows the relationship between the intra-area correlation and  $\bar{N}$  at different degrees of autocorrelation. When the variable has very low autocorrelation the standard deviation decreases as  $\bar{N}$  increases and the mean intra-area correlation was zero. When the autocorrelations of the variables increase there is a non-linear relationship between the intra-area

correlation and  $\bar{N}$  with it decreasing as the  $\bar{N}$  increases.

Figure 4.41 shows the mean of the intra-area correlation against  $\bar{N}$  at different degrees of autocorrelation. There is a non-linear trend that decreases and approaches zero as  $\bar{N}$  increases. The decrease depends on the degree of autocorrelation.

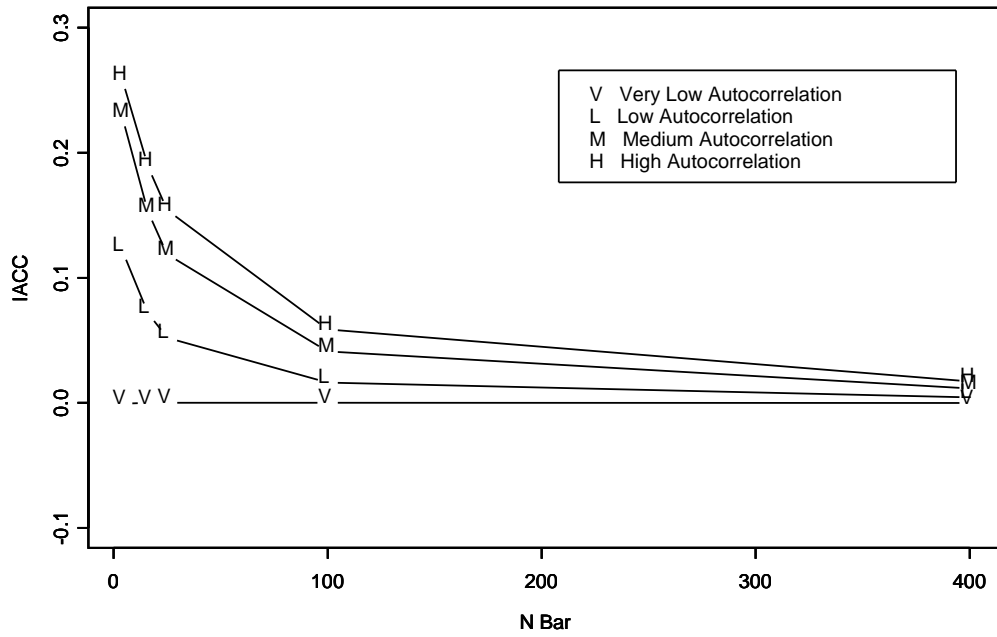
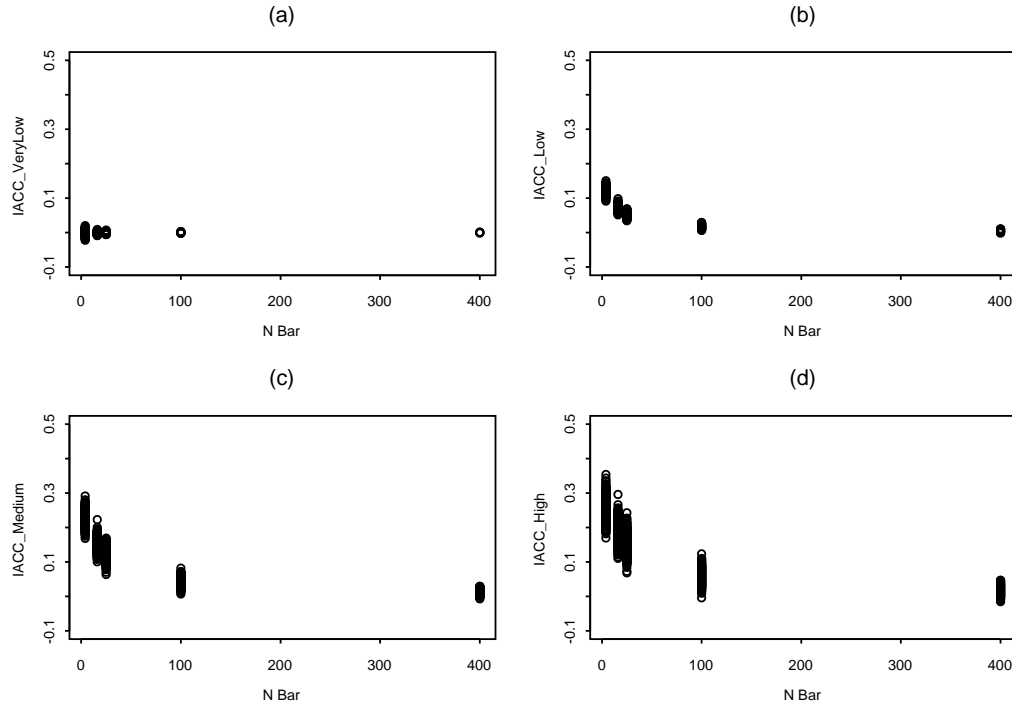


Figure 4.41: Relationship between the mean of Intra-Area Correlation and N-Bar at different degrees of autocorrelation for the two variable

### Relationship between Intra-Area Cross-Correlation and $\bar{N}$

Figure 4.42 shows the relationship between the intra-area cross-correlation and  $\bar{N}$  at different degrees of autocorrelation. The relationship is similar to the relationship between the intra-area correlation and  $\bar{N}$ . When both variable have very low autocorrelation the standard deviation decreases as  $\bar{N}$  increases and the median seems to be constant. When the autocorrelations of the variables increase there is a non-linear relationship between the intra-area correlation and  $\bar{N}$  and it decreases as the  $\bar{N}$  increases.





**Figure 4.42: Relationship between Intra-Area Cross-Correlation and N-Bar at different degrees of autocorrelation for the two variable: (a) very low, (b) low, (c) medium, and (d) high**

Figure 4.43 shows the mean of the intra-area cross-correlation against  $\bar{N}$  at different degrees of autocorrelation. Similar to the relationship between the mean of intra-area correlation there is a non-linear trend that decreases and approaches zero as  $\bar{N}$  increases. The decrease depends on the degree of autocorrelation. The decrease of high autocorrelation of the two variables is steeper because the initial mean is a higher value.

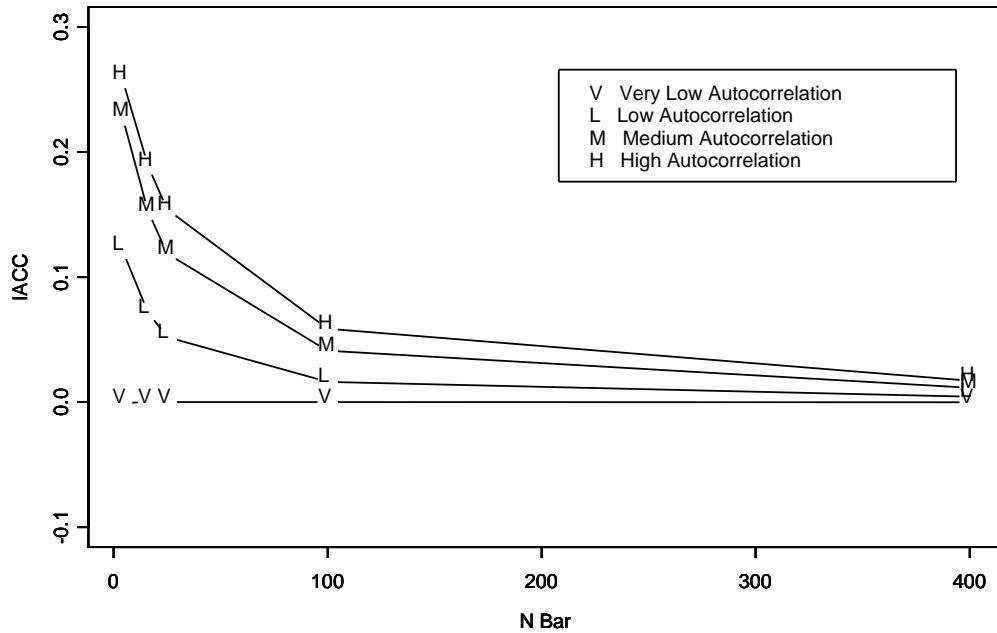
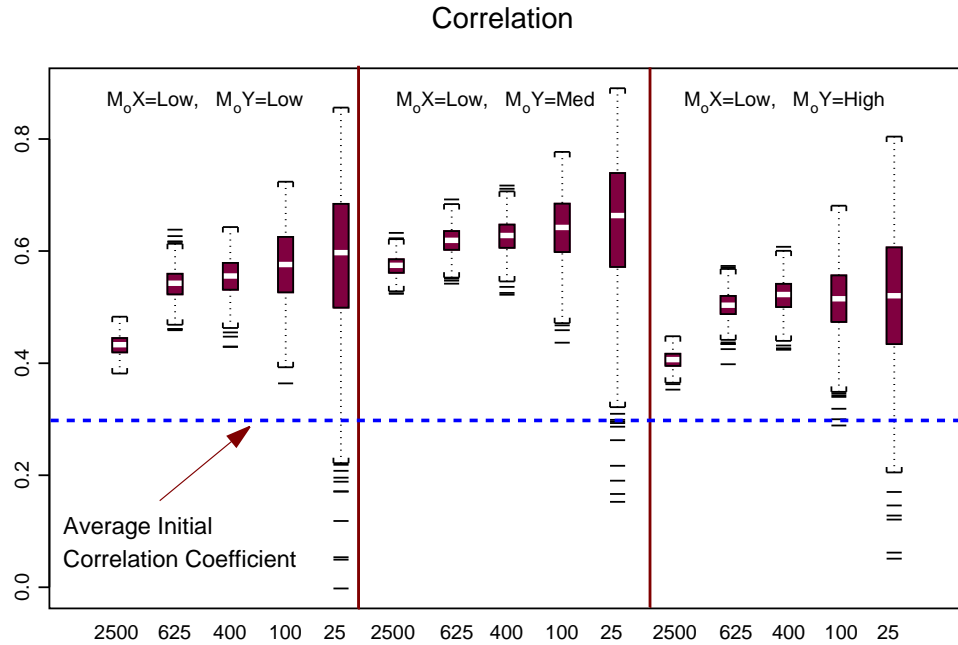


Figure 4.43: Relationship between the mean of Intra-Area Cross-Correlation and N-Bar at different degrees of autocorrelation for the two variable

## 4.5 Experiment 3: Scale effects when the variables do not have the same levels of autocorrelation

In experiment 1 the degree of autocorrelation for each variable was similar. In this experiment data are generated so that variable X will have *low* autocorrelation and the autocorrelation of variable Y varies from *low* to *medium* and to *high* autocorrelation. We tried to make the other properties of the initial individual level data consistent with the other data sets. When X has low autocorrelation and Y also has low autocorrelated, the mean initial correlation is 0.29, with mean 0.005 and 10, variance 6.0 and 8.0 for variables X and Y, respectively. The mean initial correlation when X has low autocorrelation and Y has medium autocorrelation is 0.31 and when variable X has low autocorrelated and variable Y is high autocorrelated, the mean

initial Pearson correlation is 0.28.



**Figure 4.44: Pearson Correlation variable X (low autocorrelation) and variables Y (different autocorrelation)**

Figure 4.44 shows the distributions of the estimated correlations for the three cases described above based on 500 simulations. One apparent pattern is the mean and the median and the standard deviation increase with aggregation in all cases. Most of the correlations are greater than the average initial correlation. There are some cases in which the correlations are less than the average initial correlation, especially when the data are aggregated into 25 groups. These are the cases in which the percentage loss of covariance is greater than the percentage loss of variance of the two variables X and Y, resulting in a decrease in the correlations.

For a given scale the standard deviation is little affected as the degree of autocorrelation in Y increases. The mean of the correlation increases at the autocorrelation goes from low to medium and then decreases as the autocorrelation becomes high

Figure 4.45 shows the distribution of estimated regression coefficients for the three cases being considered. The mean values of the regression coefficient at differ-

	$M_oX = Low$ $M_oY = Low$	$M_oX = Low$ $M_oY = Medium$	$M_oX = Medium$ $M_oY = High$
	Mean( <i>SD</i> )	Mean( <i>SD</i> )	Mean( <i>SD</i> )
# of Zones			
2500	0.4321(0.0185)	0.5732(0.0185)	0.4062(0.0158)
625	0.5413(0.0288)	0.6187(0.0255)	0.5038(0.0258)
400	0.5544(0.0343)	0.6265(0.0309)	0.5209(0.0318)
100	0.5746(0.0682)	0.6395(0.0609)	0.5144(0.0639)
25	0.5790(0.1426)	0.6413(0.1310)	0.5160(0.1292)
4	0.5279(0.4756)	0.5903(0.4321)	0.4194(0.5292)

Table 4.79: Summary of correlations when X have low autocorrelation and Y have different levels of autocorrelation at different levels of aggregation

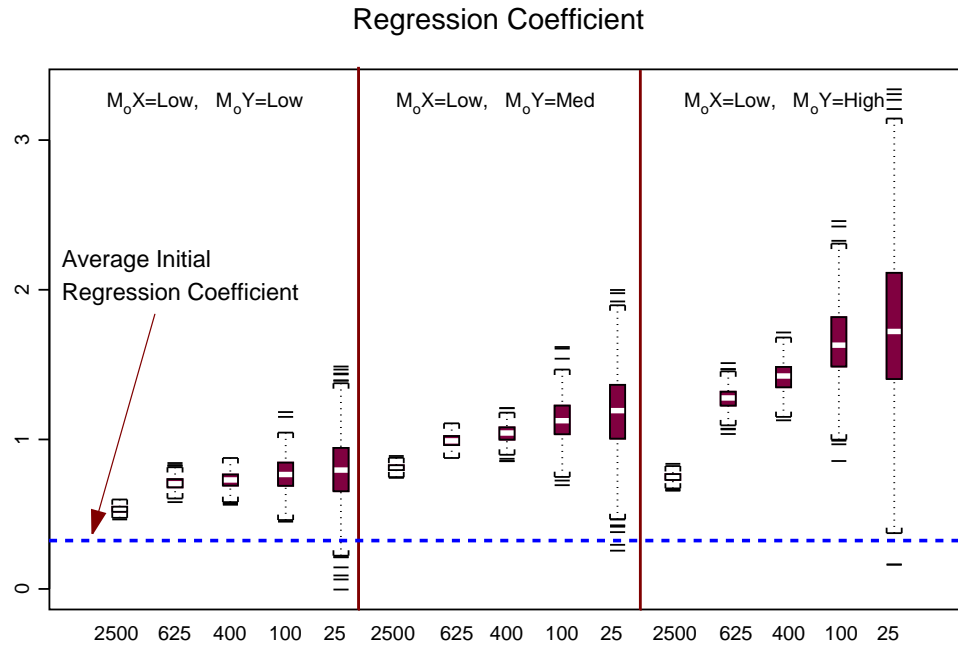
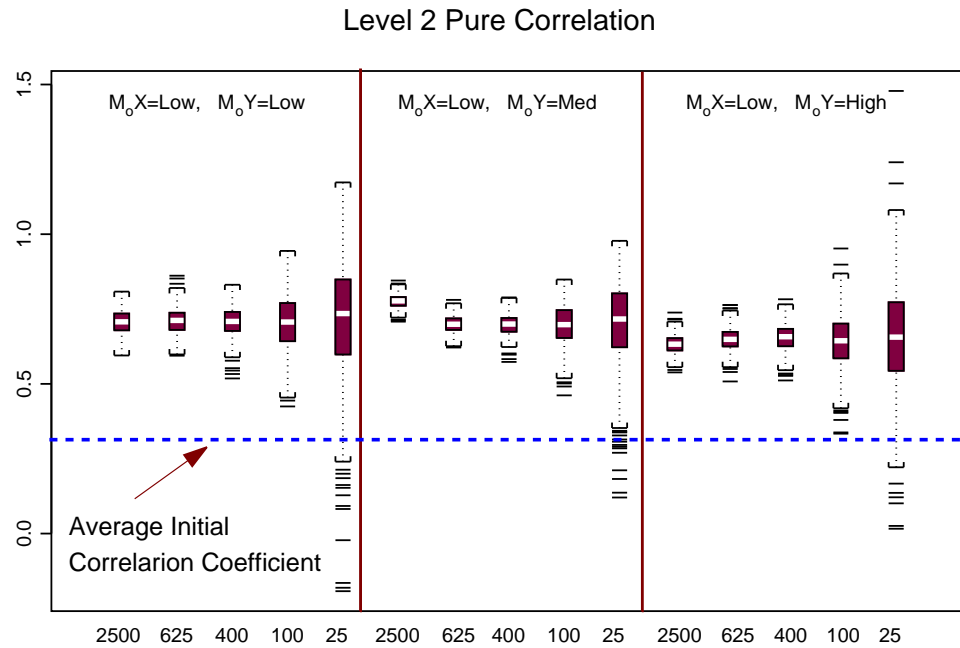


Figure 4.45: Regression Coefficient variable X (low autocorrelation) and variables Y (different autocorrelation)

ent levels of aggregation are all greater than the average initial regression coefficient. In fact, except for few regression coefficients when the data are aggregated into 25 zones, they are all greater than the average initial regression coefficient.

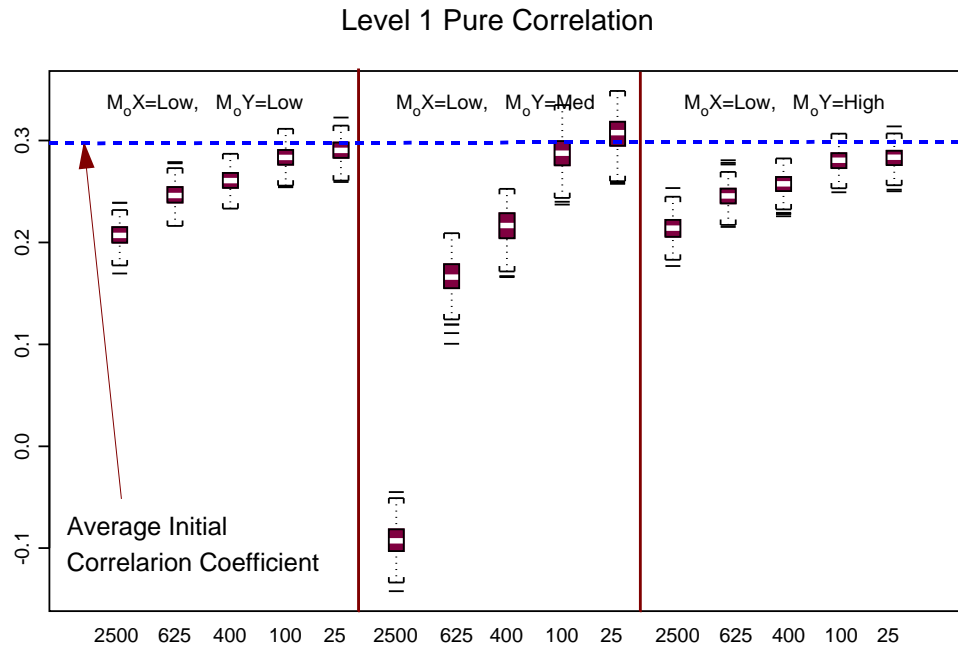
Figure 4.46 shows the distribution of the level 2 *pure* correlation. The mean when variable X has a low autocorrelation and variable Y have low but slightly



**Figure 4.46: Level 2 Pure Correlation, variable X (low autocorrelation) and variables Y (different autocorrelation)**

higher autocorrelation seems to be not affected by aggregation, the mean of pure correlation coefficients are 0.70803, 0.70934, 0.70710, 0.70360, 0.70836, 0.7871 when the data are aggregated into 2500, 625, 400, 100, 25, and 4 zones respectively. When variable X has Low autocorrelated and variable Y has medium autocorrelation, the mean also seems to be not affected by aggregation, except when the data are aggregated into 2500 and 4 zones, the mean of the pure coefficients are 0.775570, 0.699358, 0.69723, 0.69617, 0.69617, 0.7736, respectively. The mean and median of the levels 2 pure correlation coefficient of the last case also seems to be not affected by aggregation except when the data are aggregated into 4 zones, the mean are respective, 0.63239, 0.64831, 0.65488, 0.64272, 0.65487, 0.5552. In all cases, the standard deviation increases with aggregation and the mean and median is larger than the initial correlations.

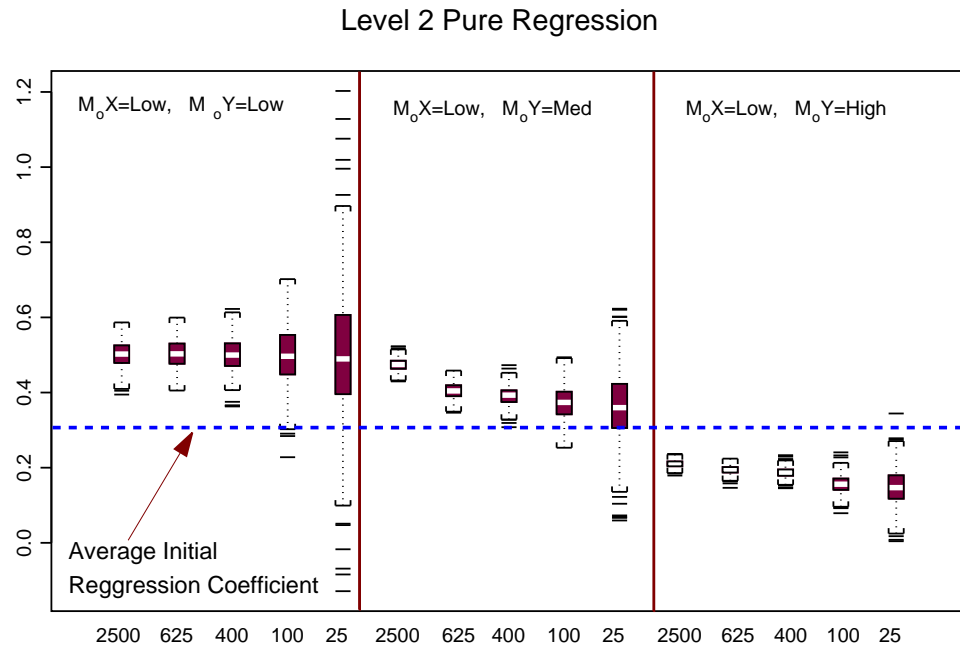
Figure 4.47 shows the distribution of the level 1 *pure* correlation. The mean in all cases increases with aggregation and tends to approach the initial correlation as



**Figure 4.47: Level 1 Pure Correlation, variable X (low autocorrelation) and variables Y (different autocorrelation)**

the number of zones decrease to 4 zones. The standard deviation of the values seems to be constant in each case. One unusual result is the case when the variable Y has medium autocorrelation and aggregated into 2500 groups, not only the mean but all of the level 1 pure coefficient is negative. This happens because all the level 1 covariance components have negative values, this is possible, but the way the level 1 and level 2 covariance component were computed, means that the level 2 variance component is greater than the initial covariance. This means that covariances increase when the data is aggregated into 2500 groups.

Figure 4.48 shows the distribution of level 2 *pure regression*. Except in the case when the two variables have almost the same degree of autocorrelation and in this experiment when variables X and Y have low autocorrelation, the mean and median correlation is not affected by aggregation although the values are higher than the initial average regression coefficient. When variable Y has medium autocorrelation, the mean correlation decreases with aggregation and approaches the average initial



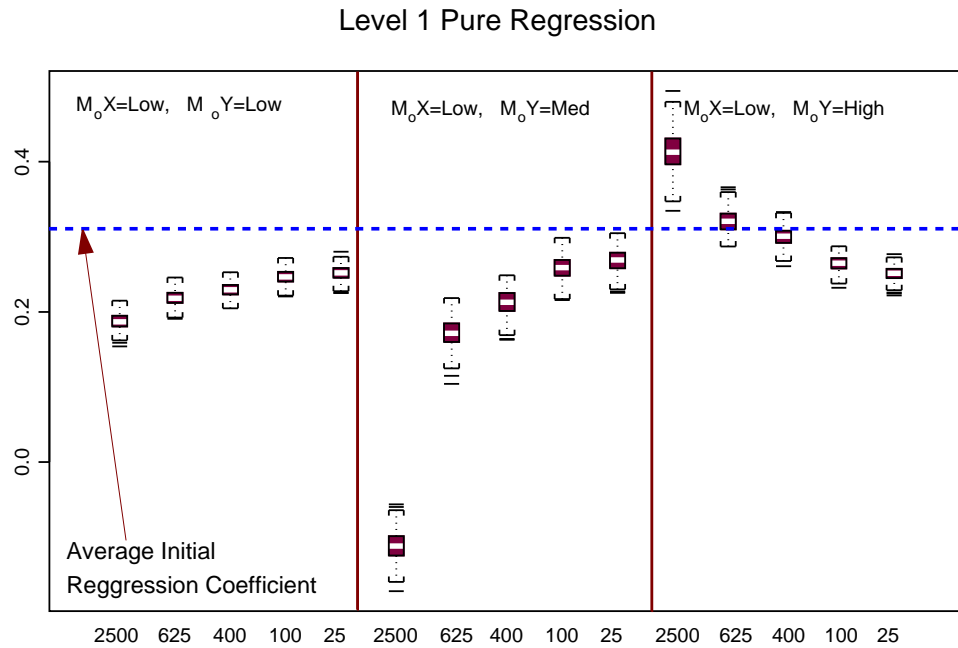
**Figure 4.48: Level 2 Pure Regression, variable X (low autocorrelation) and variables Y (different autocorrelation)**

regression coefficient. When variable Y has high autocorrelation, there is a slow decrease of the mean and median of the level 2 pure regression but these values are less than the average initial regression coefficients. In all cases the standard deviation increases with aggregation.

Figure 4.49 shows the distribution of the level 1 *pure regression* coefficient. The first two cases have similar patterns, that is, increasing and approaching the average initial regression coefficient as the number of groups is decreased but the third case has decreasing mean. In terms of the standard deviation, the first two cases seem to have constant standard deviation but in the third case, that is, variable Y is highly autocorrelated the standard deviation decreases with aggregation.

### Summary of Experiment 3

The Pearson correlation coefficient is affected by aggregation in all three cases. When the data are aggregated to 2500 zones, the minimum value of the correlation



**Figure 4.49: Level 1 Pure Regression, variable X (low autocorrelation) and variables Y (different autocorrelation)**

is less than the initial correlation coefficient and the maximum is greater in all three cases. When the data are aggregated into 625, 400, and 100 zones, the standard deviation of the values of the Pearson correlation does not include the initial Pearson correlation coefficient, in all three cases the values are greater than the initial value.

The mean of the level 2 pure correlation seems to be not affected by aggregation except for cases when the data are aggregated into 2500 and 4 zones. In all three cases, the standard deviation of values of the pure correlation increases with aggregation.

The mean of the level 1 pure correlation increase with aggregation and approaches the initial correlation as the number of zones decreases and the standard deviation seems to be constant, in all three cases.



## 4.6 Summary

Statistics derived from a simple multilevel model such as pure correlation and pure regression coefficients were investigated. Their corresponding direct statistics were also examined and compared. Other statistics derived from the simple multilevel model were also investigated. All the statistics being investigated were affected by the initial degree of autocorrelation and scale.

Regardless of the initial degree of autocorrelation of the variables the weighted and unweighted variances, correlation and regression coefficients were affected by aggregation.

Based on the results of the experiments we see that when the autocorrelation is very low the mean of the direct correlation is close to the individual level correlation. When autocorrelation is present the mean of the direct correlation increases with the degree of aggregation, that is, as the scale decreases to 625 and thereafter the increase is minimal. For a given scale the degree of autocorrelation affects the direct correlation, initially increasing as the autocorrelation increases, but then decreases when one or both of the variables has high autocorrelation. As the level of aggregation increases the dispersion of the distribution of the direct correlation increases, which is reflected in the standard deviation. When there are only 25 groups there are some values of the direct correlation less than the individual level correlation. The standard deviation is only moderately affected by the degree of autocorrelation. In going from very low to low autocorrelation the standard deviation decreases, but increases as the autocorrelation increase further.

The standard deviation of level 1 pure correlation seems to be stable regardless of the initial degree of autocorrelation and level of aggregation. The mean when the initial autocorrelation is very low also shows stability. However, when the initial autocorrelation of the variables were low, medium, and high the mean started with a low correlation (less than 0.3) when aggregated to 2500 groups and approaches the initial correlation of 0.3 as the aggregation increases to 25 groups.

For the level 2 pure correlation, the standard deviations increases with aggre-

gation except when the initial autocorrelation is very low. The mean level 2 pure correlation did not show stability when the initial correlations of the variables were very low. When the initial autocorrelation of the variables were low, medium and high, the mean level 2 pure correlation shows stability but the values are higher than the initial correlation of 0.3 and approaches 0.3 as the degree of autocorrelation increased.

Relationships between  $\bar{N}$  and the intra-area correlation, where  $\bar{N} = \frac{N}{M}$  shows what was expected when the variable had autocorrelation, that is, the intra-area correlation decreases as  $\bar{N}$  increased. When the autocorrelations of the variables increases there is a non-linear relationship between the intra-area correlation and  $\bar{N}$  and it decreases as  $\bar{N}$  increases. However, when the variable have very low autocorrelation the mean of the intra-area correlation is constant. Similar relationships were observed between the intra-area cross-correlation and  $\bar{N}$ .

The mean and the weighted mean is not affected by aggregation and autocorrelation. The distribution of the weighted group level variance becomes more disperse as the scale increases and the number of groups becomes small and the small number of degrees of freedom involved in calculating the variance, which is  $M-1$ , where  $M$  is the number of groups. The change in standard deviation with scale is more than would be expected through the change in  $M$ . If the weighted variance behaved as proportional to a  $\chi^2_{M-1}$  random variable the ratio of its standard deviation to the mean, which is its coefficient of variation (CV), would be  $\sqrt{2/(M-1)}$ , which would be the case for no autocorrelation. Results in table 4.11 table 4.29 , and table 4.45 give values a little less than these theoretical values.

When there is no spatial autocorrelation equation 4.2 suggests that the group level correlation will be close to the individual level correlation, although there is a tendency to increase in absolute value when the number of groups is quite small. Equation 4.3 gives a theoretical formula for the standard deviation of the group level correlation when no spatial correlation is present and predicts that it increases as the number of groups decreases, i.e. as scale increases. These results are confirmed empirically in table 4.67.

As the level of aggregation increases the dispersion of the distribution of the direct correlation increases, which is reflected in the standard deviation.

For of the pure coefficients, as the scale increases we should expect the estimated level 1 variance component to increase when there is spatial autocorrelation because more dissimilar units are included in each group and hence the estimated level 2 variance decreases. This is seen in tables 4.15, 4.32 and 4.48. The rate of the increase in the level 1 variance component estimates and the decrease in the level 2 estimates depends on the level of the autocorrelation, being greater with higher levels of autocorrelation.

The means of the level 1 correlation, when there is spatial autocorrelation, are affected by aggregation, starting below the individual level correlation but approaching it as the number of groups decreases. This is because, as the number of groups becomes smaller, the groups become larger and the individuals within them become more like the whole population. Spatial autocorrelation has little effect on the mean of the level 1 correlations.

## Chapter 5

# Analysis of Real Data from UK Census

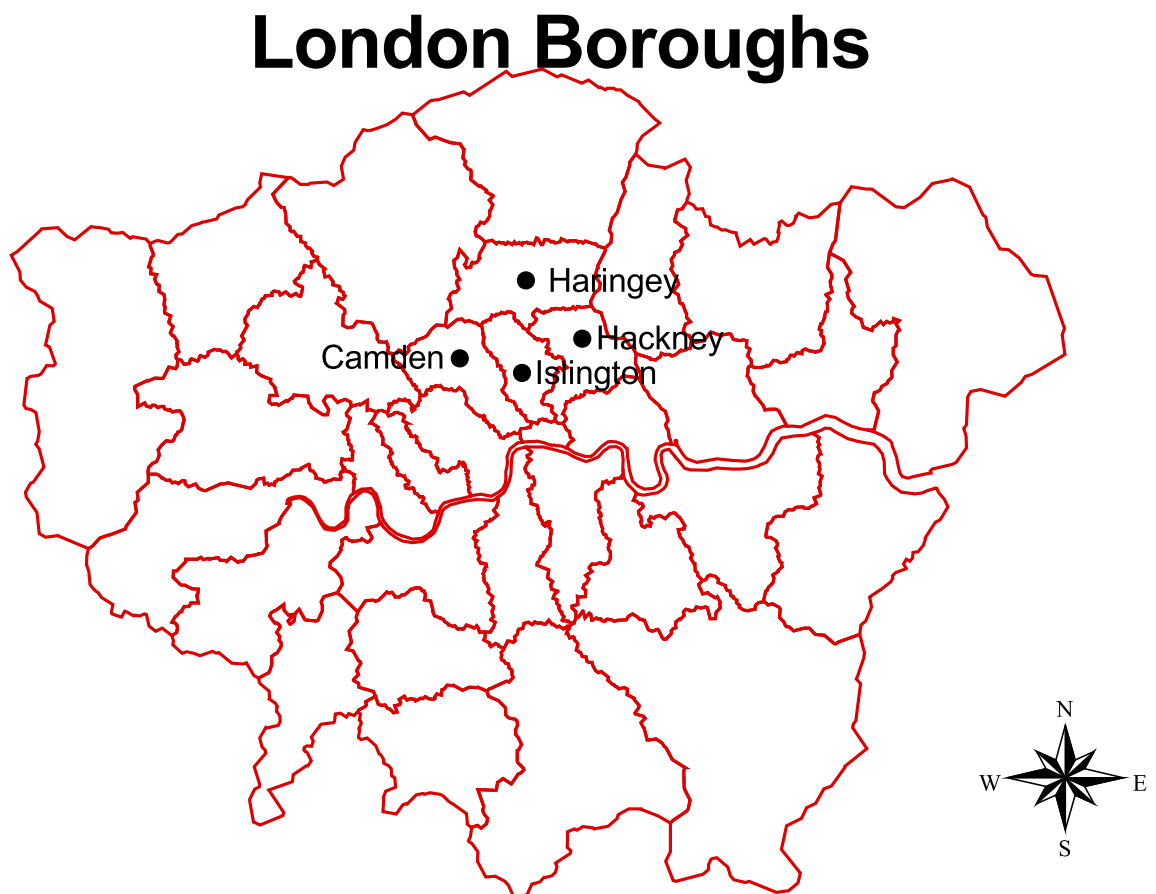
This chapter describes the results of analyses of scale effects on relevant statistics being considered in this thesis based on real data from the UK Census. Directly calculated statistics, such as correlation and regression coefficients are considered and statistics associated with a simple multilevel model. Using real data provides results for groups that vary in population size and are arranged in a less regular spatial manner than in the simulated data. The previous chapter relied entirely on simulated data.

### 5.1 Data from two sources

To further investigate the behavior of common statistics when the data are aggregated, actual data from the 1991 UK Census are used. Three levels of data are considered in this analysis; individual level, Enumeration District (ED) level, and the Ward level.

The individual level data are taken the 1991 SARs (Samples of Anonymised Records). The 1991 SARs correspond to a two percent sample of individuals counted in households and communal establishments of Great Britain. Several variable are included in the 2% SARs, some of them will be used in this experiment. The lowest

geographical indicator available is the SAR District, which is an area of at least 120000 people, to protect the confidentiality of information. The SAR Districts considered in the study are Camden, Hackney, Haringey, and Islington and are part of London boroughs. This will provide individual level data from the UK Census. The data are used to provide an estimate of the individual level covariance between variables. Figure 5.1 shows the location of the four districts considered in this study.



**Figure 5.1: Location of the four districts**

Figure 5.2 shows the boundaries of the districts, Ward, and Enumeration Districts (EDs). An enumeration district (ED) is the lowest geographical level in the 1991 UK population census for which aggregate data are released. These EDs are grouped into larger geographical areas called Wards. The Wards are also grouped into larger geographical areas called districts. The Census data from both ED level

and Ward level of the UK population census are extracted from Small Area Statistics (SAS) data base. Figure 5.2 shows the region composed of the districts Camden, Hackney, Haringey, and Islington. It comprises 1904 EDs nested into 92 Wards and 4 districts.

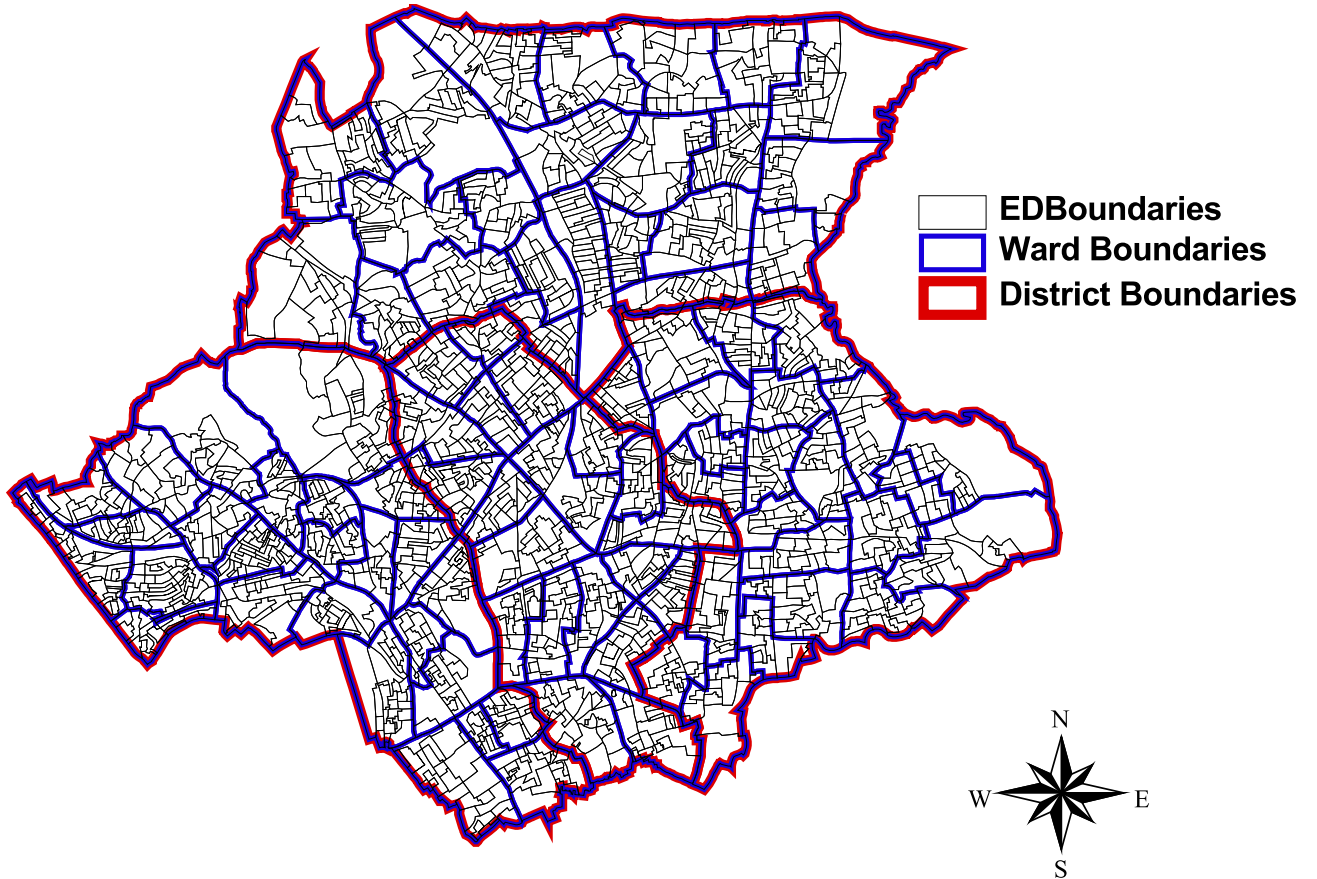


Figure 5.2: The region with its boundaries

Table 5.1 shows the number of individuals from the SAR and SAS and the number of Enumeration Districts and Wards for the districts being considered in this study. Although the SAR does not contain ED and Ward indicators, it is possible to estimate variance components at these level by combining the SAR and SAS data as shown by Tranmer and Steel (2001).

Based on Table 5.1, the average number of individuals in the districts Camden, Hackney, Haringey, and Islington per ED are 307,429, 386, and 398 respectively and the overall average number of individuals per ED is 380. The average number of

District	Number of Individuals (SAR)	Population Counts (SAS)	Number of EDs	Number of Wards
Camden	3508	165877	540	26
Hackney	3378	180540	421	24
Haringey	3832	201620	522	23
Islington	3249	163500	411	19
Total	13967	711537	1904	92

**Table 5.1: Individuals counts from SAR and SAS and number of EDs and Wards for each district**

individuals per Ward for the districts are 6380, 7523, 8766, and 8605 respectively and the overall average number of individuals per Ward is 7818.

#### **Variables to be investigated:**

In this study the following variables were considered.

**age:** percentage of individuals between 16 and 65, inclusive

**ftw:** percentage of full-time workers

**uemp:** percentage of unemployed

**liti:** percentage of individuals with limiting long term illness

**nocar:** percentage of individuals with no car

Both the percentages of full-time workers and percentage of unemployed are included as they reflect different aspects of the labor force. We shall see later that these variables are not particularly highly correlated at any of the levels considered, since people may also be part-time workers or economically inactive. Besides their direct measurement of the labor force, these variables reflect socio-economic characteristics. The percentage of individuals with no car is included as it is often used as an indicator of lower socio-economic areas. The chosen variables cover a range of measures (see table (5.2)).

For each of the  $M$  EDs, the total population in the region being considered is

$$N = \sum_{g=1}^M N_g$$

The average number of individuals per ED is

$$\bar{N} = \frac{N}{M}$$

We will use the following model from Tranmer and Steel (1998):

$$Y_i = \mu_y + \alpha_g + \varepsilon_i \quad (5.1)$$

where

$Y_i$  represent the value of Y for the  $i$ th individual in area  $g$  (ED)

$\mu_y$  is the population mean of Y

$\alpha_g$  is a random variable representing the area effect for the  $g$ th ED

$\varepsilon_i$  is a random variable representing pure individual effect

**Assumptions:**

1.  $E(\alpha_g)=0$  ,  $E(\varepsilon_i)=0$  , and  $\text{var}(\alpha_g)=\sigma_\alpha^2$  ,  $\text{var}(\varepsilon_i)=\sigma_\varepsilon^2$
2. The area effect and individual effects are uncorrelated.

$$\text{cov}(\alpha_g, \varepsilon_i) = 0$$

3. The effects for different individuals are uncorrelated, for  $i \neq j$

$$\text{cov}(\varepsilon_j, \varepsilon_i) = 0$$

Properties under Model specified by (5.1) are:

$$E(Y_i) = \mu_y$$

$$\text{var}(Y_i) = \sigma_\alpha^2 + \sigma_\varepsilon^2 = \sigma_y^2$$

$$\text{cov}(Y_i, Y_j) = \begin{cases} \sigma_\alpha^2, & \text{if } i \text{ and } j \text{ are from the same ED} \\ 0 & \text{otherwise.} \end{cases}$$



For a second variable, say X, similar variance components can be specified.

To be able to compute other statistics later, we need the following: the covariance between the area effects between two variables Y and X can be described as,

$$\text{cov}(\alpha_{(Y)g}, \alpha_{(X)g}) = \sigma_{(yx)\alpha}$$

The individual level covariance for the two variables is

$$\text{cov}(\varepsilon_{(Y)i}, \varepsilon_{(X)i}) = \sigma_{(YX)\varepsilon}$$

The covariance between the two variables is

$$\text{cov}(Y_i, X_i) = \sigma_{(YX)\alpha} + \sigma_{(YX)\varepsilon} = \sigma_{YX}$$

and the values of the two variables for two different individuals in the same group(ED) is

$$\text{cov}(Y_i, X_j) = \begin{cases} \sigma_{(YX)\alpha}, & \text{if } i \text{ and } j \text{ are from the same ED} \\ 0 & \text{otherwise.} \end{cases}$$

### 5.1.1 Case 1: Individual level from SAR and second level is Enumeration District (ED)

The individual level data used are from the SAR in which the variables have value 1 or 0. The total number of individuals in the sample is 13967. The variables were converted into dichotomous variables. The value of the variable is 1 if the description of the variable is satisfied, otherwise it is 0. For example, for variable age, age=1 if the age of the individual is between 16 and 64, inclusive otherwise age=0.

To compute the individual level statistics for variable Y from the SAR the following are used:

$$p = \frac{f_1}{n}$$

where  $f_1$  is the number of observations with value=1, n is the number of individuals and the variance is

$$S_{YY}^{(1)} = (1 - p)p.$$

The individual level mean and variance can be computed from the Census SAS data using:

$$P_g = \frac{F_g}{N_g} \quad (5.2)$$

where  $F_g$  is the number of individuals with a specific characteristics and  $N_g$  is the total number of individuals in group  $g$ .

The unweighted group mean is defined as

$$\tilde{p} = \frac{1}{M} \sum_{g=1}^M P_g \quad (5.3)$$

where  $M$  is the number of groups used in the census (ED, Ward, District)

The Unweighted group level variance is defined as

$$\tilde{S}_{YY} = \frac{1}{M-1} \sum_{g=1}^M (P_g - \tilde{p})^2 \quad (5.4)$$

The weighted mean is

$$P = \frac{\sum_{g=1}^M N_g P_g}{\sum_{g=1}^M N_g} \quad (5.5)$$

The weighted group level variance is defined by

$$S_{YY}^{(2)} = \frac{1}{M-1} \sum_{g=1}^M N_g (P_g - P)^2 \quad (5.6)$$

Note that the data from the Census are dichotomous at the individual level. The individual level variance can be computed using

$$S_{YY}^{(1)} = p(1 - p) \quad (5.7)$$

or

$$S_{YY}^{(1)} = P(1 - P) \quad (5.8)$$

from the SAR and SAS respectively.

Weighted variance and covariance are used in this study because the number of individuals in each ED or Ward differs. Also, if there is no within area correlation they give unbiased estimate for the individual level variance and covariance.

To be able to investigate some components of a simple multilevel model we need estimates for the variance and covariance components. From Tranmer and Steel (1998) group-level (eg ED-level) variance components for variables Y and X can be estimated using

$$\hat{\Lambda}_{YY}^{(2)} = \frac{S_{YY}^{(2)} - S_{YY}^{(1)}}{\bar{N}^* - 1} \quad (5.9)$$

and

$$\hat{\Lambda}_{XX}^{(2)} = \frac{S_{XX}^{(2)} - S_{XX}^{(1)}}{\bar{N}^* - 1} \quad (5.10)$$

where

$$\bar{N}^* = \bar{N} + \frac{\bar{N} - \bar{N}^0}{M - 1}, \quad \bar{N}^0 = \frac{1}{N} \sum_{g=1}^M N_g^2, \quad \bar{N} = \frac{N}{M}.$$

Similarly the ED level covariance component can be estimated using:

$$\hat{\Lambda}_{YX}^{(2)} = \frac{S_{YX}^{(2)} - S_{YX}^{(1)}}{\bar{N}^* - 1} \quad (5.11)$$

The SAR provides data to estimate the level 1 variance and the SAS provides the data to estimate the level 2 variances. Tranmer and Steel (1998) show that the variance and covariance estimates given by (5.9), (5.10) and (5.11) are unbiased for the model given by (5.1).

### Some statistics from SAR (Individual level)

Computations for the individual level statistics from the SAR are done using SPSS and R. Table 5.2 shows the mean and variance computed from the SAR and different levels from the Census SAS. Notice the small differences of the means and variances from the SAR and the means and variances of the variables computed from the Census, which are due to the SAR being a sample.

<i>SAR</i>			<i>Census(SAS)</i>					
			ED		Ward		District	
Variables	mean	var	mean	var	mean	var	mean	var
age	0.6738	0.2198	0.6818	0.2169	0.6761	0.2190	0.6745	0.2196
ftw	0.2941	0.2076	0.3042	0.2117	0.3722	0.2337	0.3721	0.2336
unemp	0.0907	0.0825	0.0860	0.0786	0.1092	0.0973	0.1104	0.0982
llti	0.1322	0.1147	0.1254	0.1096	0.1258	0.1100	0.1264	0.1104
nocar	0.4657	0.2488	0.4806	0.2496	0.4923	0.2499	0.4925	0.2499

**Table 5.2: Mean and Individual Level Variances from SAR and different levels from Census**

Table 5.3 shows the variance-covariance matrix calculated from the SAR to be used in the computations of some multilevel components. These statistics will be used later in the computation of the estimates of the variance and covariance components. The computations are done using SPSS.

	age	ftw	unemp	llti	nocar
age	0.2198	0.0931	0.0291	-0.0214	-0.0310
ftw	0.0931	0.2076	-0.0267	-0.0276	-0.0330
unemp	0.0291	-0.0267	0.0825	-0.0039	0.01219
llti	-0.0210	-0.0276	-0.0030	0.1147	0.0213
nocar	-0.0310	-0.0330	0.0121	0.0213	0.24882

**Table 5.3: Variance-Covariance matrix Individual Level: SAR**

Table 5.4 shows the correlation at the individual level computed from the SAR using SPSS.

	age	ftw	unemp	llti	nocar
age	1.0000	0.4360	0.2160	-0.1350	-0.1326
ftw	0.4360	1.00000	-0.2040	-0.1791	-0.1442
unemp	0.2160	-0.2040	1.0000	-0.0400	0.0838
llti	-0.1350	-0.1791	-0.0400	1.0000	0.1261
nocar	-0.1326	-0.1442	0.0838	0.1261	1.0000

**Table 5.4: Correlations at Individual level**

### Some statistics from ED level

From the Census, there are a total of 711537 individuals in 1904 enumeration districts, so that  $\bar{N}=373.71$  and  $\bar{N}^*=373.69$ . Table 5.5 shows the weighted means and variances calculated from the SAS at Enumeration District (ED) level using equations (5.5) and (5.6), respectively. These statistics will be used to compute the estimates of some statistics that are relevant to the present study. The table also shows the aggregation effect defined as  $S^{(2)}/S^{(1)}$ . The aggregation effect on the variance is a simple measure of the strength of the within-group autocorrelation. From (3.28) we see that the aggregation effect for variable Y is  $1 + (\bar{N}^* - 1)\hat{\Delta}_Y^{(2)}$ . We have also seen in section 3.1.8, how the intra-area correlation is related to spatial autocorrelation within groups as measured by Moran's I with appropriate spatial proximity weights. In an applied context, examination of the aggregation effect indicates which variables have greater scale effects. The aggregation effect will be equal to 1 if there is no aggregation effect, which will occur if there is no within group autocorrelation. From the results, it is evident that there are substantial aggregation effects, but they vary considerably between the variable. In particular the variable "nocar" has much longer aggregation effect than any of the other variables.

Variables	weighted mean	weighted variance	aggregation effect ( $S^{(2)}/S^{(1)}$ )
age	0.6729	2.2213	10.1060
ftw	0.2948	2.0320	9.7881
unemp	0.0876	0.4022	4.8751
llti	0.1273	0.8544	7.4490
nocar	0.4832	8.3418	33.5281

**Table 5.5: Weighted Mean and Variances from SAS (ED level)**

Table 5.6 shows the variance-covariance matrix at ED level. Compared with the individual level variance-covariance matrix, the table shows increase in the absolute values in all cases. However, the covariance between the variables *llti* and *unemp* changes from negative at the individual level to positive at the ED level and all the other combinations of the variables retain their sign.

	age	ftw	unemp	llti	nocar
age	2.2213	1.6986	0.0603	-0.7641	-1.5430
ftw	1.6986	2.0320	-0.1790	-0.5726	-1.5625
unemp	0.0603	-0.1790	0.4022	0.0934	0.9295
llti	-0.7641	-0.5726	0.0934	0.8544	1.5872
nocar	-1.5430	-1.5625	0.9295	1.5872	8.3218

**Table 5.6: Variance-Covariance matrix ED Level**

Table 5.7 shows the correlation at ED level. Comparing with the individual correlation matrix, we see that not all the correlations increase with aggregation.

	age	ftw	unemp	llti	nocar
age	1.0000	0.7995	0.0638	-0.5547	-0.3586
ftw	0.7795	1.0000	-0.1979	-0.4346	-0.3795
unemp	0.0638	-0.1979	1.0000	0.1593	0.5075
llti	-0.5547	-0.1358	0.1593	1.0000	0.5945
nocar	-0.3586	-0.3795	0.5075	0.5945	1.0000

**Table 5.7: Correlations at ED level**

There are five variables being considered in this study, so there are ten possible correlations calculated from all possible pairs of variables. Table 5.8 shows the correlations for these pairs of variables, where the second column shows the correlations at the individual level and the third column the correlation at ED level. For six of the combinations of the variables the absolute values of the correlation increases from individual level to ED level and were of the same sign, while the rest decreased or changed sign. The correlation between variables *age* and *unemp* decreased to 0.0638 at the ED level from 0.2160 at the individual level. This can be explained by the aggregation effects of the variances of the variables and the percentage increase in the covariance. The aggregation effects of the variances of both variables is relatively large compared with the aggregation effect of the covariance which is equal to 1.072, resulting in a decrease in the correlation coefficient.

Figure 5.3 shows the behavior of the correlation in going from the individual level to ED level. The figure shows that there are five pairs of variables with greater

Variable Combination	Correlations Individual	Correlations ED level
1. age-ftw	0.4360	0.7795
2. age-unemp	0.2160	0.0638
3. age-llti	-0.1350	-0.5547
4. age-nocar	-0.1326	-0.3586
5. ftw-unemp	-0.2040	-0.1979
6. ftw-llti	-0.1791	-0.1358
7. ftw-nocar	-0.1442	-0.3795
8. unemp-llti	-0.0400	0.1593
9. unemp-nocar	0.0838	0.5075
10. llti-nocar	0.1261	0.5945

**Table 5.8: Variable combinations and correlations at different levels**

correlations and five below the individual level correlations when aggregated to ED level.

Table 5.9 shows the estimated variance components and the intra-area (intra-ED) correlations for the five variables. The table shows that the level 2 variance components is very small compared with the corresponding level 1 variance components resulting in small but typical values of the intra-ED correlations. There were no problems associated with negative estimates of variance components.

Variable	level1 variance	level2 variance	intra-area correlation
age	0.21442	0.00537	0.02443
ftw	0.20271	0.00490	0.02358
unemp	0.08162	0.00086	0.01040
llti	0.11274	0.00198	0.01730
nocar	0.22711	0.02172	0.08727

**Table 5.9: Variance components and Intra-area correlation**

The aggregation effect can be defined as the ratio  $S_{YY}^{(2)}/S_{YY}^{(1)}$  (Steel, et. al. (1996)). Table 5.10 shows the aggregation effect and the corresponding intra-area correlation of the variables being considered. Note that even though the estimate of intra-area correlation is small, aggregation effects are substantial because  $S_{YY}^{(2)} \approx S_{YY}^{(1)} (1 + (\bar{N}^* - 1)\delta_{YY})$  and  $\bar{N}^*$  is large.

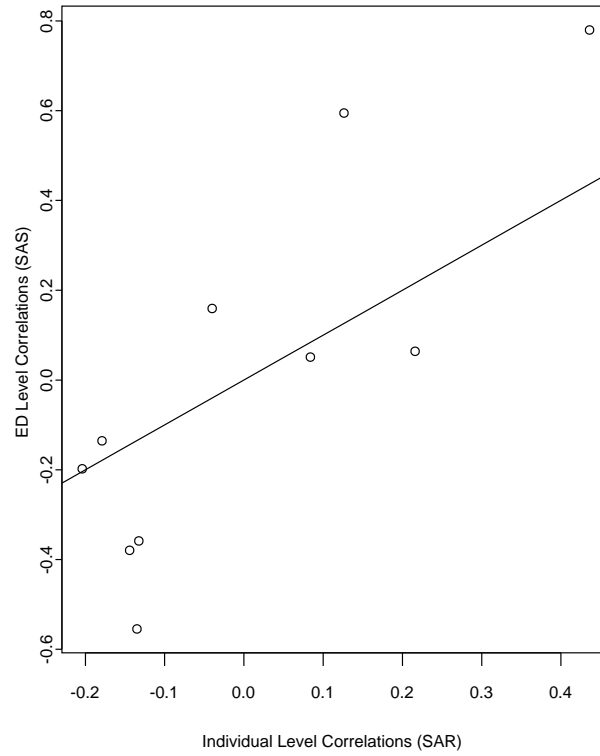


Figure 5.3: Correlations: Individual level (SAR) and ED level (SAR)

Variables	Aggregation Effect	Intra-area Correlation
age	10.1060	0.0244
ftw	9.7881	0.0236
unemp	4.6767	0.0104
liti	7.4490	0.0173
nocar	33.5281	0.0873

Table 5.10: Aggregation Effect and Intra-Area Correlation

Table 5.11 shows the intra-area cross-correlations. The signs of the intra-area cross correlations are the same as the signs of the correlations at the ED level. Recall that the intra-area cross-correlation is a measure of the within group homogeneity of a pair of variables and that similarity of the values of two variables within areas can be measured using this statistic. Looking at the values, there seems to be low



	age	ftw	unemp	llti	nocar
age	*	0.0202	0.0006	-0.0126	-0.0174
ftw	0.0202	*	-0.0003	-0.0095	-0.0181
unemp	0.0006	-0.0003	*	0.0027	0.0172
llti	-0.0126	-0.0095	0.0027	*	0.0249
nocar	-0.0174	-0.0181	0.0172	0.0255	*

**Table 5.11: Intra-area Cross-correlation**

similarity within areas for the values of the pairs of variables being considered here.

### Pure correlation

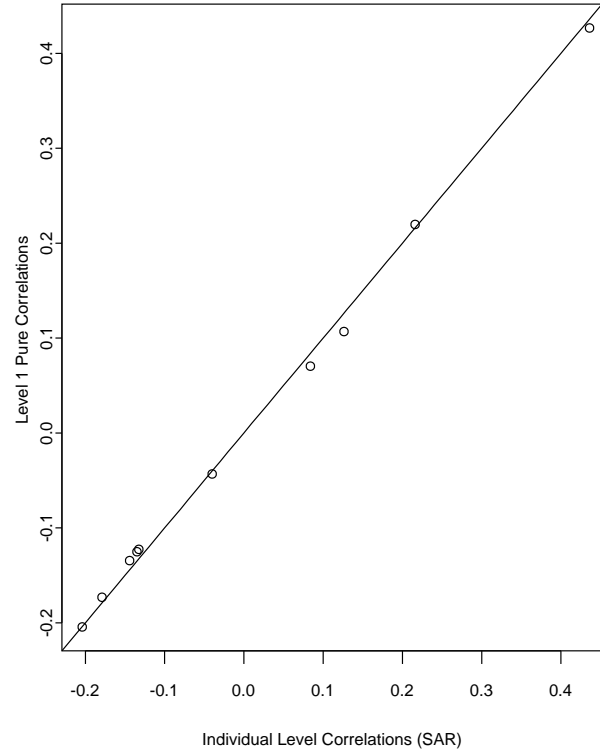
Table 5.12 shows the estimated level 1 pure correlation coefficients for each of the pairs of variables. The level 1 pure correlations are similar to the corresponding values of the correlations computed from the SAR. The signs are the same as the signs of the corresponding correlations at the individual level computed from the SAR.

	age	ftw	unemp	llti	nocar
age	*	0.4266	0.2192	-0.1251	-0.1226
ftw		*	-0.2044	-0.1731	-0.1344
unemp			*	-0.0433	0.0703
llti				*	0.1068
nocar					*

**Table 5.12: Level 1 Pure correlation**

Figure 5.4 shows level 1 pure correlations plotted against the corresponding individual level correlations computed from SAR. The level 1 pure correlations are either higher or lower than the corresponding individuals level correlations but the differences are very small. The individual level, the direct correlation and pure correlation are effectively the same.

Table 5.13 shows the estimated level 2 pure correlations. The level 2 pure correlations are not similar to the individual level correlations; the values are either larger or smaller in absolute values and one even changed sign, that is, from negative to



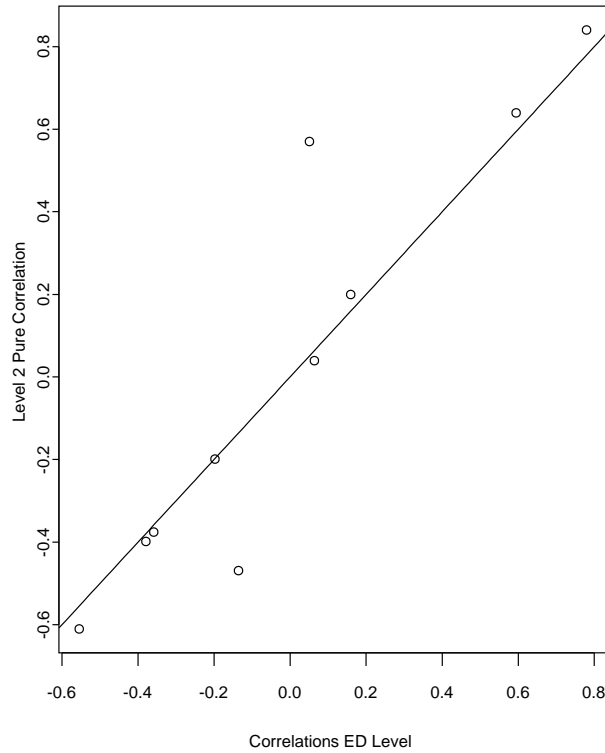
**Figure 5.4: Individual level Correlations(SAR) versus Level 1 Pure Correlation**

positive. The level 2 pure correlations have the same sign as the corresponding correlations at ED level computed from SAS. Except for one pair of variables (age and unemp), the pure correlations are larger in absolute values compared with the corresponding correlations at ED level.

	age	ftw	unemp	llti	nocar
age	*	0.8401	0.0390	-.6104	-0.3757
ftw		*	-0.1993	-0.4691	-0.3981
unemp			*	0.2001	0.5703
llti				*	0.6400
nocar					*

**Table 5.13: Level 2(ED level) Pure correlation**

Figure 5.5 shows the relationship when the ED level correlations are plotted



**Figure 5.5: ED Level Correlations versus Level 2 Pure Correlation**

against the corresponding level 2 pure correlations. It shows that  $r^{(2)} \approx \rho^{(2)}$ , that is, the ED correlations are essentially estimating the level 2 pure correlation in general except for two cases. Hence in general, correlations calculated at the ED level are almost entirely determined by relationships at ED level, and have very little to do with individual level relationships.

The study of the regression and pure regression will be considered in the later part of this chapter.

### 5.1.2 Case 2: Individual level from SAR and second level is Ward

A similar study is conducted, this time the Ward level serves as level 2. However, there is discrepancy in terms of the total number of individuals reported at the Ward level and the total number of individuals reported at ED level. This is because confidentiality was protected by "*adding +1 or -1 to the census counts in a quasi-random manner*" (Blake and Openshaw, 1994, page 2). The ED level has 711537 total individuals reported while the Ward level has a total of 718614, a difference of 7077. In this subsection Ward level data are created by summing the relevant ED level data. For these data, we have  $\bar{N}=7811.02$  and  $\bar{N}^*=7806.01$ . Table 5.14 shows the weighted mean and variance from the SAS (Ward level) and the aggregation effect defined as  $S^{(2)}/S^{(1)}$  for the variables being considered. The aggregation effect will be equal to 1 if there is no aggregation effect. From the results, it is evident that the aggregation effects that are much larger than the corresponding effects when the analysis is done at ED level.

Variables	weighted mean	weighted variance	aggregation effect ( $S^{(2)}/S^{(1)}$ )
age	0.6729	2.2213	59.3735
ftw	0.2948	2.0320	65.0882
unemp	0.0876	0.4022	39.5261
llti	0.1273	0.8544	35.5850
nocar	0.4832	8.3418	288.9365

Table 5.14: Weighted Mean and Variances from SAS (Ward level)

Table 5.15, shows the variance-covariance matrix at Ward level. The values are larger than the individual level and the ED level counterparts. There are covariances that change sign; the covariance of *age* and *unemp* have positive sign at individual level, positive at ED level but negative in Ward level. Another pair of variables is *unemp* and *llti*; negative at individual level, positive at ED level and Ward level. These changes affect all the statistics computed involving these pairs of variables.

Table 5.16 shows the correlations at Ward level. The absolute values of the

	age	ftw	unemp	llti	nocar
age	13.0502	12.0819	-0.8334	-5.0064	-12.6456
ftw		13.5123	-2.3247	-4.9404	-15.8028
unemp			3.2609	1.4193	8.2889
llti				4.0816	13.5355
nocar					71.8874

**Table 5.15: Variance-Covariance matrix Ward Level**

correlations of all pairs of variables are larger than the corresponding correlations at individual level except for one. This is the correlation of the pair *age* and *unemp*, where the sign changes.

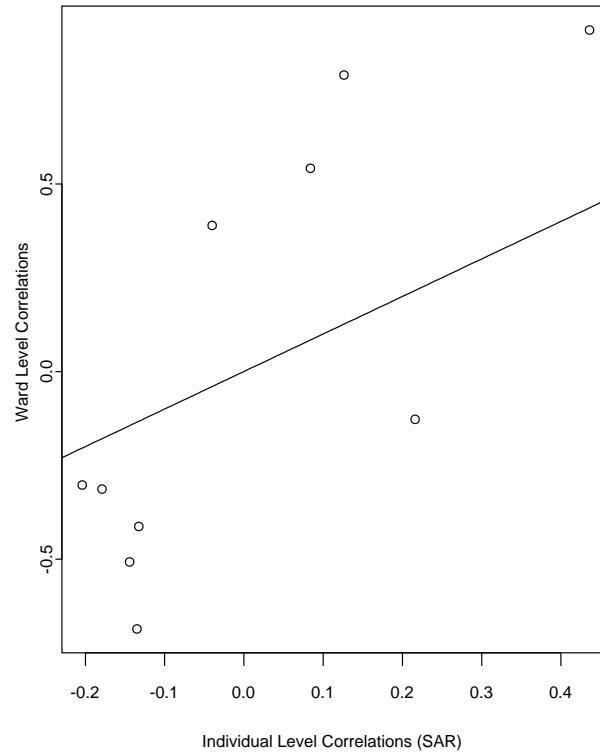
Figure 5.6 shows the graph for the Ward level correlations plotted against the individual level correlations. There are four correlations at Ward level that are greater than the corresponding individual level correlations, the rest are below the corresponding individual level. All of the Ward level correlations differ appreciably from their corresponding individual level correlations.

	age	ftw	unemp	llti	nocar
age	1.0000	0.9098	-0.1278	-0.6860	-0.4129
ftw		1.0000	-0.3502	-0.6652	-0.5070
unemp			1.0000	0.3890	0.5414
llti				1.0000	0.7902
nocar					1.0000

**Table 5.16: Correlations at Ward level**

Table 5.17 displays the variance components and the intra-Ward correlations. The level 1 variance component is much larger than the corresponding level 2 variance component for each variable, resulting in a very small intra-area correlations. The intra-area correlations are smaller than the corresponding ED level intra-area correlations as Wards are larger. However, the aggregation effect is larger because  $\bar{N}$  is larger.

Tables 5.18 shows the intra-Ward cross-correlations. As noted in the ED analysis, the intra-area cross-correlation is a measure of the within group homogeneity of a



**Figure 5.6: Individual Level Correlations versus Ward Level Correlation**

Variable	level1 variance	level2 variance	intra-area correlation
age	0.2181	0.0017	0.0076
ftw	0.2059	0.0017	0.0083
unemp	0.0821	0.0004	0.0050
llti	0.1142	0.0005	0.0045
nocar	0.2396	0.0093	0.0373

**Table 5.17: Variance components and Intra-area correlation**

pair of variables and that similarity of the values of two variables within areas can be measured using this statistic. The results shows that similarity within areas for the values of the pairs of variables are small.

Level 1 pure correlations are shown in Table 5.19. Note the similarity of the level 1 pure correlations to the corresponding individual level correlations. The signs of

	age	ftw	unemp	llti	nocar
age	*	0.0073	-0.0008	-0.0040	-0.0070
ftw		*	-0.0023	-0.0041	-0.0090
unemp			*	0.0019	0.0075
llti				*	0.0104
nocar					*

**Table 5.18: Intra-Ward Cross-Correlations**

the correlations coincide. This is due to the similarity of the values of the level 1 variance components, which is part of the computation of the pure correlation. Figure 5.6 shows the individual level correlation plotted against the corresponding level 1 pure correlations. The figure shows the similarity of the individual correlations computed from the SAR and the level 1 pure correlation from the simple multi-level model being considered.

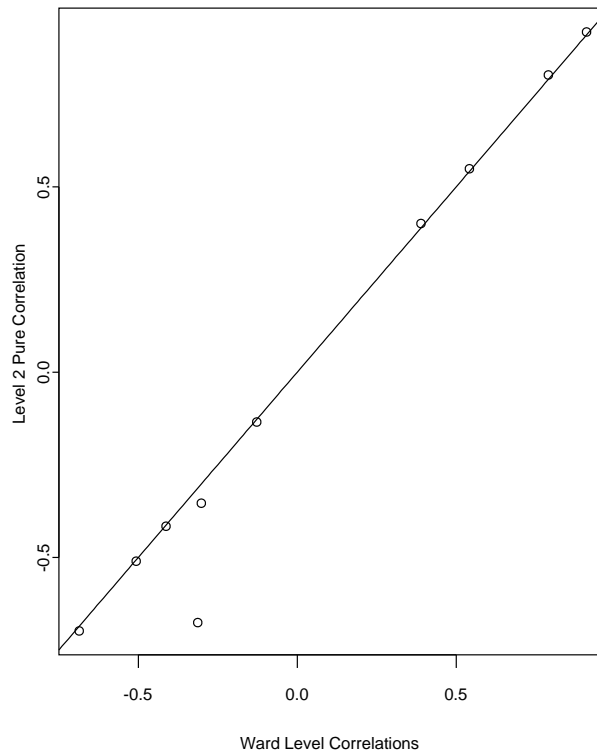
	age	ftw	unemp	llti	nocar
age	*	0.4322	0.2182	-0.1317	-0.1289
ftw		*	-0.2031	-0.1760	-0.1382
unemp			*	-0.0421	0.0845
llti				*	0.1276
nocar					*

**Table 5.19: Level 1 Pure correlation**

	age	ftw	unemp	llti	nocar
age	*	0.9176	-0.1351	-0.6987	-0.4161
ftw		*	-0.3534	-0.6762	-0.5108
unemp			*	0.4008	0.5485
llti				*	0.8016
nocar					*

**Table 5.20: Level 2 (Ward level) Pure correlation**

As before, the regression analysis will be in the latter part of the chapter.



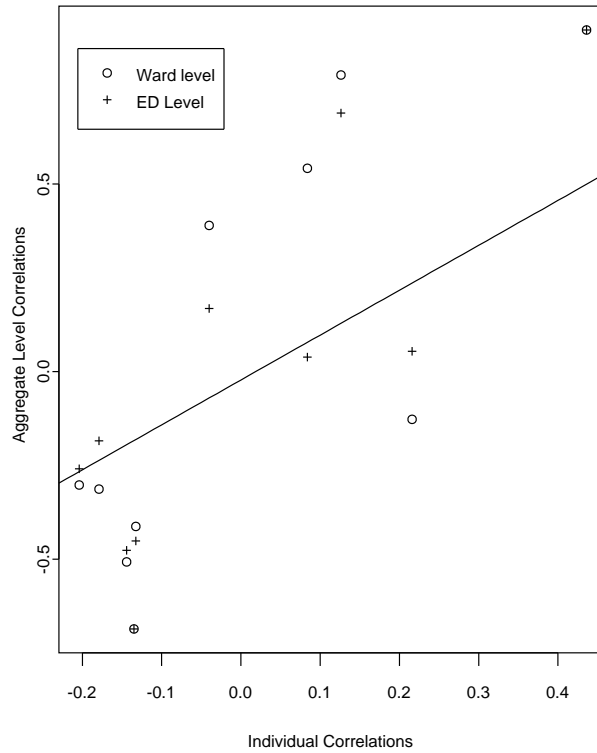
**Figure 5.7: Ward Level Correlations versus Level 2 Pure Correlations**

### Summary

Figure 5.8 shows the plot of individual level correlations versus the ED and Ward correlations together. The figure shows aggregation effects of the correlations at different levels of aggregation. The aggregation effects of the correlations at different levels do not display a predictable pattern. Some correlation increase with aggregation in both ED and Ward levels and in the aggregate levels others decrease.

Table 5.21 shows the correlation coefficients at different levels. Notice that not all of the correlations increase with aggregation. This shows again that correlation does not necessarily increase with aggregation. Some even change signs. The individual level correlations were computed from the SAR. The ED level and the Ward level correlations were computed from other source of data, the SAS. The computations are weighted according to the number of individuals included in each ED and Ward.





**Figure 5.8: Individual level correlations vs Ward and ED level correlations**

As mentioned earlier, to have consistency in the number of individuals for ED and Ward levels, the Ward level data are based on the ED level data. Comparing the correlations from ED level to Ward level, all the corresponding correlations increase in absolute values but one changes sign. The pair *age* and *unemp* have positive correlation at ED level, but negative at Ward level.

Figure 5.9 shows the individual level correlations plotted against the level 1 pure correlations for the simple two-level model where the level 2 is Ward and ED levels respectively. The figure shows that there is not much change for level 1 pure correlations and for both cases the aggregation effects are minimal.

Table 5.22 shows the level 1 *pure* correlation at different levels. In almost all cases the level 1 pure correlation are very similar to the correlation at individual level when ED and Ward levels are used as level 2. Figure 5.8 shows the individual level

	age	ftw	unemp	llti	nocar
Individual (SAR)					
age	1.0000	0.4356	0.2160	-0.1350	-0.1326
ftw		1.0000	-0.2040	-0.1791	-0.1442
unemp			1.0000	-0.0400	0.0838
llti				1.0000	0.1261
nocar					1.0000
ED Level (SAS)					
age	1.0000	0.7795	0.0638	-0.5547	-0.3586
ftw		1.0000	-0.1979	-0.4346	-0.3795
unemp			1.0000	0.1593	0.5075
llti				1.0000	0.5945
nocar					1.0000
Ward Level (SAS)					
age	1.0000	0.9098	-0.1276	-0.6860	-0.4129
ftw		1.0000	-0.3502	-0.6652	-0.5070
unemp			1.0000	0.3890	0.5414
llti				1.0000	0.7902
nocar					1.0000

Table 5.21: Correlations at Different Levels

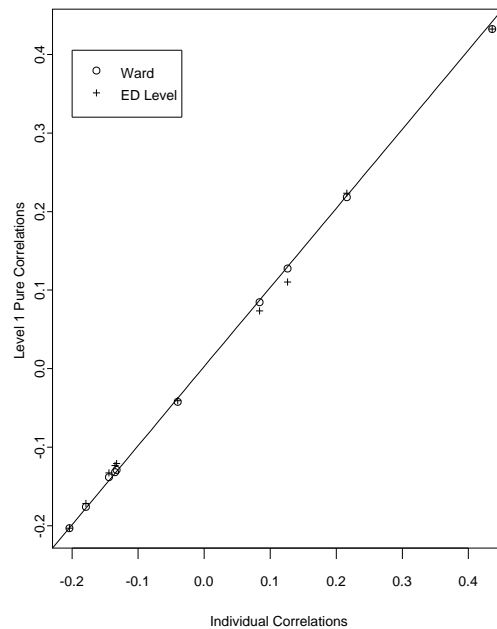


Figure 5.9: Individual level correlations vs Ward and ED Level 1 Pure correlations

correlations plotted against the corresponding Ward level and ED level correlations.

	age	ftw	unemp	llti	nocar
a. Correlation(SAR)					
age	1.0000	0.4356	0.2160	-0.1350	-0.1326
ftw		1.0000	-0.2040	-0.1791	-0.1442
unemp			1.0000	-0.0400	0.0838
llti				1.0000	0.1261
nocar					1.0000
b. Level 1 Pure Correlation(ED)					
age	*	0.4261	0.2192	-0.1251	-0.1226
ftw			-0.2044	-0.1731	-0.1344
unemp				-0.0433	0.0703
llti					0.1068
nocar					
c. Level 1 Pure Correlation Ward					
age		0.4322	0.2182	-0.1317	-0.1289
ftw			-0.2031	-0.1760	-0.1382
unemp				-0.0421	0.0845
llti					0.1276
nocar					

**Table 5.22: Correlations at Individual Level (from SAR) and Level 1 Pure Correlations when level 2 are ED and Ward Levels (fromSAS)**

Figure 5.10 shows the individual level correlations plotted against both the level 2 pure correlations. The figure clearly shows the aggregation effects on level 2 pure correlations which in both cases are far from the individual level correlations.

Table 5.23 shows the level 2 pure correlations. The level 2 pure correlation differ greatly in values in both ED and Ward Levels compared to the individual level correlation coefficient. Even sign of the correlations changed for one pair of variables.

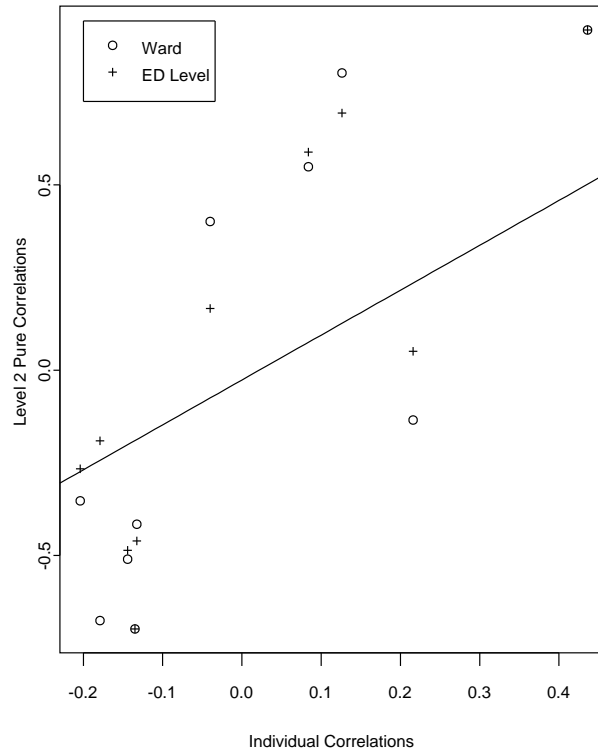


Figure 5.10: Individual level correlations vs Ward and ED Level 2 Pure correlations

### 5.1.3 Linear Regressions and Pure Regressions of Percentage of Full-Time Workers and Other variables

To look into the regression coefficients and the pure regression coefficients derived from the simple two-level model, the variable *ftw* is paired with the other variables namely: *age*, *liti*, and *nocar*. Thus, our dependent variable will be  $Y = \text{ftw}$  and the independent variables  $X$  will be the variables *age*, *liti*, and *nocar*.

#### 1. Ftw-Age

Table 5.24 shows the correlation and regression coefficients when the analysis is done at each level. An increase in the correlation and regression coefficients is observed as the number of zones is decreased.

	age	ftw	unemp	lti	nocar
a. Correlations Ward Level					
age	1.0000	0.9098	-0.1278	-0.6860	-0.4129
ftw		1.0000	-0.3502	-0.6652	-0.5070
unemp			1.0000	0.3890	0.5414
lti				1.0000	0.7902
nocar					1.0000
b. Level 2 Pure Correlations					
<i>WardLevel</i>					
age	*	0.9176	-0.1351	-0.6987	-0.4161
ftw		*	-0.3020	-0.6762	-0.5108
unemp			*	0.4008	0.5485
lti				*	0.8016
nocar					*

**Table 5.23: Correlations at Ward Level (from SAS) and Level 2 Pure Correlations when level 2 is Ward (from SAS) and level 1 is Individual (from SAR)**

age ~ ftw	correlation	regression
Individual	<b>0.4356</b>	<b>0.4237</b>
ED	0.7795	0.7647
Ward	0.9098	0.9257

**Table 5.24: Correlation and regression coefficients at different scales**

Table 5.25 displays some statistics derived from the simple multilevel model. Notice that the level 1 *pure correlation*, when ED level is considered as level 2 in the model, is almost equal to the Pearson correlation at the individual level and that the value goes nearer to the individual level when Ward is considered as the level 2 in the model. The level 2 *pure correlation* is larger than the individual level Pearson correlation. The level 1 and level 2 *pure regression* coefficient display characteristics similar to the level 1 and level 2 *pure correlation*. Less aggregation effect are observed for pure coefficient when going from ED to Ward but an aggregation effect is still present.

ftw $\sim$ age	Pure Correlation	Pure Regression
Level 1 (Individual)	<b>0.4266</b>	<b>0.4143</b>
Level 2 (ED Level)	0.8401	0.8021
Level 1 (Individual)	<b>0.4322</b>	<b>0.4198</b>
Level 2 (Ward Level)	0.9176	0.9344

Table 5.25: Some Statistics derived from multilevel model

## 2. Ftw-Llti

Table 5.26 shows the correlation and regression coefficients at different levels of analysis. The values increase in absolute values. Again, looking at the results in Table 5.27, the level 1 *pure correlation* and *pure regression* coefficients have values almost equal to the initial Pearson correlation and the regression coefficients, respectively. Level 2 *pure correlation* and *pure regression* coefficients have values different from the initial Pearson correlation and the regression coefficients, respectively. When the Ward level is considered a level 2 in the model being considered, the value is nearer to the initial characteristics of the individual level data. The level 2 pure coefficient correlations have values similar to the corresponding aggregate level correlations and the level 2 pure regression have values not far from the aggregate regression coefficients.

ftw $\sim$ llti	correlation	regression
Individual	<b>-0.1791</b>	<b>-0.2408</b>
ED	-0.4346	-0.6702
Ward	-0.6652	-1.2104

Table 5.26: Correlation and regression coefficients at different scales

ftw $\sim$ llti	Pure Correlation	Pure Regression
Level 1 (Individual)	<b>-0.1731</b>	<b>-0.2321</b>
Level 2 (ED Level)	-0.4691	-0.7378
Level 1 (Individual)	<b>-0.1760</b>	<b>-0.2363</b>
Level 2 (Ward Level)	-0.6762	-1.2384

Table 5.27: Some Statistics derived from multilevel model

### 3. Ftw-Nocar

Table 5.28 shows the correlation and regression coefficients at different levels. The Pearson correlation and regression coefficient increase in absolute values. Table 5.29 shows statistics derived from the simple two-level multilevel model.

The level 1 pure correlations and regressions are similar to the individual level correlations and regressions, respectively. The level 2 *pure* correlations and regressions are similar to the values of the ED level and Ward level Pearson correlation and regression coefficients, respectively.

ftw $\sim$ nocar	correlation	regression
Individual	<b>-0.1442</b>	<b>-0.1315</b>
ED	-0.3795	-0.1878
Ward	-0.5070	-0.2089

**Table 5.28: Correlation and regression coefficients at different scales**

ftw $\sim$ nocar	Pure Correlation	Pure Regression
Level 1 (Individual)	<b>-0.1334</b>	<b>-0.1260</b>
Level 2 (ED Level)	-0.3981	-0.1890
Level 1 (Individual)	<b>-0.1382</b>	<b>-0.1281</b>
Level 2 (Ward Level)	-0.5108	-0.2201

**Table 5.29: Some Statistics derived from multilevel model**

### Summary

Level 1 pure correlations and regressions have values similar to the direct coefficient obtained from individual level data. The values differ by a small fraction and in all cases, it seems that the values go nearer to the individual level statistics as the number of zones is decreased.

The level 2 pure correlations and regressions coefficient have values similar to the correlation and regression coefficients at the aggregate level.

The Pearson correlation and the regression coefficients increase in absolute value as the number of zones is decreased.

## 5.2 Data from one source (SAS)

### 5.2.1 Some Statistics from 1991 UK Census

To further examine the behavior of the simple multilevel model under aggregation another experiment is conducted. This time the data involves no individual level data. The same variables as those used in the analysis of the previous section were analyzed. The same districts were considered. This time the percentages of the variables as defined previously are from enumeration districts (EDs) and are considered as the level 1 data and for level 2 the percentage from the Ward level are used. The Ward level data are derived from the original ED level data.

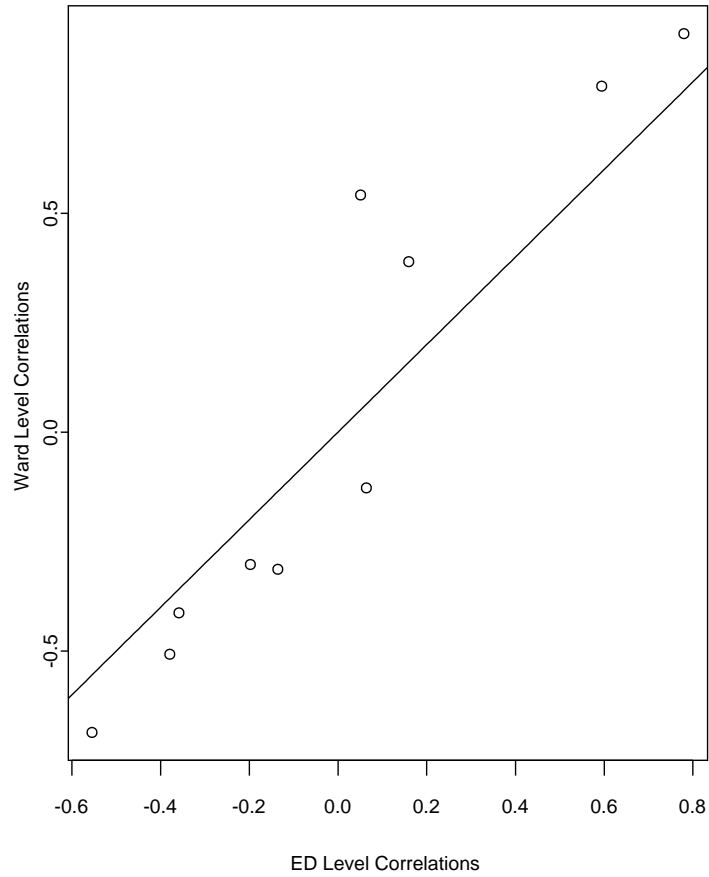
	age	ftw	unemp	liti	nocar
ED Level					
age	1.0000	0.7995	0.0638	-0.5547	-0.3586
ftw		1.0000	-0.1979	-0.4661	-0.3589
unemp			1.00000	0.1649	0.4862
liti				1.0000	0.6005
nocar					1.0000
Ward Level					
age	1.0000	0.9098	-0.1278	-0.6860	-0.4129
ftw		1.0000	-0.3502	-0.6652	-0.5070
unemp			1.0000	0.3890	0.5414
liti				1.0000	0.7902
nocar					1.0000

**Table 5.30: Pearson Correlations at ED and Ward Levels**

Table 5.30 shows the Pearson correlation coefficients at ED and Ward levels. All correlation coefficients either increase when the values are positive and increase in absolute values when the correlations are negative except for correlation between Age and Unemp where the value change from positive at ED level and negative at Ward level. Figure 5.11 shows the plot of ED level correlations against the Ward level correlations. The plot shows that the ED level correlations changes with aggregation, some increase and some decrease, but they are of the same sign and greater in absolute value.

Table 5.31 shows the level 1 and level 2 *pure correlation*. The values of the





**Figure 5.11: ED level correlations vs. Ward level correlations**

entries in the table are similar to the corresponding Pearson correlations at each level. All except one of the level 1 pure correlations is less in absolute value than the corresponding ED level Pearson correlation. This is the case when variables Age and Unemp is analyzed. All the level 2 pure correlation are greater in absolute value than the corresponding Ward level correlations. Figure 5.12 shows the plot of ED level correlations against level 1 and level 2 pure correlations.

Table 5.32 shows the aggregation effects on the variances and the covariances of the pairs of variables. The positive sign of a covariance means that the sign of a covariance is the same at the ED level and Ward level. There is one negative aggreg-

	age	ftw	unemp	llti	nocar
Level 1					
age	1.0000	0.7526	0.1613	-0.5108	-0.3351
ftw		1.0000	-0.1145	-0.3507	-0.3082
unemp			1.0000	0.0618	0.4848
llti				1.0000	0.5216
nocar					1.0000
Level 2					
age	1.0000	0.9312	-0.1606	-0.7176	-0.4232
ftw		1.0000	-0.3746	-0.7176	-0.5272
unemp			1.0000	0.4365	0.5460
llti				1.0000	0.8343
nocar					1.0000

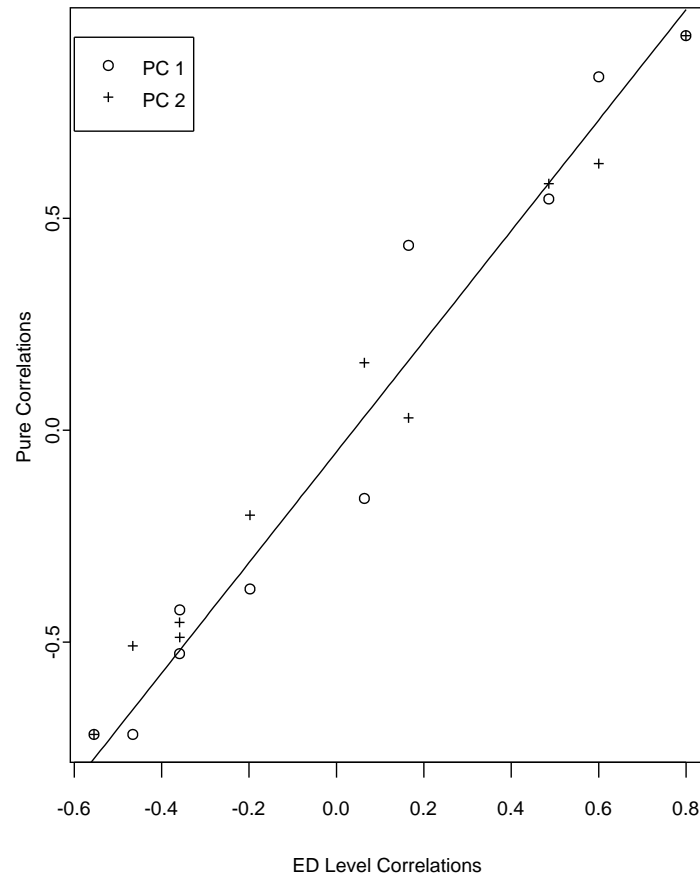
Table 5.31: Pure Correlations

gation effect, which means that the sign of the covariance changes. In this particular case, it is the pair *age* and *unemp*; the covariance changes from positive at ED level to negative at Ward level. Looking at the table and disregarding the sign of the aggregation effects on the variances and covariance, in most cases, the aggregation effect of the variances of the pairs of variables are lesser than the corresponding covariances, in effect, the corresponding correlations increase with aggregation. In some cases, one of the aggregation effects of the variance is smaller than the corresponding covariance, but the aggregation effect on the variance of the other variable is greater than the corresponding covariance resulting in increase of the correlation at Ward level.

	age	ftw	unemp	llti	nocar
age	<b>5.8751</b>	7.1130	-13.8206	6.5514	8.1921
ftw		<b>6.6498</b>	12.9937	8.6273	10.1142
unemp			<b>8.1077</b>	15.1986	8.9168
llti				<b>4.7772</b>	8.5283
nocar					<b>8.6177</b>

Table 5.32: Aggregation effects on the variances (diagonal, bold) and covariances (off-diagonal)

The intra-area correlation was computed using the Tranmer and Steel (1998) method. Table 5.33 shows the variance component and the intra-area (intra-Ward) correlation. The level 1 variance component is less than the level 2 variance com-



**Figure 5.12: ED level correlations vs. Level 1 and Level Pure Correlations**

ponents for each of the variables. The intra-area correlation of the variable varies between 0.19 and 0.39.

Variables	Level 1	Level 2	Lev 2 Var Com	Level 1 Var Com	IAC
Age	ED	Ward	1.6710	0.5503	0.2477
Ftw	ED	Ward	1.4486	0.5835	0.2871
Unemp	ED	Ward	0.2569	0.1453	0.3611
Llti	ED	Ward	0.6904	0.1640	0.1919
Nocar	ED	Ward	5.1128	3.2300	0.3871

**Table 5.33: Variance component and Intra-Ward Correlation**

Table 5.34 shows the intra-area cross-correlation of the variables. The values on

the diagonal are equivalent to the intra-area correlations of the variables. Intra-area cross-correlation is a measure of within-area homogeneity for a pair of variables. These values represent the similarities of the values of two different variables within areas. The values ranges from -0.0480 to 0.2483. By (3.40) the pure level 2 correlation is the intra-area cross correlation divided by the square root of the relevant intra-area correlations.

	age	ftw	unemp	llti	nocar
age	<b>0.2477</b>	0.2483	-0.0480	-0.1565	-0.1311
ftw		<b>0.2871</b>	-0.1206	-0.1684	-0.1758
unemp			<b>0.3612</b>	0.1149	0.2042
llti				<b>0.1919</b>	0.2274
nocar					<b>0.3871</b>

Table 5.34: Intra-Area Cross Correlation

### 5.2.2 Regression Analysis of Variable Ftw and Other variables

Similar to subsection (5.1.3) the variable *ftw* was paired with the other variables in bivariate regression analyses; *ftw* is the dependent variable and the independent variables are *age*, *llti*, and *nocar*. Table 5.35 shows the estimates of the regression coefficients computed at each level. In all pairs of variables, the values increase in absolute values from ED level to Ward level.

Table 5.36 shows pure regression coefficients when ED level is the level 1 and Ward level is the level 2 for the model. There level 1 pure regression coefficients are of the same sign and smaller than the ED level regression coefficient, whereas the level 2 pure regression coefficients are slightly larger in absolute values than the Ward level regression coefficients.

### 5.2.3 Spatial autocorrelation of the Variables

The role of autocorrelation is evident from the results based on simulated data in chapter 4. Here we examine the evidence on spatial autocorrelation in the real data

	Regression Coefficient
Ftw $\sim$ Age	
ED Level	0.7647
Ward Level	0.9258
Ftw $\sim$ Llti	
ED Level	-0.6702
Ward Level	-1.2104
Ftw $\sim$ Nocar	
ED Level	-0.1873
Ward Level	-0.2198

**Table 5.35: Regression coefficients at ED and Ward levels**

	Pure regression
Ftw $\sim$ Age	
Level 1	0.7007
Level 2	0.9588
Ftw $\sim$ Llti	
Level 1	-0.5080
Level 2	-1.3534
Ftw $\sim$ Nocar	
Level 1	-0.1641
level 2	-0.2241

**Table 5.36: Level 1 and Level 2 Pure Regression Coefficients**

used in this chapter. As the SAR does not give geographic indicators below the ED level, it is not possible to analyze spatial autocorrelation within ED in detail, although it is possible to assess the average level of within ED spatial autocorrelation using the intra-area correlation, which in effect equivalent to Moran's I with block proximity weights. Spatial autocorrelation at ED and Ward levels can be directly analyzed since the geographic location of these units are available.

Table 5.37 shows the degree of autocorrelation as measured using Moran's I statistic for the ED level. Several types of proximity weights were used. Lag 1 denotes that each ED is a neighbor to each immediate surrounding EDs. Lag 2

means that each ED is a neighbor to each immediate surrounding EDs and the next immediate surrounding EDs, and so on. The results in column 6 of Table 5.37 are done using a ‘*block*’ proximity matrix in which EDs within a Ward are considered neighbors. The computations are done using GeoDa, a freeware developed by Luc Anselin and co-workers. The software ‘*is designed implement techniques for exploratory spatial data analysis (ESDA) on lattice data (points and polygons)*’ (<http://sal.agecon.uiuc.edu/cssi/geoda.html>).

Variable	lag 1	lag 2	lag 3	lag 4	ED w/in Ward
age	0.3425	0.2262	0.1681	0.0764	0.2270
ftw	0.3640	0.2589	0.2074	0.1251	0.2614
unemp	0.4326	0.3771	0.3374	0.2679	0.3586
liti	0.2659	0.1872	0.1441	0.0877	0.1782
nocar	0.5288	0.4175	0.3607	0.2591	0.3967

**Table 5.37: Moran’s I at different weight definition**

Table 5.38 show the Moran’s I at Ward level with different definitions of proximity matrices.

Variable	Lag 1	Lag 2
Age	0.2192	<b>0.1006</b>
Ftw	0.3872	<b>0.2013</b>
Unemp	0.6650	<b>0.4574</b>
Liti	0.3817	<b>0.2330</b>
Nocar	0.5711	<b>0.3336</b>

**Table 5.38: Moran’s I with different proximity matrices Ward level**

Tables 5.39 shows the bivariate Moran’s I with block proximity. The values in the diagonal are Moran’s I for the variables and the off-diagonals are the bivariate Moran’s I for the pairs of variable. The Bivariate Moran’s I measures the degree of spatial association of two variables. The values are very similar to the corresponding intra-area cross-correlations. In fact, there is almost perfect linear correlations between the statistics (0.999) as shown in figure 5.13.

When the block proximity matrix used: (1) The Intra-area correlation is equal to the Moran’s I and (2) Intra-Area Cross-Correlation is equal to the Bivariate Moran.

	age	ftw	unemp	llti	nocar
age	0.2270	0.2292	-0.0540	-0.1459	-0.1228
ftw		0.2614	-0.1148	-0.1571	-0.1617
unemp			0.3586	0.1141	0.1992
llti				0.1782	0.2227
nocar					0.3967

Table 5.39: Bivariate Moran using GeoDa (EDs within Ward)

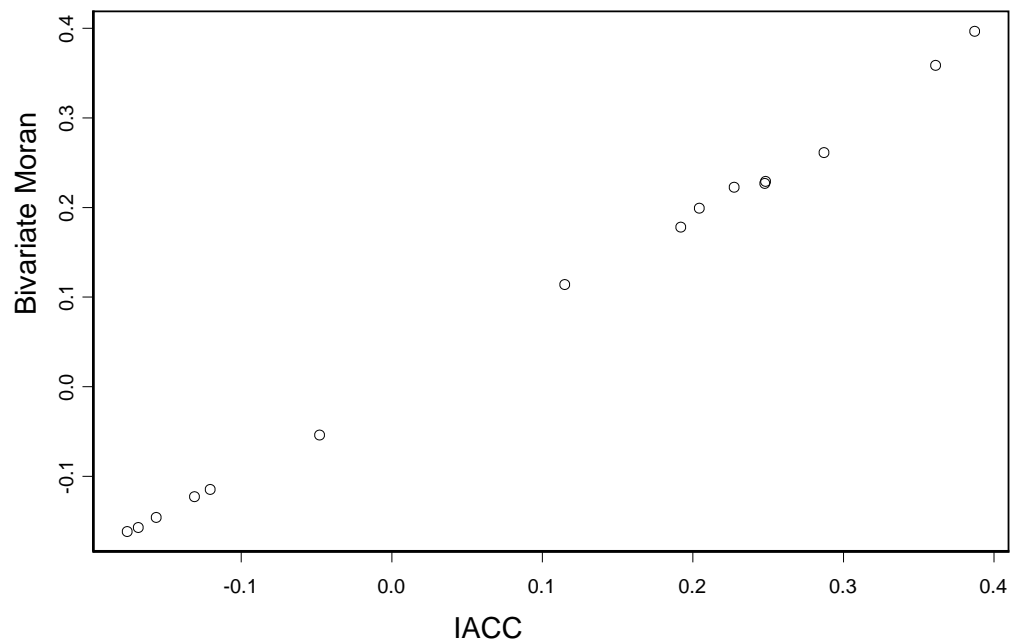


Figure 5.13: Bivariate Moran vs. Intra-area Cross-correlation

The intra-area -correlation is a special type of measure of spatial autocorrelation using the block proximity matrix.

## Pictorial Representation of the 5 Variables

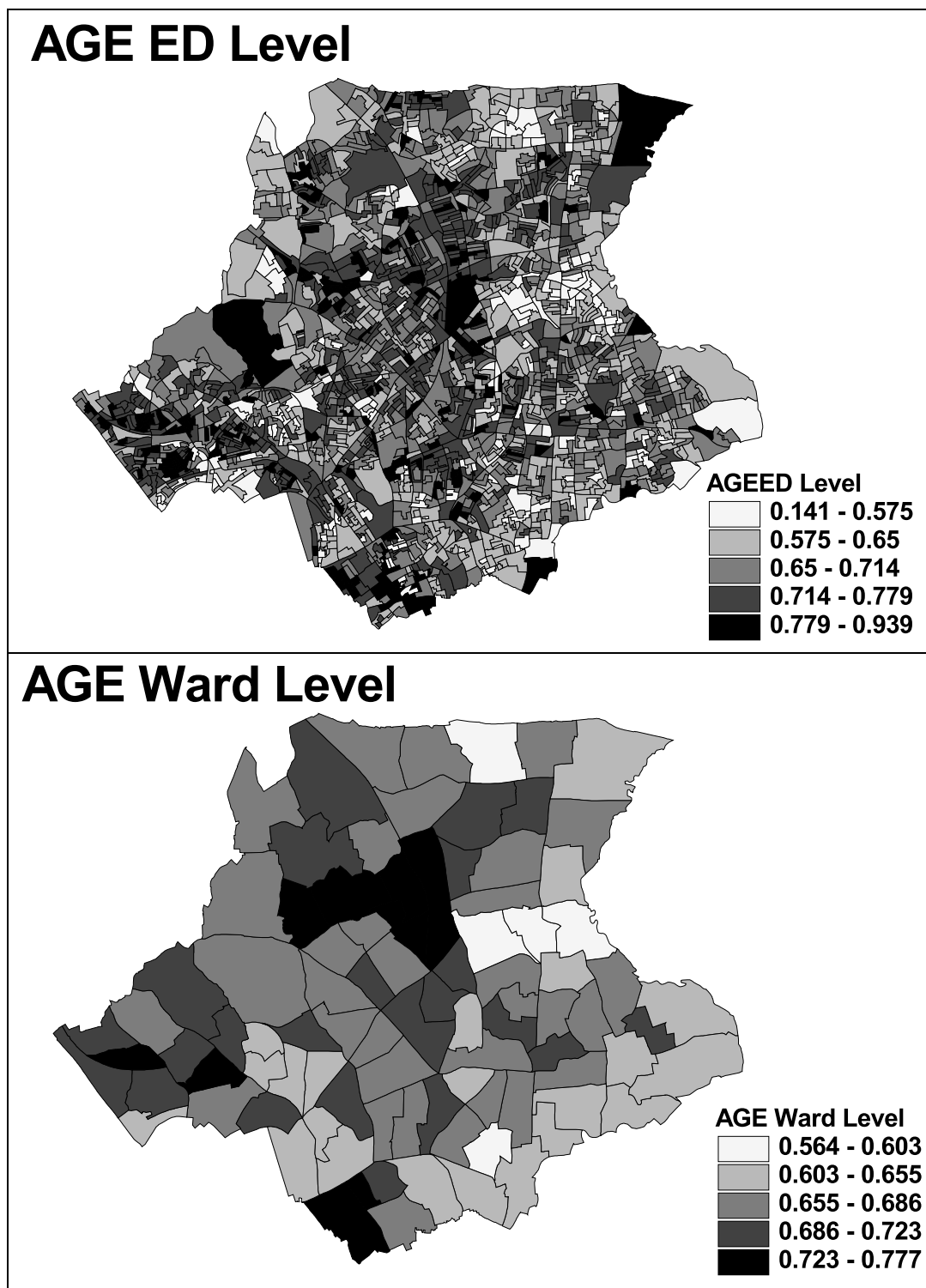


Figure 5.14: Graphical representation of the variable AGE at different levels



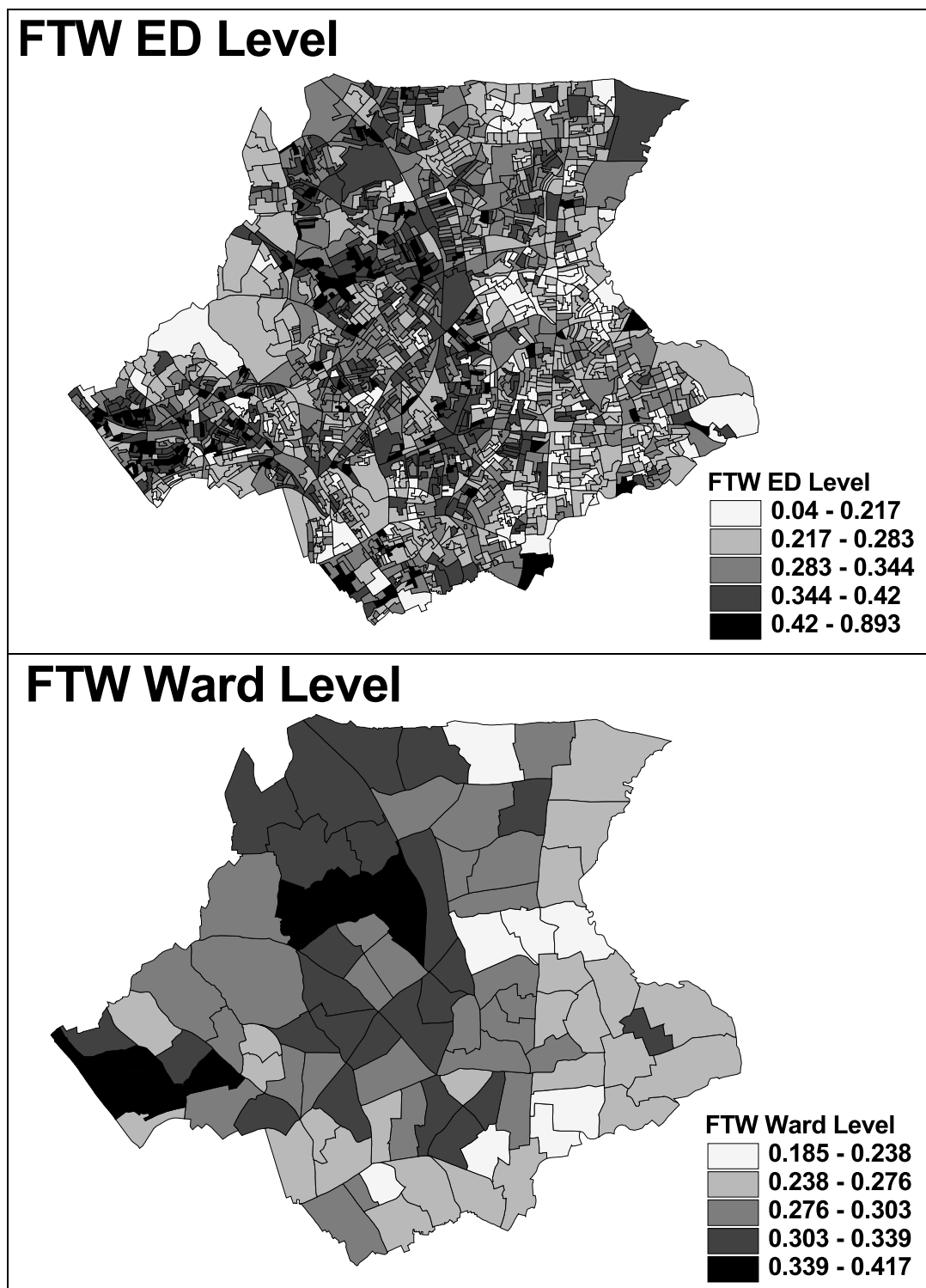


Figure 5.15: Graphical representation of the variable FTW at different levels

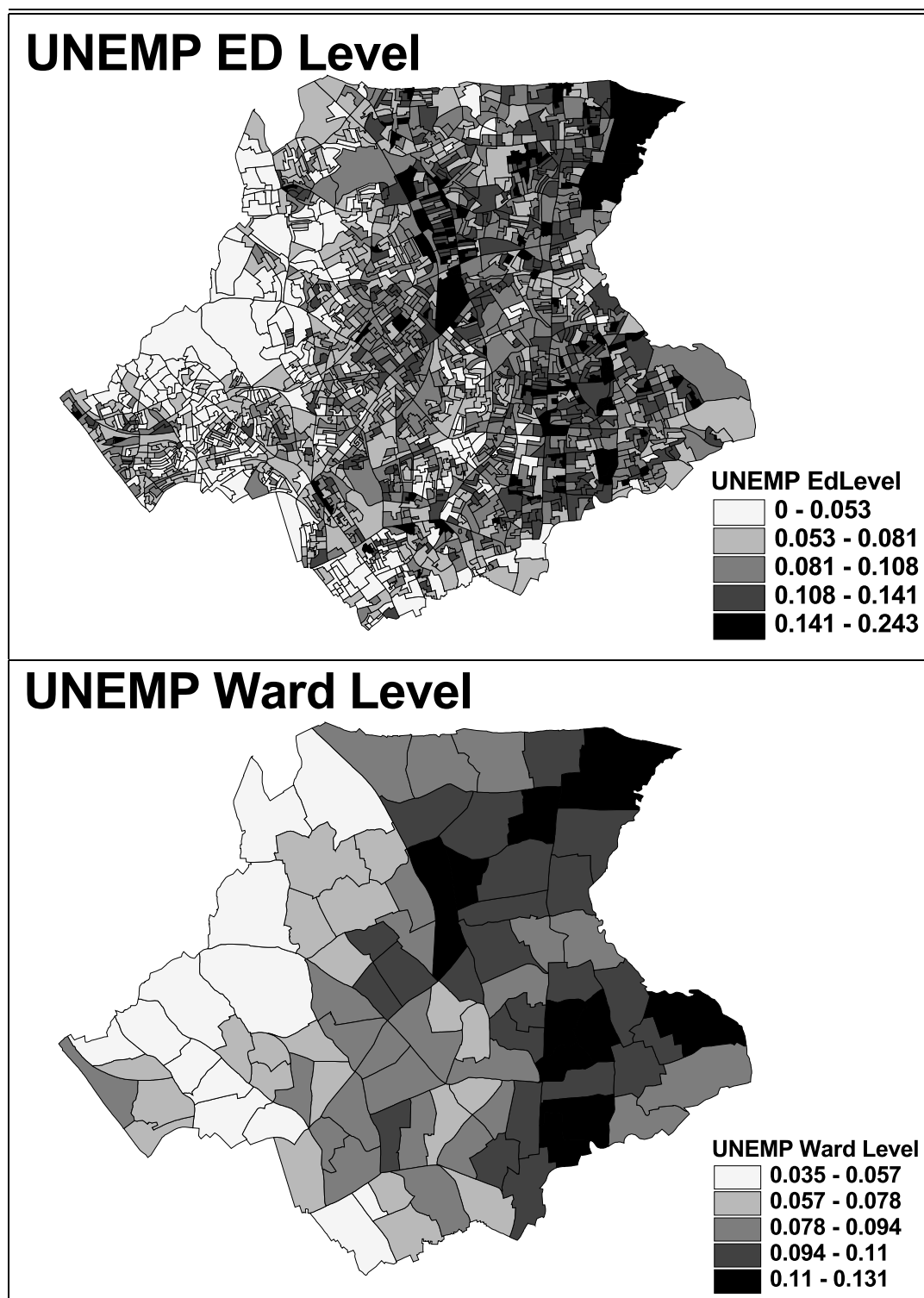


Figure 5.16: Graphical representation of the variable UNEMP at different levels

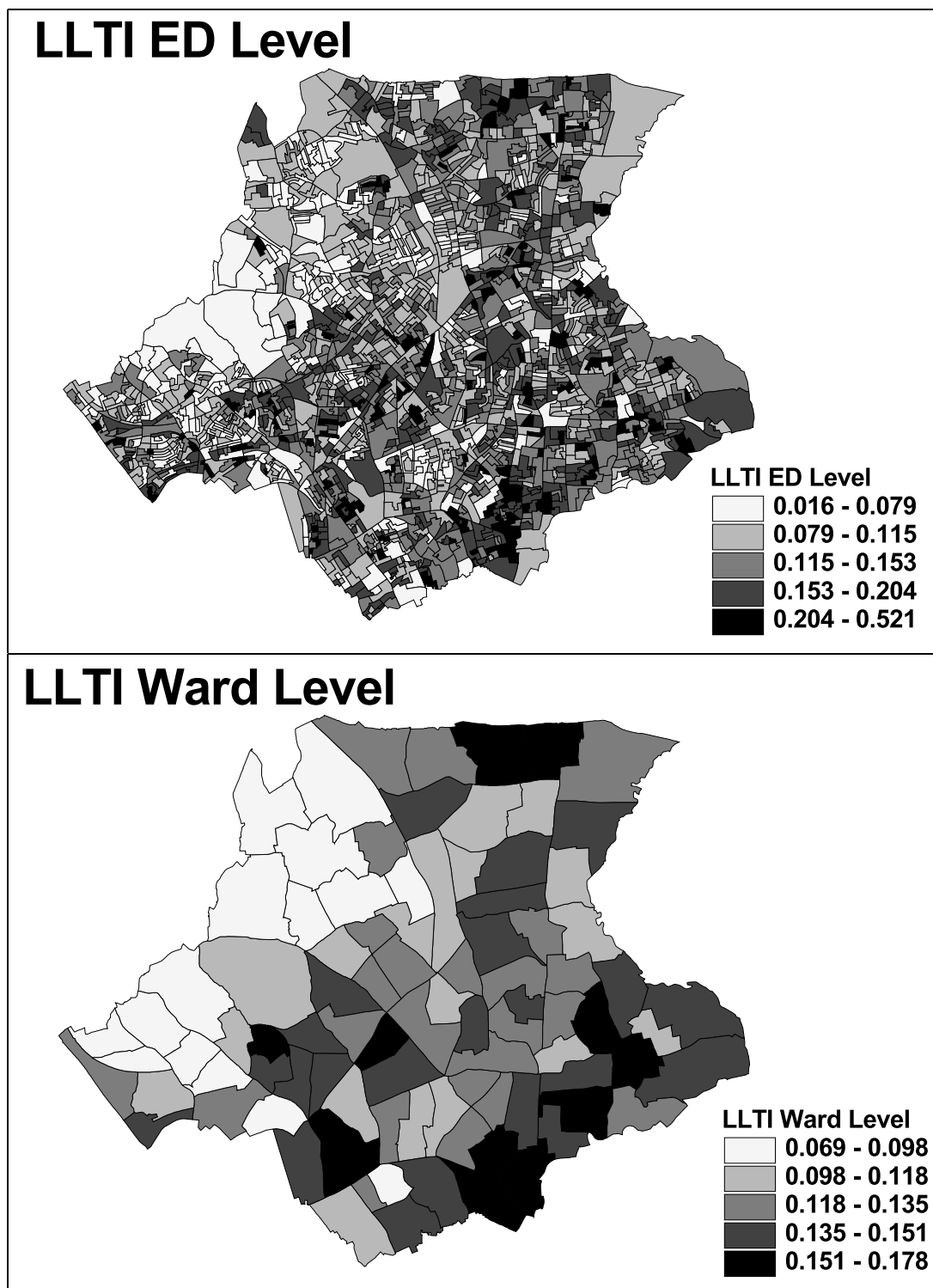


Figure 5.17: Graphical representation of the variables at different levels

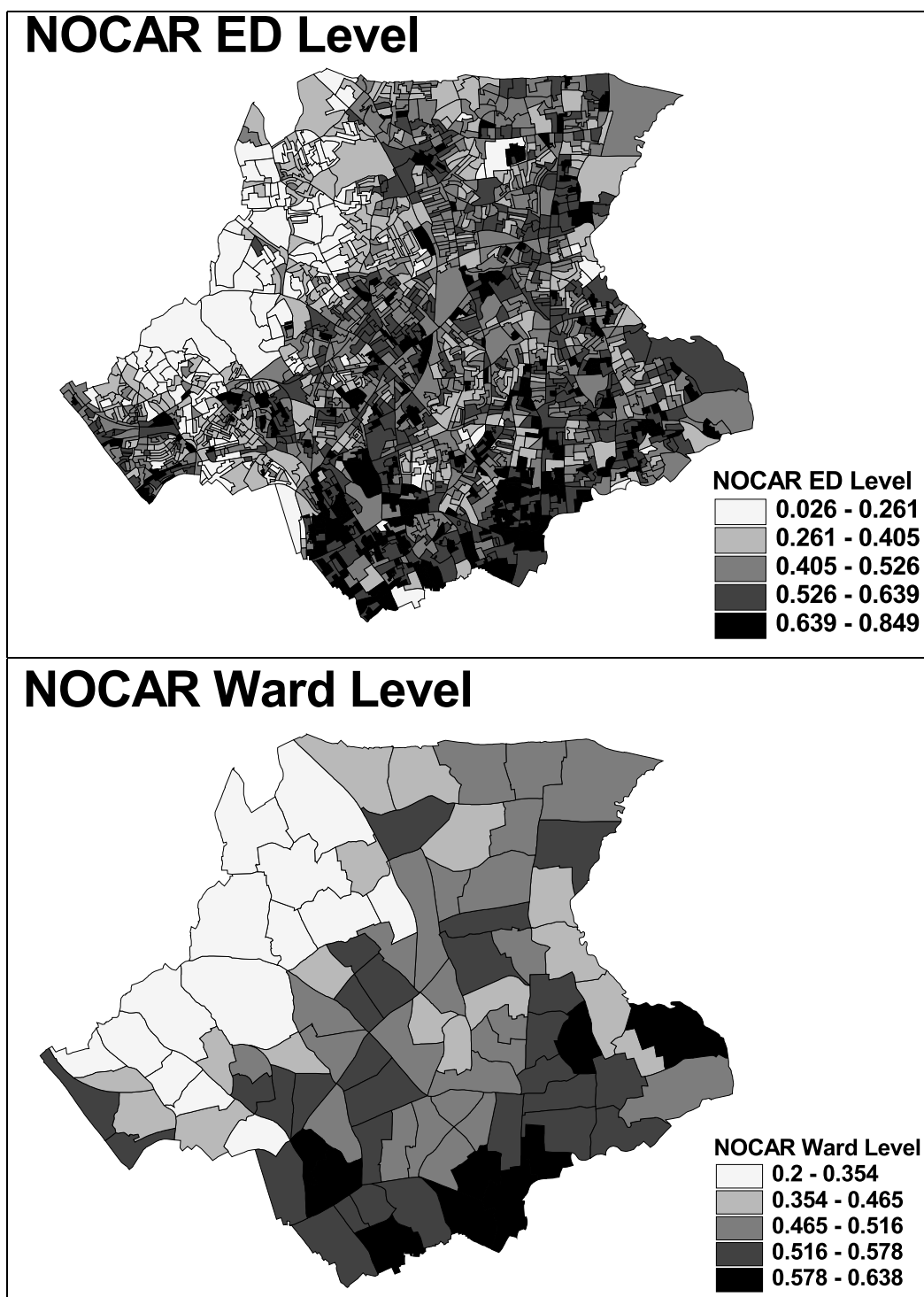


Figure 5.18: Graphical representation of the variable NOCAR at different levels

### 5.3 Summary

To investigate the aggregation effects of common statistics actual data from the 1991 UK Census is used. Three levels of data are considered in this analysis; individual level, Enumeration District (ED) level, and the Ward level. The individual level data are taken from the 1991 SARs (Samples of Anonymised Records) that correspond to a two percent sample of individuals counted in households and communal establishments in Great Britain. The geographical indicator available is only the district level to protect the confidentiality of information. The SAR Districts considered in the study are Camden, Hackney, Haringey, and Islington and are part of London boroughs. For the aggregate level data, both the ED level and the Ward level of the UK population census are extracted from the Small Area Statistics (SAS) data base. Five variables were considered: **age** (percentage of individuals between 16 and 65, inclusive), **ftw** (percentage of full-time workers), **uemp** (percentage of unemployed), **liti** (percentage of individuals with limiting long term illness), **nocar** (percentage of individuals with no car).

There are three cases considered in this chapter, namely: Case 1 the individual level data are from the SARs and second level data are from the ED level obtained from the SAS; Case 2 the individual level data are from the SARs and the second level data are from Ward the level obtained from the SAS; and Case 3 the ED level and Ward level aggregate data both from the SAS.

Since there are five variables, there are ten possible correlations calculated from all possible pairs of variables. The correlations increased in absolute values in going from individual level to ED level except the correlation between *age* and *unemp*. The decrease of the correlation between two variables can be explained by the aggregation effects of the variances and covariance. The variance aggregation effect can be defined as the ratio  $S_{YY}^{(2)}/S_{YY}^{(1)}$  and the covariance aggregation effect is  $S_{YX}^{(2)}/S_{YX}^{(1)}$  (Steel, et. al. (1996)). If the aggregation effects of the variances of both variables is smaller than the aggregation effect on the covariance, the result is an increase in the correlation coefficient.

The aggregation effects of the correlations at different levels do not display a predictable pattern. Some correlations increase with aggregation in both ED and Ward levels and at the aggregate levels others decrease. Not all of the correlations increase with aggregation. This shows again that correlation does not necessarily increase with aggregation. Some even change signs. However, if we look at the absolute values and signs we see a different picture. Generally the correlation increase in absolute value, but can change sign.

In both Case 1 and Case 2 level 1 pure correlations and regressions have values similar to the correlations from individual level. Also the level 2 pure correlations and regressions have values similar to the correlation and regression coefficients at the aggregate level. The Pearson correlation and the regression coefficients increase in absolute value as the number of zones is decreased.

In the third case when the data are from one source, all correlation coefficients either increase when the values are positive and increase in absolute values when the correlations are negative, except for correlation between age and llti where the value change from positive at ED level and negative at Ward level.

The level 1 and level 2 pure correlation were similar to the corresponding Pearson correlations at each level. All except one of the level 1 pure correlation are greater in absolute value than the corresponding ED level Pearson correlation. This is the case when variables age and unemp is analyzed.

In terms of the relationship between the intra-area correlation and the spatial autocorrelation, the results shows that the Moran's I with block proximity equal the corresponding intra-area correlation at one decimal point.

These results suggest that in applied setting it is important to calculate the aggregation effects for the set of variables of interest for the scales of analysis being considered. Even if individual level data are not available to calculate covariances at the individual level, it is often possible to calculate estimates of the individual level variances. The aggregation effects can be used to obtain estimates of the intra-area correlation for each variable at each scale. The results here emphasize that even small intra-area correlations can lead to major aggregation effects when

the population sizes of the areal units are large. The results demonstrate that in practice level 2 correlations need to be interpreted as reflecting relationships at that level and have almost nothing to do with individual level relationships. They also suggest that correlation should be calculated at the level of substantive interest, as they are specific to the level used in the calculation. In an application the spatial autocorrelation of the variables can be analyzed using the estimates of the intra-area correlation, which reflect the average with-in areal unit spatial autocorrelation. The spatial autocorrelation at the aggregate level can be analyzed directly.

## Chapter 6

# Data generation based on actual boundaries

This chapter describes a second series of simulation experiments to look into the characteristics of the relevant statistics considered in this thesis. While the values of the variable are simulated, the areal units are real and so reflect at least some of the complexity of the real world. The ED is used as the lowest level unit, effectively acting as the individual level in the experiments.

### 6.1 Data Set Generator

To look deeper into the behavior of the distribution of pure statistics and other common statistics, data sets are constructed with pre-determined characteristics based on the ED and Ward boundaries used in Chapter 5. The data set generator used in this study is based on Reynolds (1998). One desirable property of the data set generator is that *‘it allows the user to create a set of variables with specific levels of spatial autocorrelation (as measured by the Moran Coefficient) and Pearson correlations’* (Reynolds, 1998, pp.10).

Aside from creating data sets with specified autocorrelation, means, and variances of variables, Reynold’s data set generator can also generate *‘the entire matrix of correlations between the variables’* (Reynolds and Amrhein, 1997). However, there



are certain combinations of Moran's I and Pearson correlations that are not possible. For a more detailed description of the data set generator, refer to Reynolds (1998, pp. 10-17) and Reynolds and Amrhein(1997).

The principles behind the Reynolds' data generator are used to generate data based on an actual region that is divided into enumeration districts (ED), the lowest geographical level in the 1991 UK population census for which aggregate data are released. These EDs are grouped into larger geographical areas called Wards. The data generator is used to produce the values of the variables and the areal units are the EDs and the Wards of the real data set described in Chapter 5. This enables us to examine the properties of various statistics and analyses using more realistic spatial structures than the data considered in Chapter 4. The region in Figure 6.1 is composed of the districts; Camden, Hackney, Haringey, and Islington. It comprises 1904 EDs nested into 92 Wards, so that the average number of EDs per Ward is 20.7.

## The Region

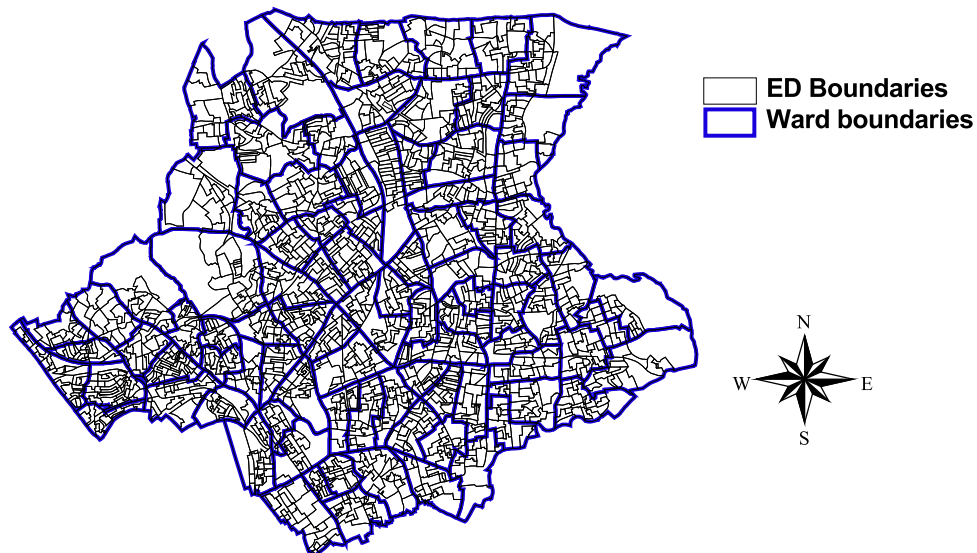


Figure 6.1: The region with its boundaries

Based on the boundaries of the EDs sets of data are generated with some specified properties. Figure 6.2 presents some realizations of variable X using the data set

generator. Light colors indicate small values and, as values become larger the color approaches black. The Moran's I is denoted by  $M_o$  at ED level and uses the weight matrix corresponding to 'block' proximity, that is,  $w_{ij}=1$  if ED  $i$  and ED  $j$  are in the same Ward, and 0 otherwise. Looking at the figure, when the  $M_o$  at ED level is 0.002, it seems that the distribution is random. A pattern emerged when the Moran's I is 0.1 as it seems that clusters form, especially the larger values. As the degree of autocorrelation increases, clustering of values is observed. This clustering is not unique, given a specific degree of autocorrelation different patterns emerged.

Figure 6.3 shows different clustering at autocorrelation equal to 0.8 as measured by the Moran's I ( $Mo$ ). We will examine the distributions of some statistics for different degrees of autocorrelation. Holding a specific degree of autocorrelation constant, as well as some other properties such as the mean, variance, and initial Pearson correlation will enable us to observe the aggregation effects of statistics such as intra-area correlation, intra-area-cross correlation, pure regression, pure correlation and some other statistics.

## 6.2 Case 1: Variables have the same spatial autocorrelation

A set of data is generated with the same mean and variances, and specific Pearson correlation for different degrees of autocorrelation. The first data set generated is composed of two variables X and Y and to make the analysis simple, the variables have the same mean (40) and the same variance (16) at the ED level. The variables also have a fixed Pearson correlation equal to 0.3 but they are generated in such a way that the autocorrelation of the variables are equal but varies (0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) as measured by Moran's I and uses the weight matrix corresponding *queen's case lag 1*, ie  $w_{ij}=1$  if ED  $i$  and ED  $j$  are immediate neighbors and 0 otherwise. To look at the distribution of the direct and pure statistics and other statistics, the generation of the data is repeated 3000 times for each degree of autocorrelation. The data sets are then filtered to select those data sets that

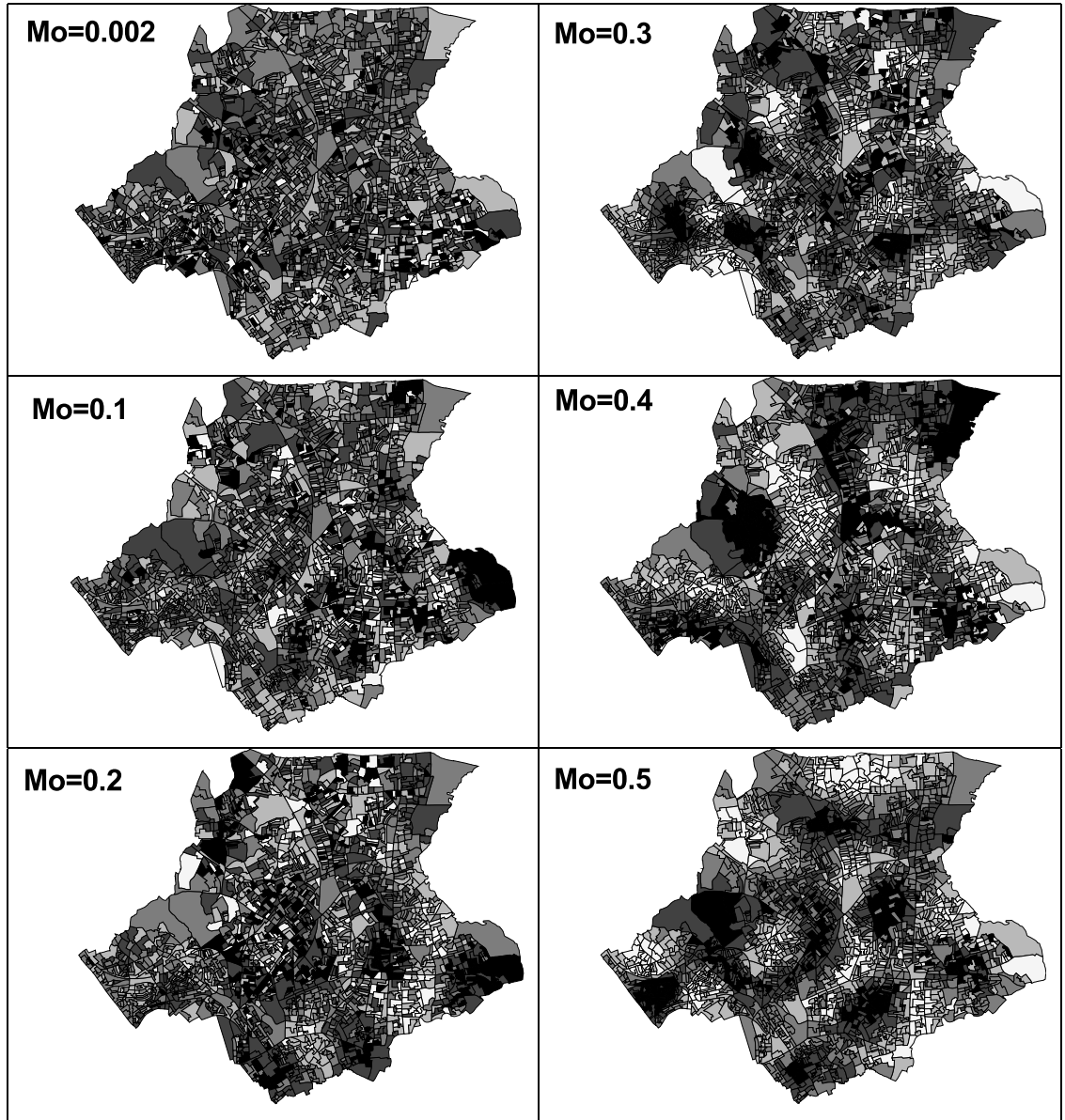
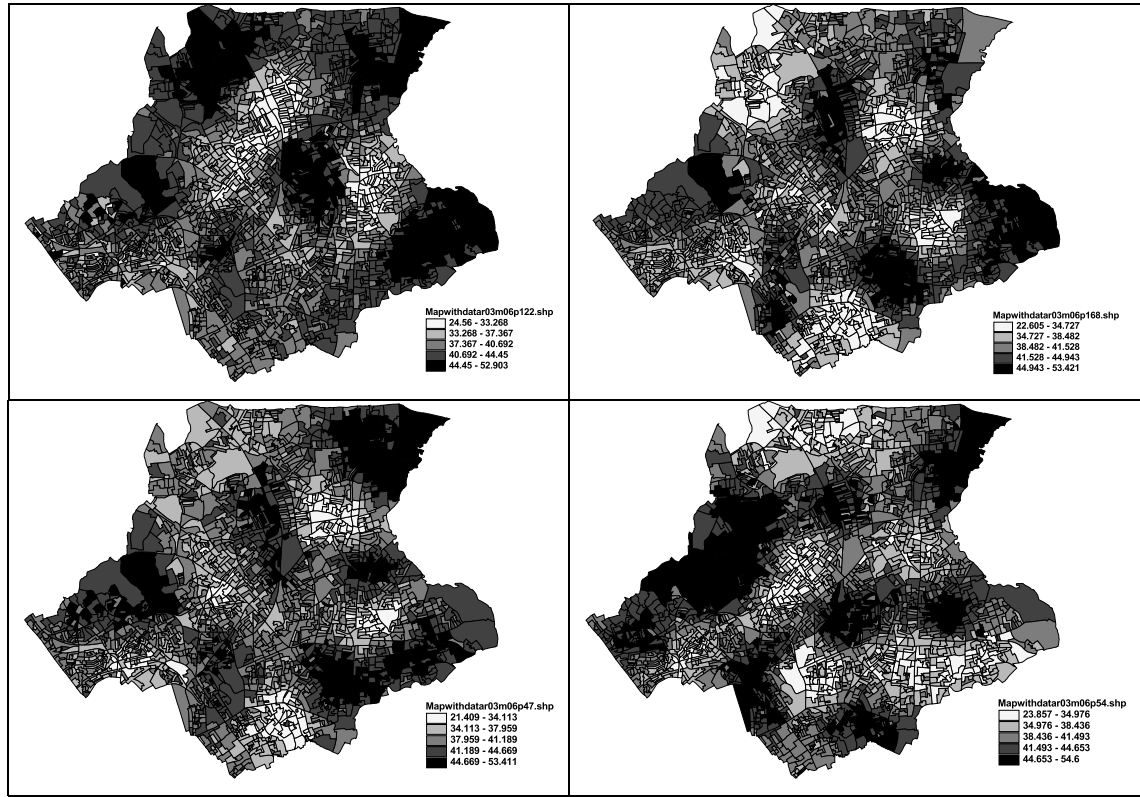


Figure 6.2: Some realizations of the data generator

satisfied the specified population parameter values. Most of the data sets generated are able to generate between 1090 to 2050 out of 3000 repetitions of the data set generator that satisfy the required properties. Some combinations of Moran's I and the required Pearson correlation generated data for which only a few repetitions satisfy the required properties. An example is when the required Moran's I is 0.02 and the Pearson correlation is 0.3, out of 3000 repetitions, 2046 satisfied the requirement. When the required Moran's I is 0.5 there are 1094 out of 3000 that satisfy the

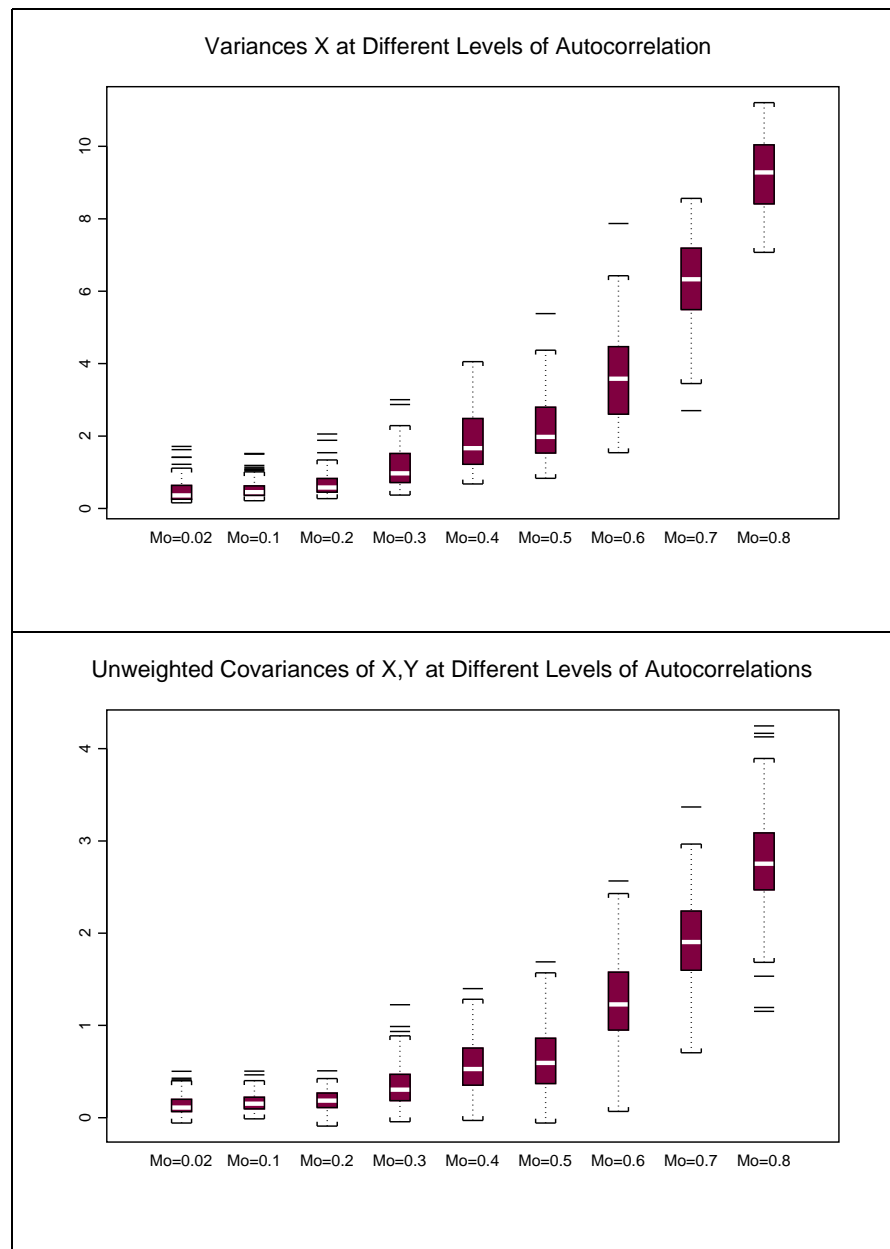


**Figure 6.3:** Some realizations of the data generator the same degree of autocorrelation  $Mo=0.8$

required values. The number of data points that satisfied the respective requirement decreases as the required Moran's I is increased. For each degree of autocorrelation described above, 1000 data points were selected at random from those that satisfy the required statistics and are analyzed in this subsection. All analyses are made with 1000 data points for each level of autocorrelation.

### 6.2.1 Behavior of Some Statistics

Figure 6.4 shows the distributions of the unweighted variances of variable X and covariances of variables X and Y at the Ward level for different degrees of autocorrelations. The variance of Y have characteristics similar to the variance of X. These statistics are not used in the computations of the statistics derived from the simple multilevel model. They are used to show that the variance is affected by the degree of autocorrelation of the variable. Recall that the initial variance of the variable



**Figure 6.4: Unweighted Variance of X and Covariance of X and Y at Ward level**

X is 16.0. When the data are aggregated into Wards, the mean or median of the variance reduce and the reduction is dependent on the degree of autocorrelation of the variable as depicted by the figure. It can be noted also that the standard deviation of the variance decreases as the autocorrelation of the variable decreases. The distribution of the covariance of variables X and Y display similar pattern.

Figure 6.5 shows the distribution of the weighted variance of variable X at Ward

level. Recall that the variance of variables X and Y is 16. The horizontal axis displays the degree of autocorrelation denoted by ( $Mo$ ) at ED level and the values of the Moran's I using the weight matrix corresponding to *queen's case lag 1*.

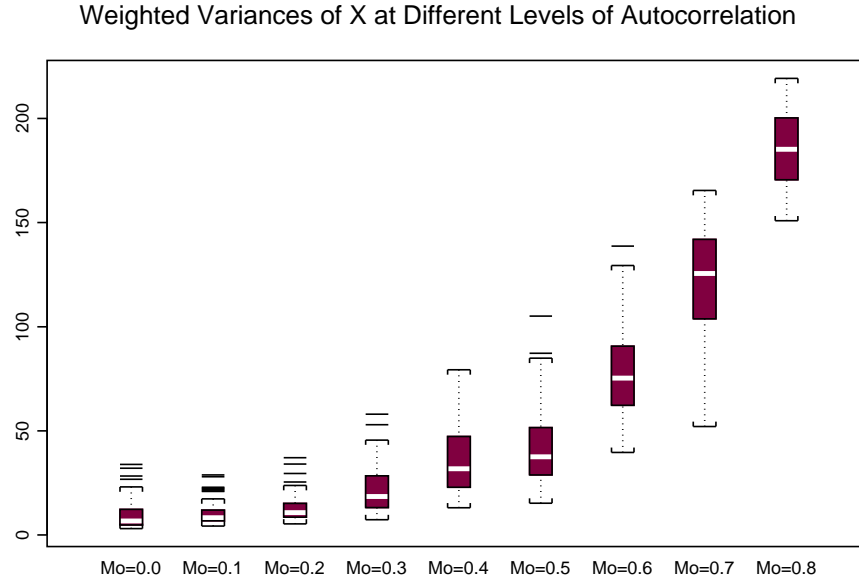


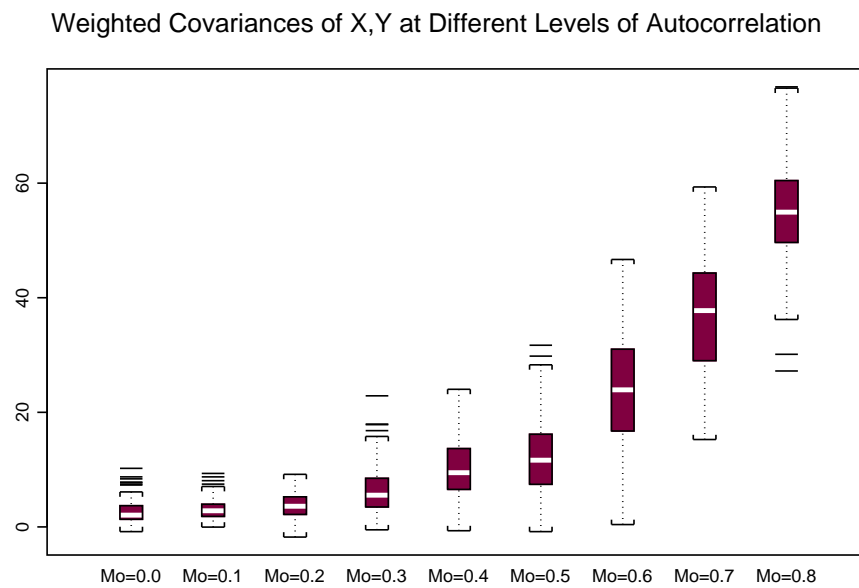
Figure 6.5: Weighted Variance X at Ward level

Moran I	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean	9.27	10.00	12.56	21.72	35.38	42.90	78.85	122.27	185.29
Median	6.67	8.26	10.75	18.45	31.77	37.53	75.36	125.60	185.22
Minimum	3.06	4.26	5.29	7.32	13.04	15.26	39.66	52.10	150.88
Maximum	33.90	28.79	37.11	58.02	79.34	105.10	138.72	165.44	219.25
Stan Dev	6.40	5.05	5.99	11.19	15.72	18.74	22.49	24.317	17.48

Table 6.1: Description of the distribution of Weighted Variances of X at Ward level

From here on, the Moran's I denoted by ( $Mo$ ) found in the horizontal axis of the figures will mean that the proximity or weight matrix used corresponds to *queen's case lag 1*. The figure shows a non-linear trend in the increase of the weighted variance as the level of autocorrelation increases. Beginning when ( $Mo$ )=0.02, the standard deviation increase with the increase of the degree of autocorrelation up to ( $Mo$ )=0.7 and decrease when ( $Mo$ )=0.8.

Figure 6.6 shows the distribution of the weighted covariance of variables X and Y. Similar to the weighted variance of X and weighted variance of Y (not shown), there is a non-linear increase of the covariance as the level of autocorrelations increase. The differences between the means and the medians are small. The standard deviations of the weighted covariances increase with the levels of autocorrelations except when Moran's I is 0.8.



**Figure 6.6: Weighted Covariance at Ward level**

Moran I	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean	2.76	3.10	3.68	6.57	10.90	12.74	23.67	36.94	54.94
Median	2.07	2.81	3.62	5.53	9.45	11.66	23.94	37.76	54.96
Minimum	-0.82	-0.03	-1.77	-0.51	-0.66	-0.82	0.42	15.24	27.21
Maximum	10.22	9.35	9.17	22.88	24.02	31.73	46.70	59.34	76.81
Stan Dev	2.09	1.84	2.01	4.52	5.60	6.70	9.75	9.96	8.85

**Table 6.2: Description of the Weighted Covariances at Ward level**

Recall that the initial correlations of the variables at the ED level is 0.3. Figure 6.7 shows the distribution of the correlations at Ward level at different degrees of autocorrelation. Not much differences are observed in the means and medians of

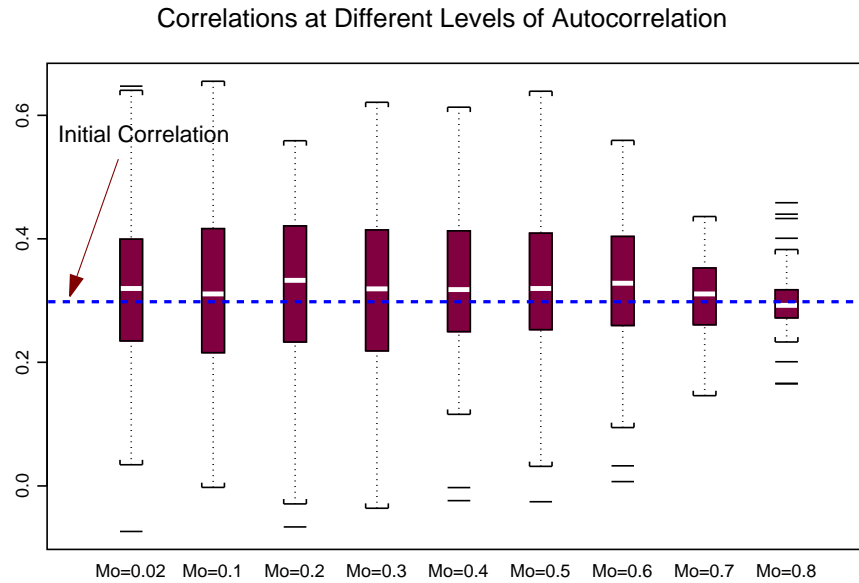


Figure 6.7: Correlations at Ward level

Moran I	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean	0.32	0.31	0.32	0.31	0.32	0.32	0.32	0.30	0.30
Median	0.32	0.31	0.33	0.32	0.32	0.32	0.33	0.31	0.29
Minimum	-0.07	0.00	-0.07	-0.04	-0.02	-0.03	0.01	0.15	0.17
Maximum	0.65	0.66	0.56	0.62	0.61	0.64	0.56	0.45	0.46
Stan Dev	0.13	0.14	0.14	0.15	0.13	0.12	0.11	0.062	0.05

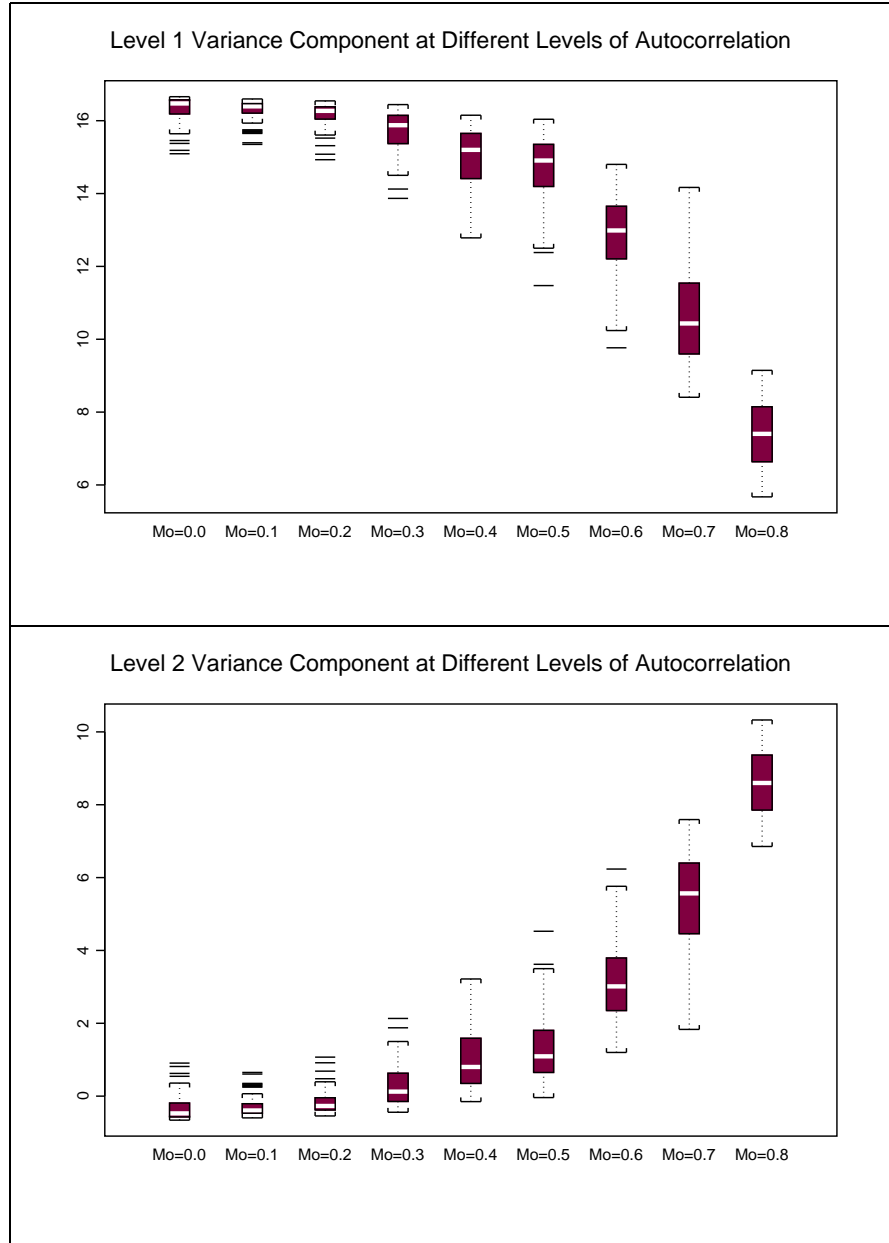
Table 6.3: Description of the direct correlations at different levels of autocorrelations

the Ward level correlations at different levels of autocorrelations which are all close to or equal to the initial correlation of 0.3. There is not much difference in the standard deviation when the Moran's I of the variables are from 0.02 to 0.5 but it starts to decrease starting when Moran's I equals 0.6. These standard deviations are higher than those given in Table 6.4, suggesting that variation in the number of units within each zone may increase the standard deviation of the direct correlation coefficient.

Figure 6.8 shows the distributions of the level 1 (ED level) and level 2 (Ward level) variance components of variable X. When the degree of autocorrelations are 0.02, 0.1, 0.2, and 0.3 then 85%, 85%, and 78% and 41% respectively have weighted



variances at Ward level that are less than the initial variance of 16.0. From Chapter 3, the estimate for the level 2 variance component denoted by  $\hat{\Lambda}_{XX}^{(2)}$  is  $\frac{S_{XX}^{(2)} - S_{XX}^{(1)}}{N^* - 1}$ , where  $S_{XX}^{(2)}$  is the weighted variance at Ward level and  $S_{XX}^{(1)}$  is the variance at ED level and is equal to 16.0. This explains why the same percentages in each degree



**Figure 6.8: Level 1 and Level 2 Variance components of X**

of autocorrelation will have negative level 2 variance components. A similar phenomenon is also observed in variable Y. When the degree of autocorrelations are 0.4

and 0.5, there are 7% and 2% level 1 variances greater than the initial variance of 16.0. When the degree of autocorrelation is 0.6, 0.7, and 0.8, the variances are all less than the initial variance. The negative estimated level 2 variance component phenomenon will be examined later in the chapter. The mean and median of the level 1 variance component decrease in a non-linear way as the degree of autocorrelation increases. It can be seen from the figure that the standard deviations increase with the degree of autocorrelation. In terms of the mean and median, the level 2 variance components behaves in the opposite way from that of the level 1 variance component. However, the standard deviations of the level 2 variance component behave in the same way as that of the level 1 variance component. These results are due to the result (3,31) in which the estimated variance components equal the original variance calculated using level 1 data.

Figure 6.9 shows the distribution of the level 1 and level 2 *pure* correlation coefficients. The mean level 1 *pure* correlation at different degrees of autocorrelation is not far from the Pearson correlation at ED level but the standard deviation becomes larger as the degree of autocorrelation increases. This means that this statistic has less chance of having a value far from the initial correlation when aggregated when the degree of autocorrelation is low and this chance increases as the degree of autocorrelation increases. For level 2 pure correlations, only three cases are presented since the rest do not make sense because, as stated before, there are level 2 variance components for either or both variables X and Y having negative results. There are even cases in which the estimated correlation at Ward level have values greater than 1.0 or less than -1.0 which is not a characteristic of a correlation coefficient. From these three cases level 2 pure correlation we see the reverse pattern when it comes to the standard deviation, namely it decreases as the degree of autocorrelation increases. The standard deviation of values of the statistics decrease as the degree of autocorrelation increase.

Figure 6.10 shows the distribution of level 1 and level 2 *pure* regression coefficient. A pattern similar to the *pure* correlation coefficients is observed.

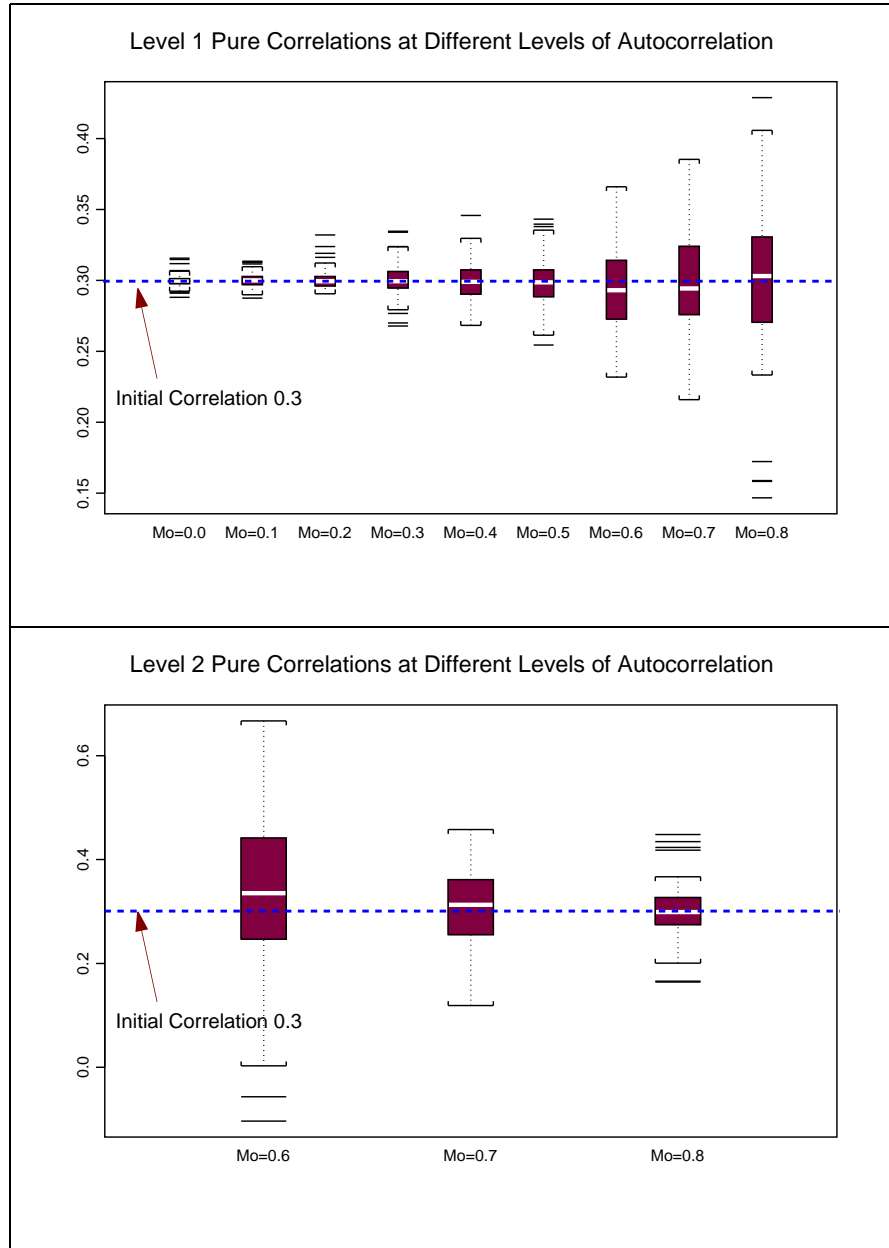


Figure 6.9: Level 1 and Level 2 Pure correlations

### 6.2.2 Aggregation Effects

Figure 6.11(a) shows the distributions of the variance aggregation effects ( $S_{XX}^{(2)}/S_{XX}^{(1)}$ ) for the weighted variances of X at different degrees of autocorrelations. Here  $S_{XX}^{(2)}$  is the weighted variance at Ward level and  $S_{XX}^{(1)}$  the variance at ED level. Figure 6.11(b) shows the covariance aggregation effects ( $S_{YX}^{(2)}/S_{YX}^{(1)}$ ) at different degrees of autocorrelations, where  $S_{YX}^{(2)}$  and  $S_{YX}^{(1)}$  are the covariances at Ward level and ED

Moran I	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Median	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Minimum	.29	0.29	0.29	0.27	0.27	0.25	0.23	0.22	0.15
Maximum	0.32	0.31	0.33	0.33	0.35	0.34	0.36	0.39	0.43
Stan Dev	0.004	0.005	0.01	0.01	0.01	0.02	0.03	0.03	0.05

**Table 6.4: Description of the Level 1 Pure Correlation at different levels of autocorrelations**

Moran I	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Mean	-0.24	-0.27	-0.19	0.15	0.50	0.43	0.34	0.32	0.30
Median	-0.30	-0.28	-0.20	0.08	0.46	0.36	0.34	0.31	0.30
Minimum	-0.63	-1.27	-2.64	-3.62	-0.69	-0.36	-0.10	0.12	0.16
Maximum	1.06	0.69	6.00	4.68	3.90	6.21	0.67	0.46	0.45
Stan Dev	0.23	0.25	0.88	0.97	0.60	0.66	0.16	0.07	0.05

**Table 6.5: Description of the Level 2 Pure Correlation at different levels of autocorrelations**

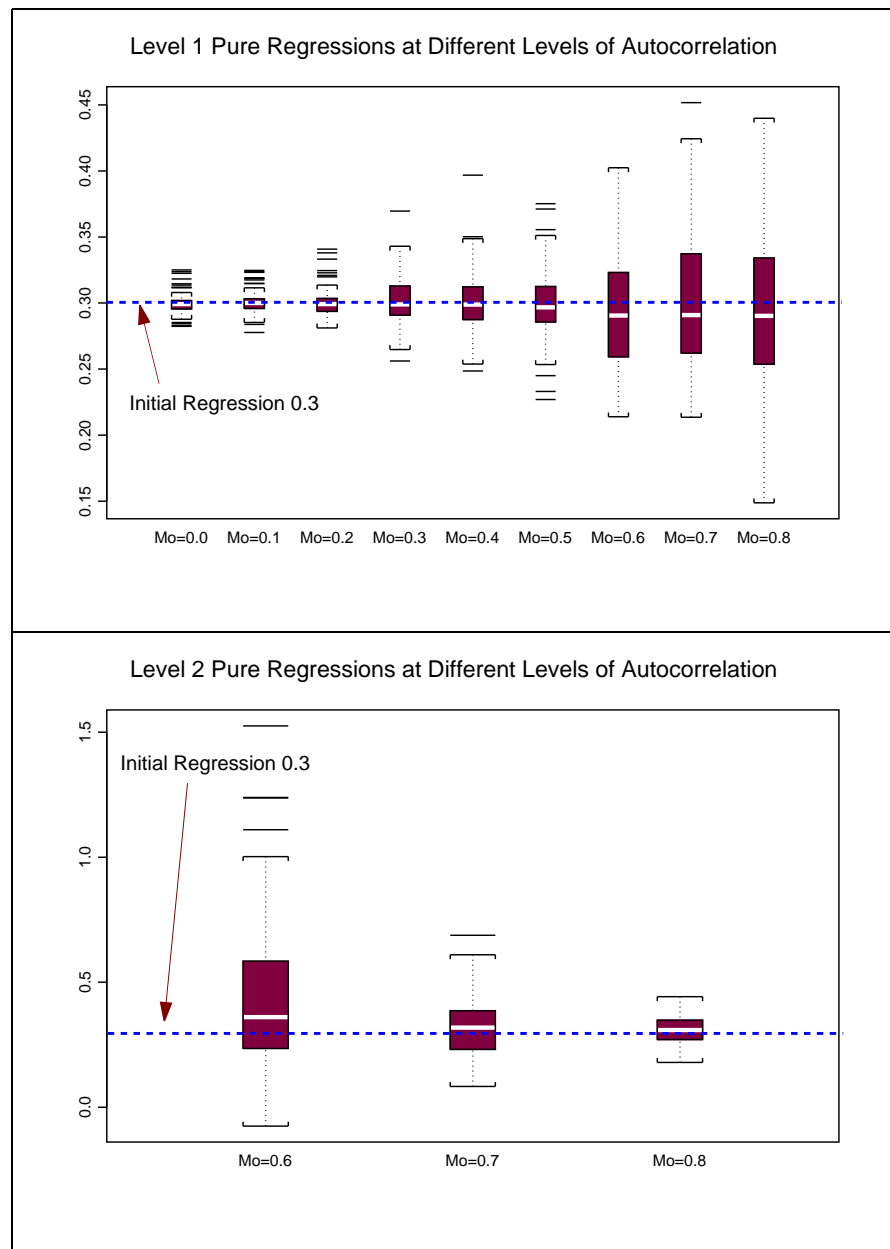
level respectively.

Moran I	0.02	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
<b>Var(X)</b>									
Mean	0.58	0.63	0.78	1.36	2.21	2.68	4.93	7.64	11.58
Median	0.42	0.52	0.67	1.15	1.98	2.35	4.71	7.85	11.58
Minimum	0.19	0.27	0.33	0.46	0.81	0.95	2.48	3.26	9.43
Maximum	2.12	1.80	2.32	3.63	4.96	6.57	8.67	10.34	13.70
Stan Dev	0.40	0.32	0.37	0.70	0.98	1.17	1.41	1.52	1.09
<b>Cov(Y,X)</b>									
Mean	0.57	0.65	0.77	1.37	2.27	2.66	4.93	7.70	11.45
Median	0.43	0.59	0.75	1.15	1.97	2.43	4.99	7.87	11.45
Minimum	-0.17	-0.01	-0.37	-0.11	-0.14	-0.17	0.09	3.18	5.67
Maximum	2.13	1.95	1.91	4.77	5.00	6.61	9.73	12.36	16.00
Stan Dev	0.44	0.38	0.42	0.94	1.17	1.40	2.03	2.08	1.84

**Table 6.6: Description of Variance of X and Covariance (Y,X) at different degrees of autocorrelations**

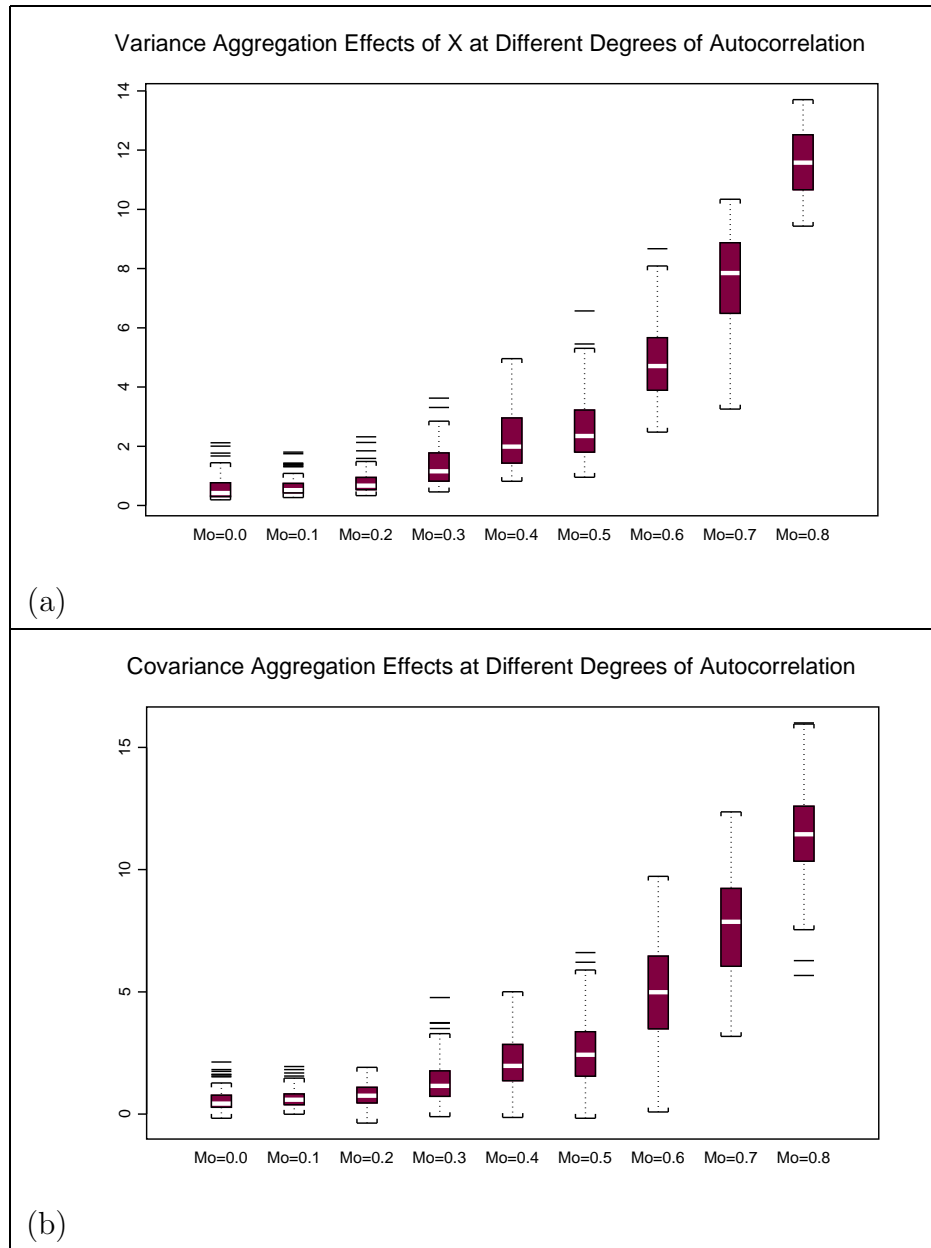
Both Figure 6.11 and Table 6.6 show a non-linear increasing trend in the variance and covariance aggregation effects as the degree of autocorrelation increases. As before the proximity matrix used in the computation of the Moran's I is the *queen lag 1*.

Figure 6.12 shows the aggregation effects of the variances of variable X at different



**Figure 6.10: Level 1 and Level 2 Pure Regression Coefficients**

levels of autocorrelations with two different proximity matrices used in computing the Moran's I. The group on the left shows the relationship between the variance effects of X with Moran's I calculated using a proximity matrix of queen lag 1. There are low negative linear correlations between aggregation effects and the corresponding Moran's I. The right group shows the relationship between the aggregation effect and the Moran's I with block proximity. This time there are very strong positive

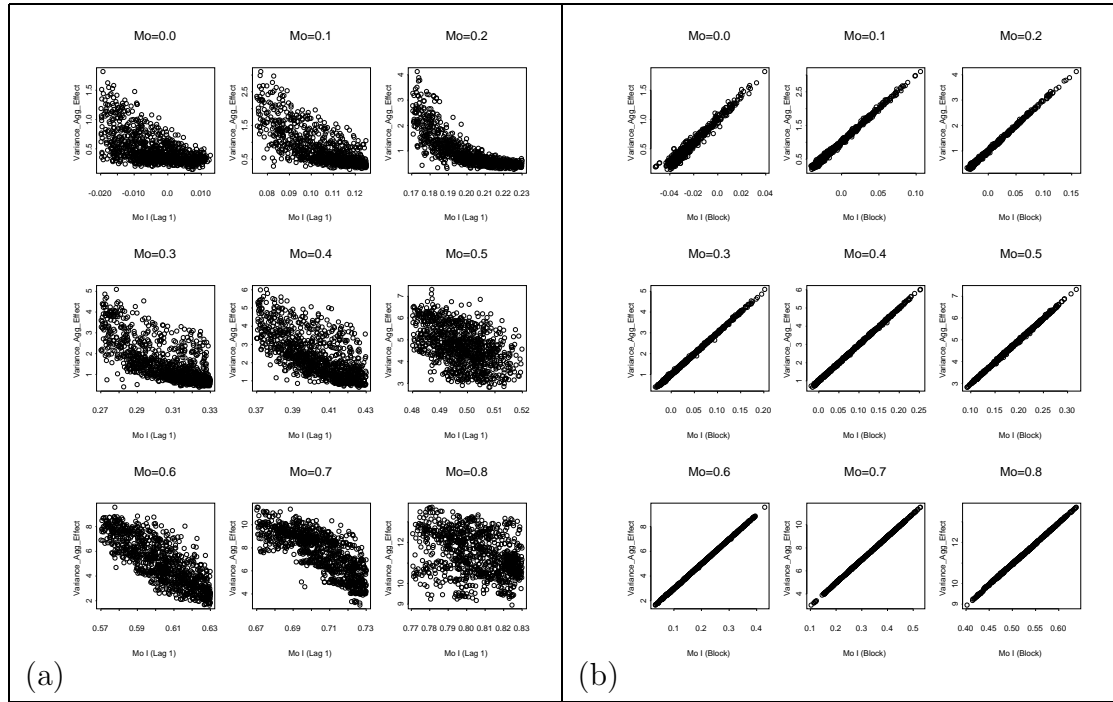


**Figure 6.11: Variance and Covariance Aggregation Effects at Different degrees of Autocorrelations**

linear associations between the aggregation effects and Moran's I. In fact, the first three on the top of the groups in the right have correlations of 0.99 and the rest have perfect positive linear correlations of 1.0.

Similar results were observed with the covariance aggregation effect.

Figure 6.13 shows the relationship between aggregation effects on the variances



**Figure 6.12: Variance Aggregation Effects versus Moran's I with Proximity matrix (a) Lag1 (b) "block"**

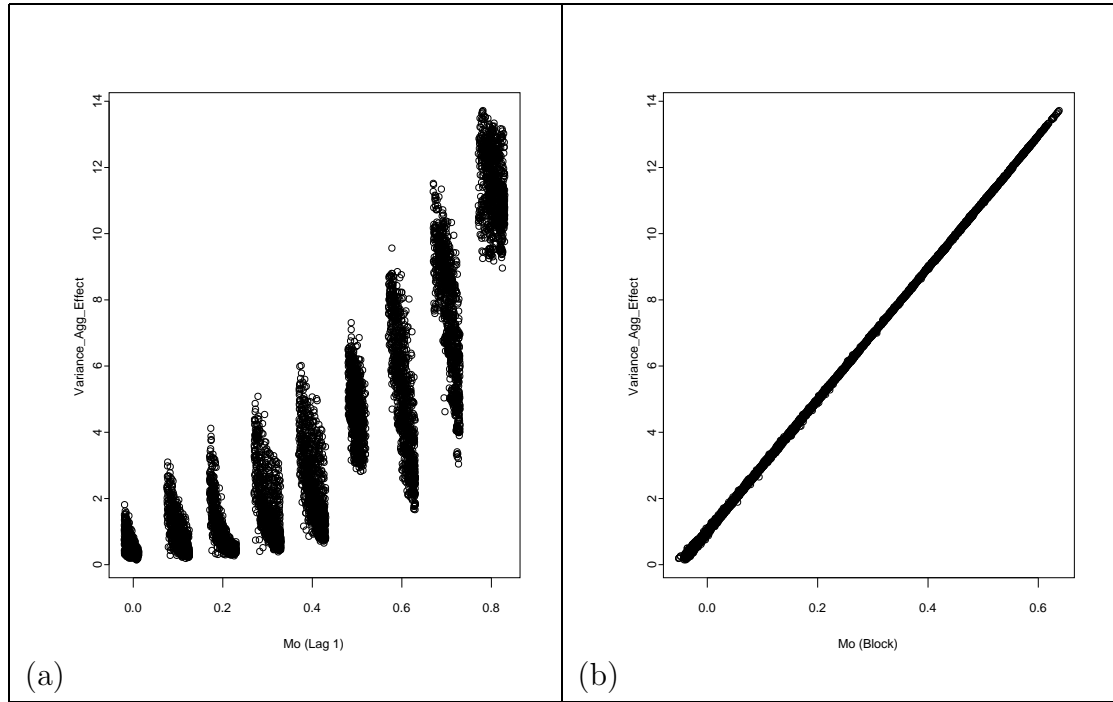
and the corresponding Moran's I on the full range of autocorrelations. Figure 6.13(a) shows the plots of the aggregation effects on the variances against the Moran's I with lag 1 proximity. Figure 6.13(b) shows the aggregation effects on the variances of X against Moran's I with block proximity and displays a perfect positive correlation.

These results show that it is the average autocorrelation within an areal unit that determines the aggregation effect on variances.

### 6.2.3 Intra-Area Correlations and Intra-Area Cross-Correlations

Figure 6.14 shows the relationship between intra-area correlations of variable X and the Moran's I using block proximity at different levels of autocorrelations. The correlations ranges from 0.992 to 1.00 which supports the relationship between the intra-area correlations and the Moran's I using block proximity matrix.

Figure 6.15 shows the relationship between the intra-area cross-correlations and the bivariate Moran's I. This shows the almost perfect positive correlations between



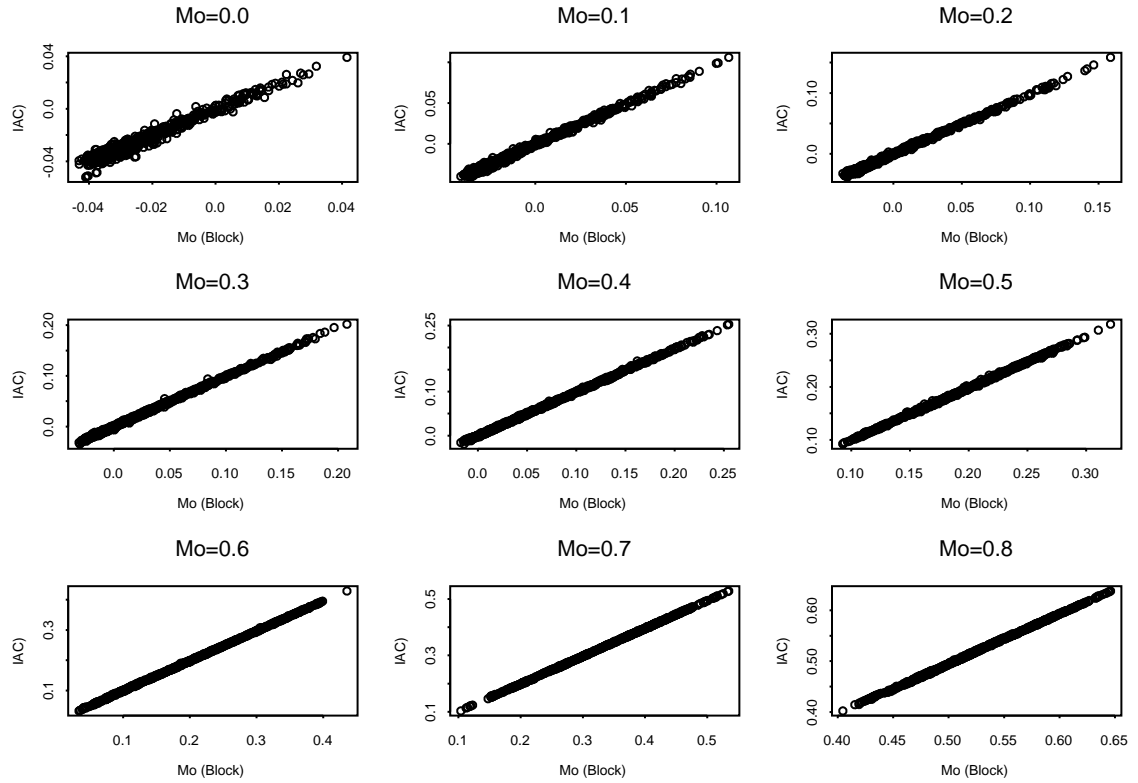
**Figure 6.13: Combined Variance Effects versus Moran's I with Proximity matrix (a) Lag1 (b) "block"**

the two statistics of the pairs of variables with low degrees of autocorrelations, and in the cases where the pairs of variables have high degree of autocorrelations, there is a perfect positive linear relationships.

### Summary

In this section data are generated with the same mean and variances, and specific Pearson correlation for different degrees of autocorrelation. The first data set is composed of two variables X and Y and to make the analysis simple, the variables have the same mean (40) and the same variance (16) at the ED level. The variables also have a fixed Pearson correlation equal to 0.3 but they are generated in such a way that the autocorrelation of the variables are equal but varies (0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) as measured by Moran's I that uses the weight matrix corresponding *queen's case lag 1*, ie  $w_{ij}=1$  if ED  $i$  and ED  $j$  are immediate neighbors and 0 otherwise. The following were observed:

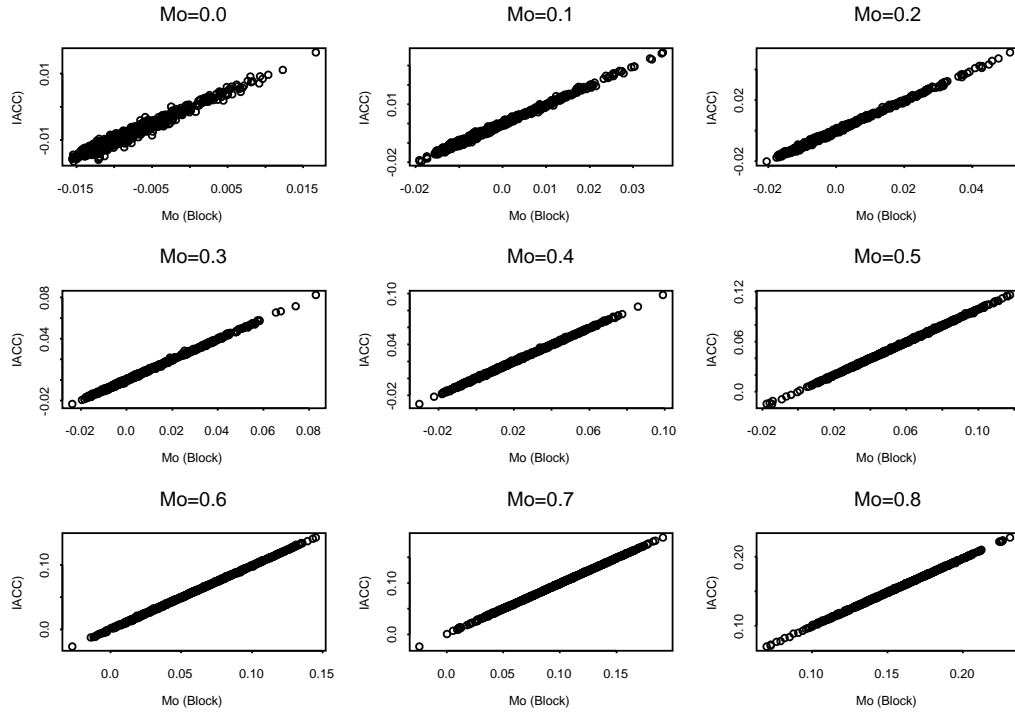




**Figure 6.14: Intra-area correlations versus Moran's I with block proximity**

The mean level 1 *pure* correlation at different degrees of autocorrelation is not far from the Pearson correlation at ED level but the standard deviation becomes larger as the degree of autocorrelation increase. This means that the level 1 *pure* correlation has less chance of having a value far from the initial correlation when aggregated when the degree of autocorrelation is low and this chance increases as the degree of autocorrelation increases. In the case of the level 2 *pure* correlation, when both variables have low to medium autocorrelations some cases are observed in which the estimated correlation at Ward level have values greater than 1.0 or less than -1.0, which is not a characteristic of a correlation coefficient. However, the standard deviation of the level 2 pure correlations decrease as the degree of autocorrelation increase.

There is a very strong positive relationships between the variance aggregation effects defined by  $(S_{XX}^{(2)}/S_{XX}^{(1)})$  and the intra-area correlation with "block" proximity.



**Figure 6.15: Intra-area cross-correlations versus Bivariate Moran's I with block proximity**

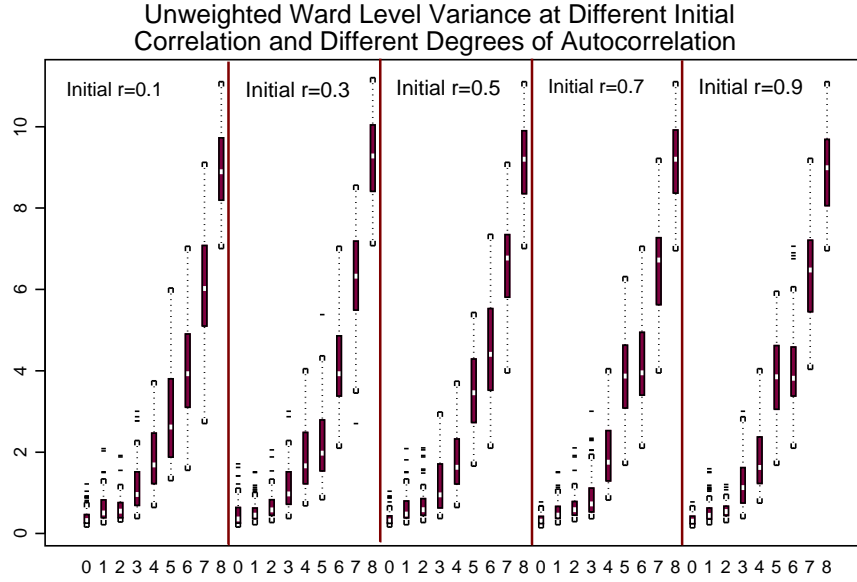
There is almost perfect positive correlations between the intra-area cross-correlations and the bivariate Moran's I with low degrees of autocorrelations, and in the cases where the pairs of variables have high degree of autocorrelations, there is a perfect positive linear relationships.

### 6.3 Different Initial correlations

Data sets are generated in such a way that the variables X and Y have the same mean (40) and the same variance (16) at the ED level, the same degrees of autocorrelations but the initial correlations differ. The values on the initial correlations are 0.1, 0.3, 0.5, 0.7, and 0.9.

For the next figures, the following numbers in the horizontal axis of the the figure means: 0 implies Mo=0.0, 1 implies Mo=0.1, 2 implies Mo=0.2, 3 implies Mo=0.3, 4 implies Mo=0.4, 5 implies Mo=0.5, 6 implies Mo=0.6, 7 implies Mo=0.7, 8 implies

$\text{Mo}=0.8$ .



**Figure 6.16: Unweighted Variance of X at Ward level with different initial correlations at different degrees of autocorrelations**

Figure 6.16 shows the distributions of the unweighted variances of variable X at Ward level with different initial correlation and different degrees of autocorrelation. The variance is not affected by the initial correlation. In all cases the unweighted variance at ward level, regardless of the initial correlation at ED level, depends on the degree of autocorrelation. As the degree of autocorrelation increases, the change of the variance decreases but the standard deviation increase with the degree of autocorrelation.

Figure 6.17 shows the distributions of the unweighted covariances of variables X and Y at Ward level with different initial correlation and different degrees of autocorrelation. Unlike the variance, the Ward level covariance is affected by the initial correlation. Figure 6.18 shows the distributions of the covariance at Ward level at different initial correlations at different degrees of autocorrelations. The distributions have the same trend as the unweighted covariance at Ward level.

Figure 6.19 shows the behavior of the aggregated correlations (Ward level corre-

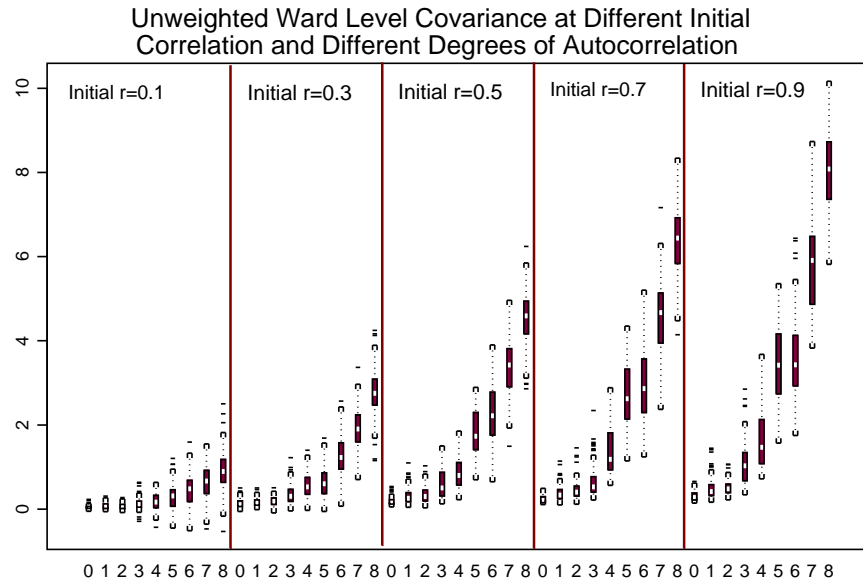


Figure 6.17: Unweighted Covariance of X and Y at Ward level with different initial correlations at different degrees of autocorrelations

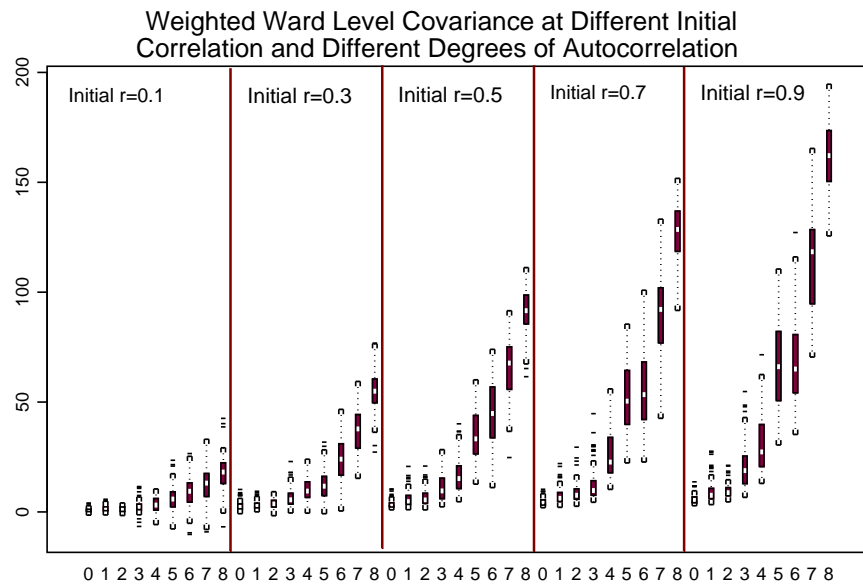


Figure 6.18: Aggregated weighted covariance with different initial correlations at different degrees of autocorrelations

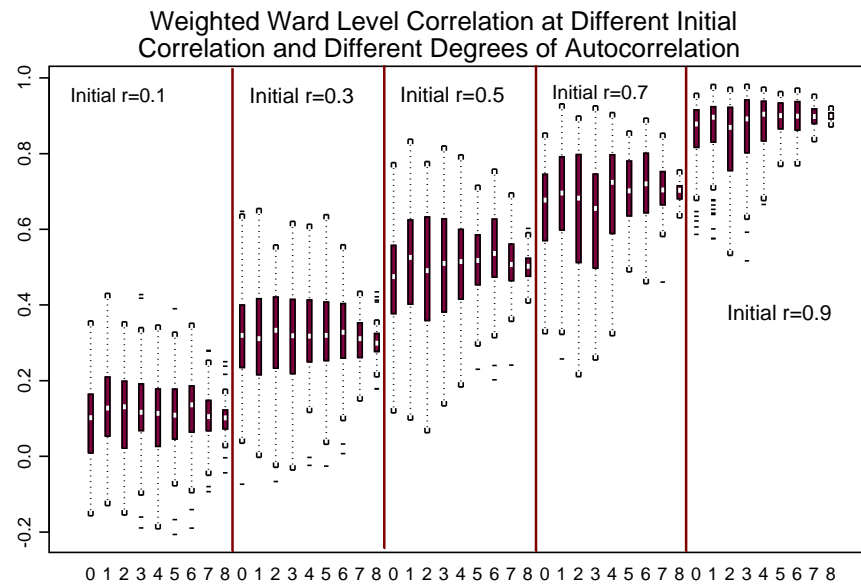


Figure 6.19: Weighted Correlations at Ward level with at different initial correlations and different degrees of autocorrelations

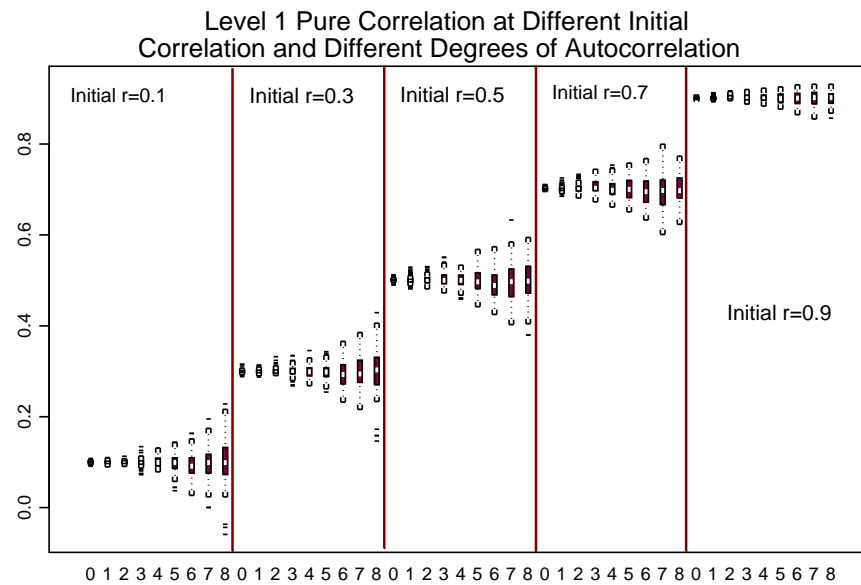
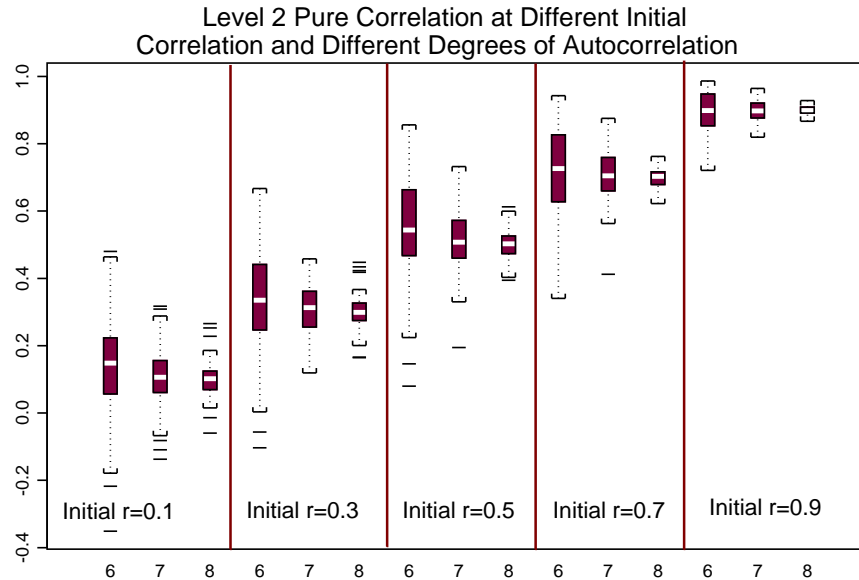


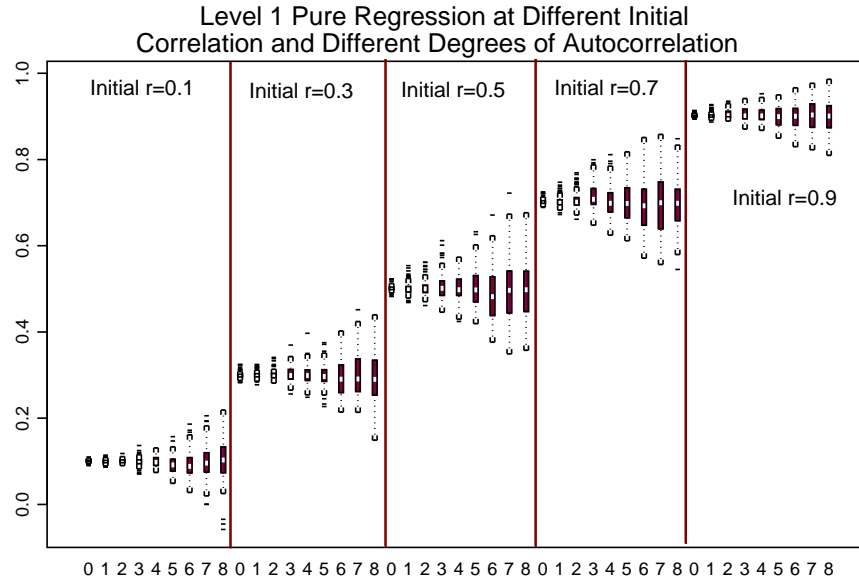
Figure 6.20: Level 1 Pure Correlations at different initial correlations at different degrees of autocorrelations

lations) with different initial correlations at the ED level. In terms of the standard deviations of the different groups (different initial correlations), the patterns are similar, the standard deviations tend to decrease when the autocorrelations get larger. When the data have high initial correlation (0.9), the magnitude of the corresponding standard deviations is a smaller than the rest of the group. The mean and median for each group fluctuate around their corresponding initial correlation.

Figure 6.20 shows the distribution of level 1 pure correlations with different initial correlations at different levels of autocorrelations. The figure is grouped according to the initial correlations given by the values of  $r$  in the figure. The boxplots in each group represents the distributions at different levels of autocorrelation of 0.02, 0.1,...,0.8 respectively. Looking at the figure, each group of the groups have similar characteristics, one of which is that the variation of the level 1 pure correlations increases as the degree of autocorrelation increases. It can also be observed that as the initial correlation gets bigger, the corresponding variation in each degree of autocorrelation of the level 1 pure correlations decrease.



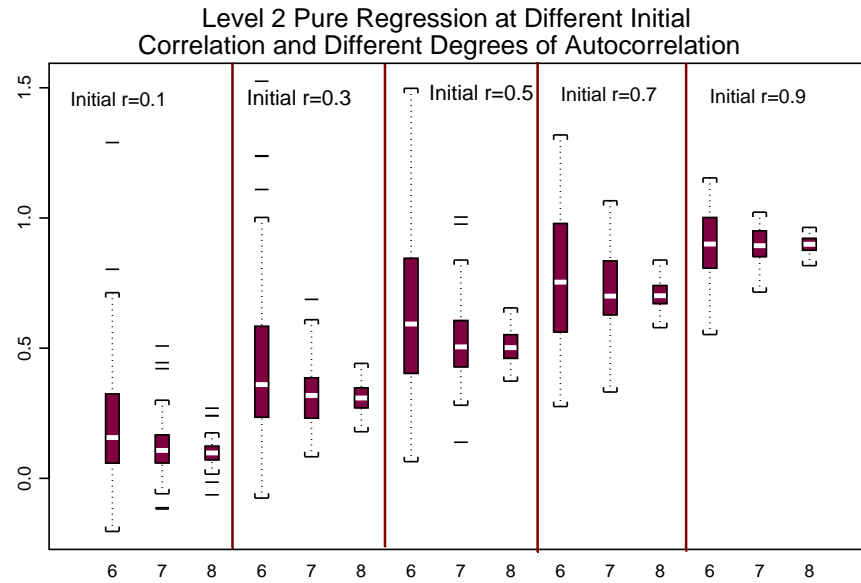
**Figure 6.21:** Level 2 Pure Correlations at different initial correlations at different degrees of autocorrelations



**Figure 6.22: Level 1 Pure Regression at different initial correlations at different degrees of autocorrelations**

Figure 6.21 shows the distributions of level 2 pure correlations for data with initial correlations equal to the values of  $r$  statistics in the figure. There are only three boxplots in each group, corresponding to autocorrelation measures of 0.6, 0.7, and 0.8, respectively. They are the ones shown because the variables with autocorrelation measure of below 0.6, have several undefined level 2 pure correlations and values more than 1 and less than 1 which is not a characteristic of a correlation coefficient. It can be observed that regardless of the initial correlations of the variables, the variation decreases as the degree of autocorrelation increases.

Figure 6.22 shows the distribution of level 1 pure regressions with different initial correlations at different levels of autocorrelations. Looking at the figure, each group has similar characteristics, one of which is that the variation of the level 1 pure correlations increases as the degree of autocorrelation increases. It can also be observed that as the initial regression gets bigger, the corresponding variations in each degree of autocorrelation of the level 1 pure correlations decreases. The behavior is very similar to the behavior of level 1 pure correlation

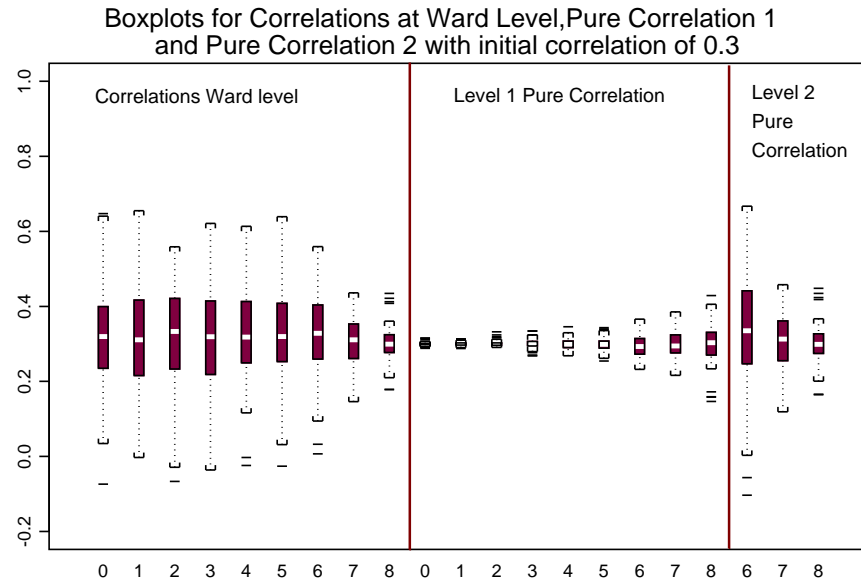


**Figure 6.23: Level 2 Pure Regression at different initial correlations at different degrees of autocorrelations**

Figure 6.23 shows the distributions of level 1 pure regression for data with initial correlations equal to the values of  $r$  statistics in the figure. Looking at the figure, each group has similar characteristics, one of which is that the variation of the level 1 pure correlations increase as the degree of autocorrelation increase. It can also be observed that as the initial correlation gets bigger, the corresponding variations in each degree of autocorrelation of the level 1 pure regression decreases.

To compare the behavior of correlation at Ward level, level 1 pure correlation, and level 2 pure correlation, the three statistics were combined in one graph. Figure 6.24 shows the distributions of the three statistics for the data sets where the initial correlation at ED level is 0.3 at different degrees of autocorrelation. The first group of boxplots is the distribution of the correlations at Ward level with measures of autocorrelations 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9, respectively. Similarly, the second group are the boxplots for level 1 pure correlation at corresponding degrees of autocorrelations. The third group consists of the plots of level 2 pure correlation. Only the variables with measures of autocorrelation equal to 0.6, 0.7,





**Figure 6.24:** Distributions of the three statistics at initial correlations of 0.3 at different degrees of autocorrelations

and 0.8 are included.

Figure 6.25 shows the distributions of aggregation effects of the weighted covariances at different initial correlations and different degrees of autocorrelation. The mean or the median have similar nonlinear increasing trend even if the data sets have different initial correlations. The corresponding standard deviations as the degree of autocorrelation increase seems to increase with aggregation except when the degree of autocorrelation is 0.9 at different initial correlations.

To look deeper the behavior of the generated data, the graph for the aggregation effect was revised so that the boxplots are categorized in terms of the Moran's I. Figure 6.26 shows the behavior of the generated data. The standard deviations decrease as the initial correlations becomes bigger but the corresponding magnitude of the standard deviations become bigger as the degree of autocorrelation increase.

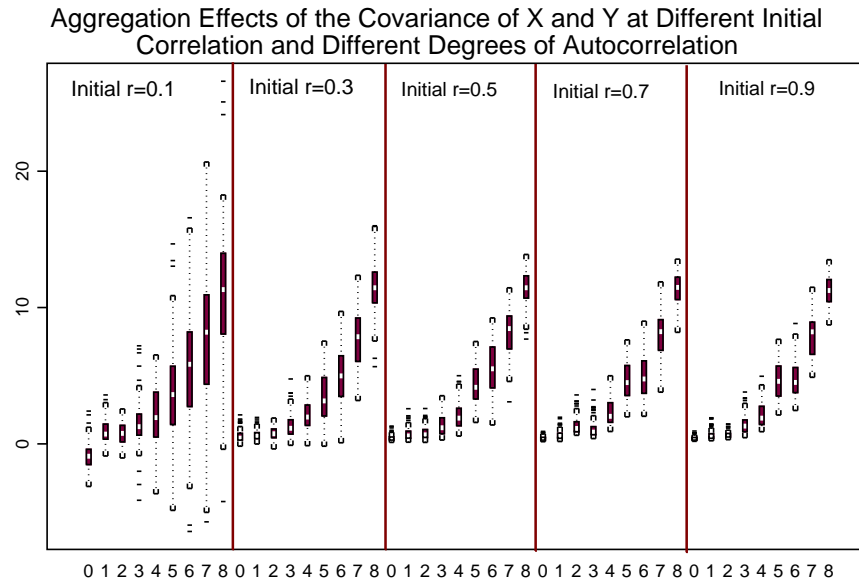


Figure 6.25: Aggregation Effects at different initial correlations at different degrees of autocorrelations

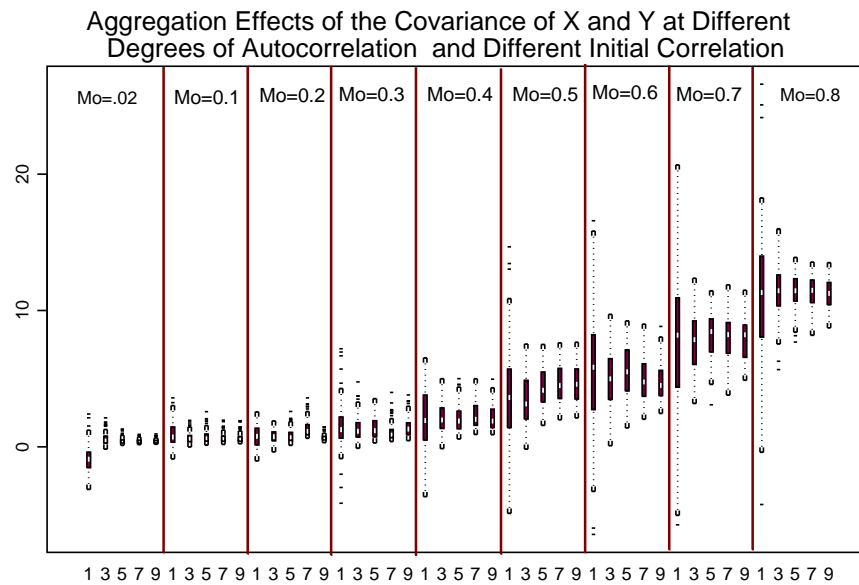


Figure 6.26: Aggregation Effects at different degrees of autocorrelations at different initial correlations

### 6.3.1 Relationship between some statistics and the Moran's I with different proximity matrices:

Figure 6.27 shows the unweighted variance plotted against Moran's I. The upper portion of the figure shows the relationship between the unweighted variance and Moran's I with *Lag1* proximity at different initial correlations at ED level while the lower portion the relationship between the unweighted variance and the Moran's I with *Block* proximity. Recall that '*block*' proximity corresponds to proximity weights,  $w_{ij}=1$  if ED  $i$  and ED  $j$  are in the same Ward, and 0 otherwise.

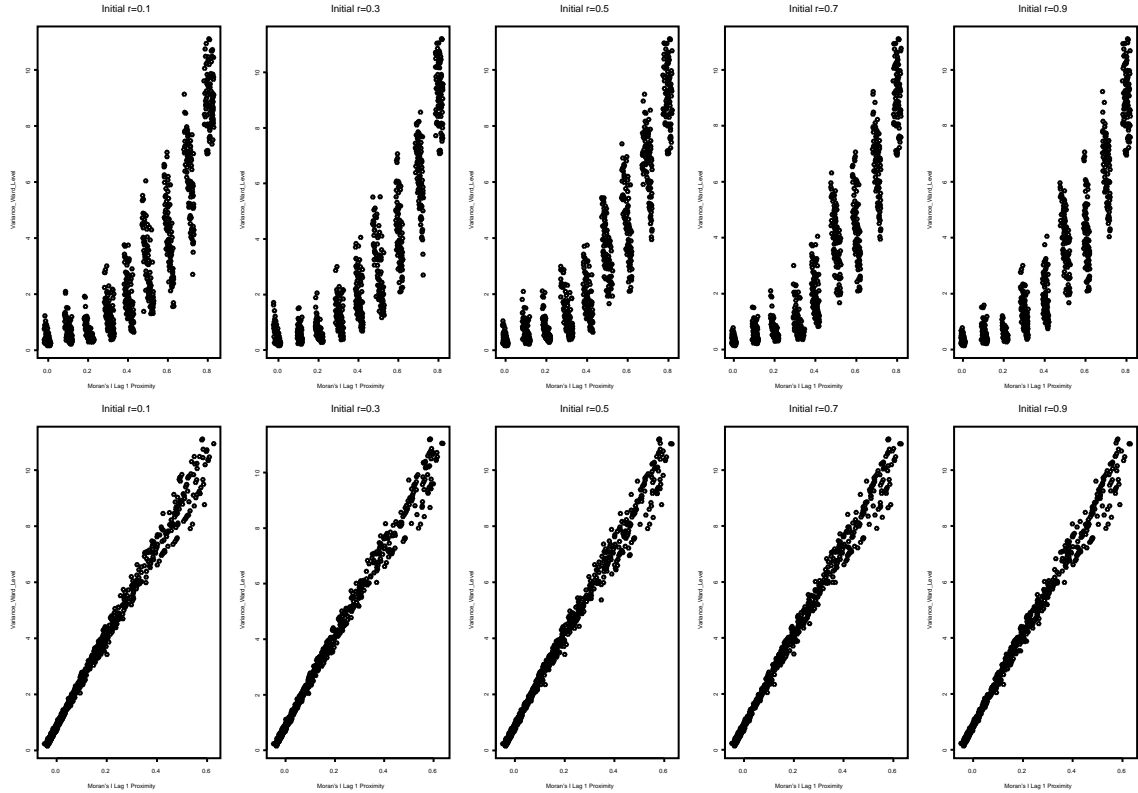
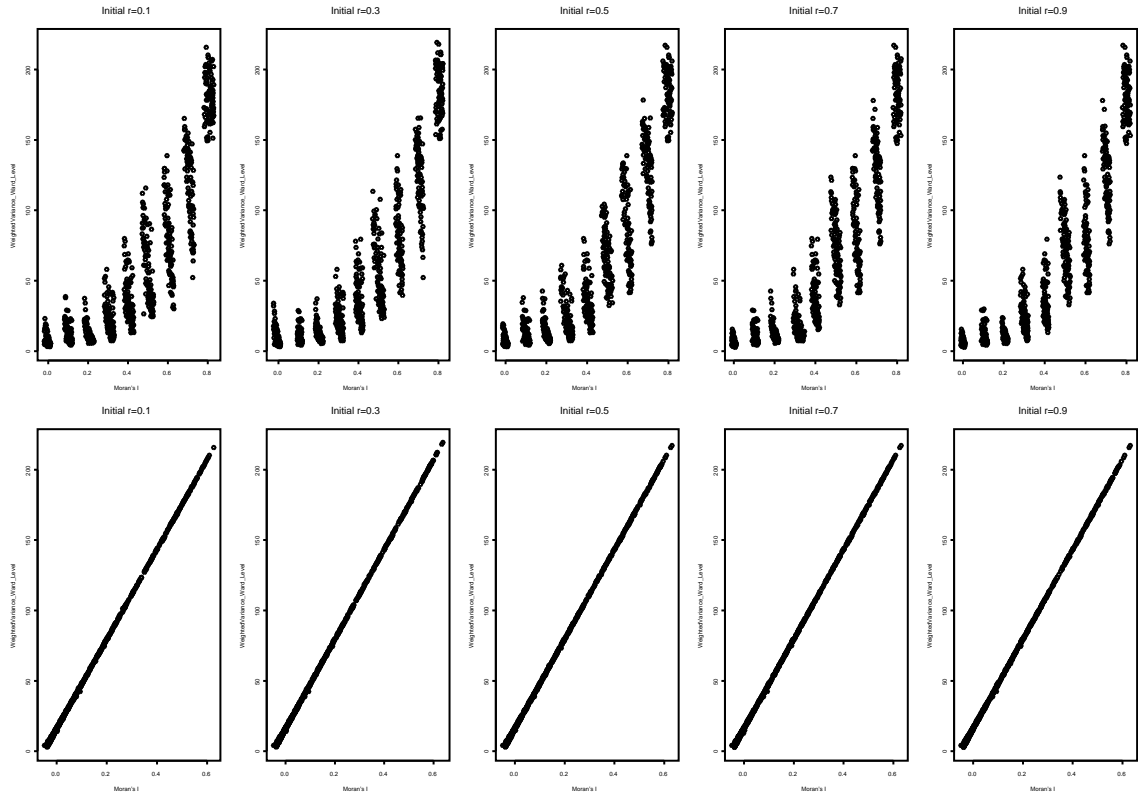


Figure 6.27: Relationship between Unweighted Variance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right  $r=0.1$ ,  $r=0.3$ ,  $r=0.5$ ,  $r=0.7$ ,  $r=0.9$ ; The vertical axes are the variances and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity

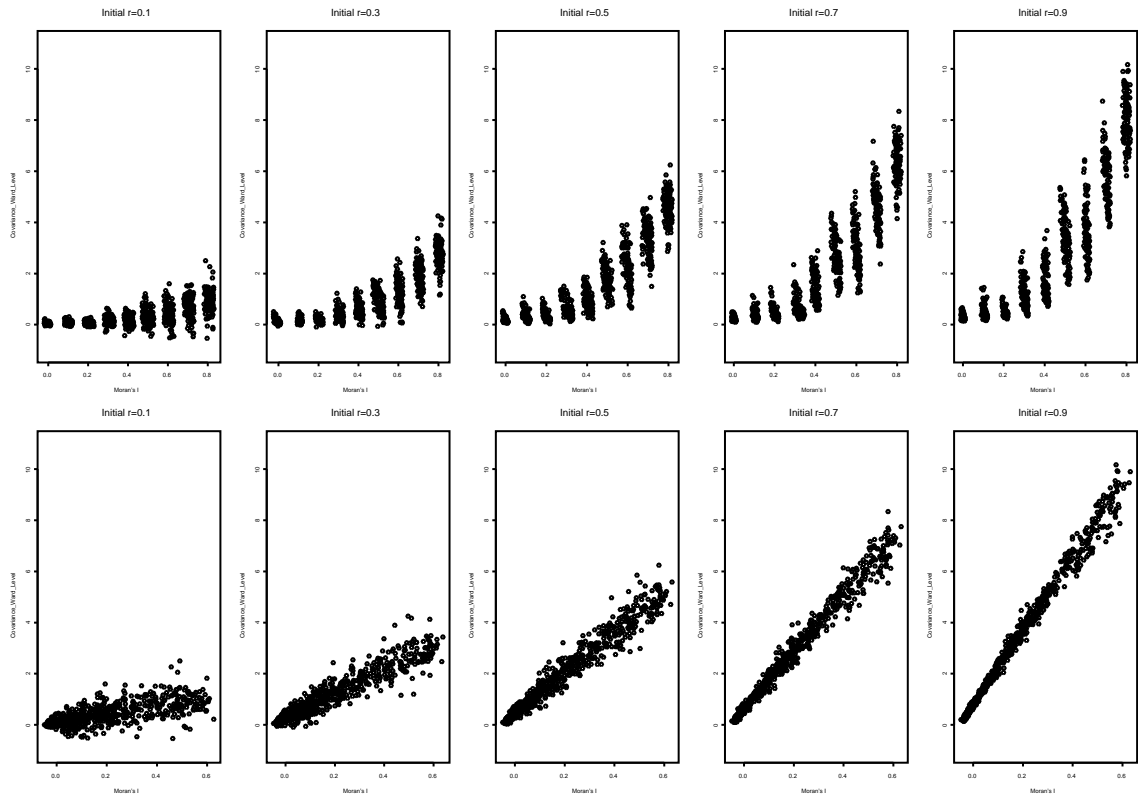
Figure 6.28 shows the relationship between weighted variances plotted against the



**Figure 6.28: Relationship between Weighted Variance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of auto-correlations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right  $r=0.1$ ,  $r=0.3$ ,  $r=0.5$ ,  $r=0.7$ ,  $r=0.9$ ; The vertical axes are the Weighted Variance and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity**

Moran's I at different proximity matrices. In both cases of the proximity matrices, there is a relationship. The big difference is that when the proximity matrix used is Block proximity, there is a perfect correlation between the weighted variance and the Moran's I regardless of the initial correlation at ED level.

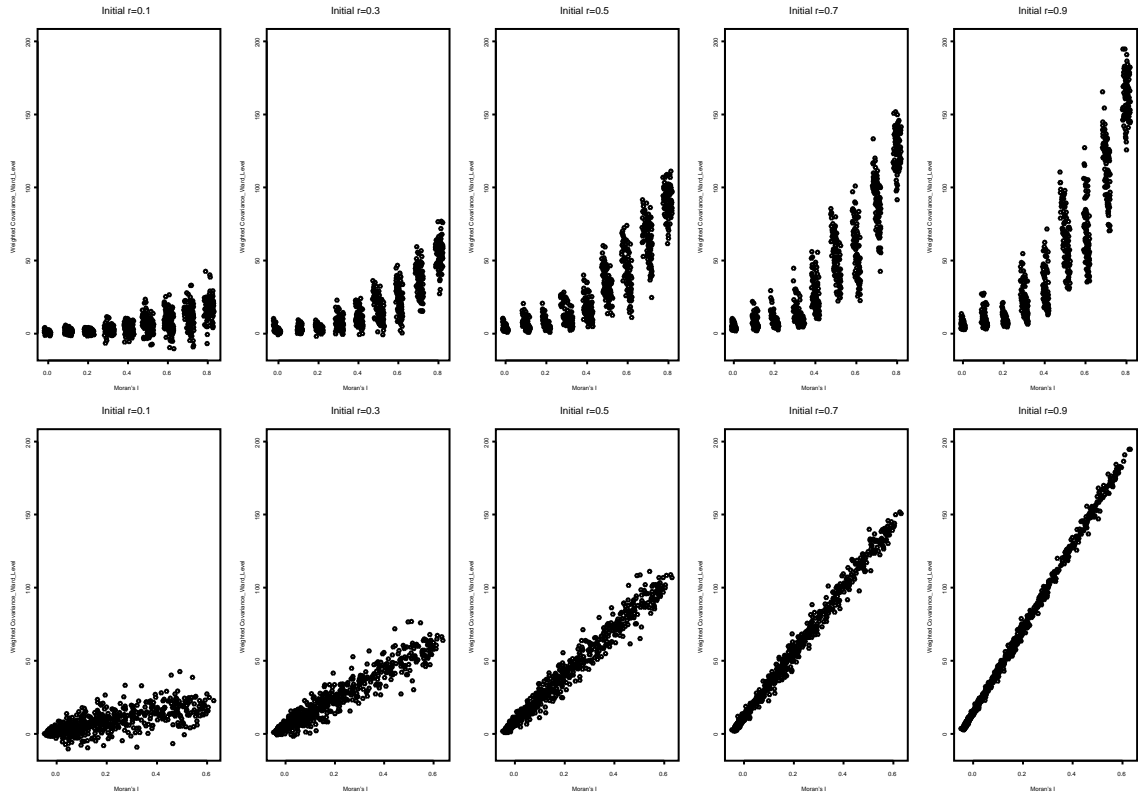
Figure 6.29 shows the relationship between the unweighted covariance and the Moran's I. When the proximity matrix is Lag 1 there is a non-linear relationship between the two statistics. When the proximity matrix is Block proximity a strong linear relationship is displayed.



**Figure 6.29: Relationship between Unweighted Covariance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right  $r=0.1$ ,  $r=0.3$ ,  $r=0.5$ ,  $r=0.7$ ,  $r=0.9$ ; The vertical axes are the Unweighted Covariance and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity**

Figure 6.30 shows the relationship between the weighted covariance and the Morans'I with different proximity matrices.

Figure 6.31 shows the relationship between the Ward level correlation and the Morans'I with different proximity matrices. We see that the Ward level correlation is affected by the ED level correlation. For a given ED level correlation the relationship with Moran's I with Lag 1 proximity is evident in the SD which decreases as the Moran's I increases. This effect is more pronounced in the Block proximity case. There does not appear to be a relationship between the mean of the Ward level



**Figure 6.30: Relationship between Weighted Covariance and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of auto-correlations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right  $r=0.1$ ,  $r=0.3$ ,  $r=0.5$ ,  $r=0.7$ ,  $r=0.9$ ; The vertical axes are the Weighted Covariance and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity**

correlation and the Moran's I for either of the choices of proximity matrices.

### Summary

A set of data is generated in such a way that variables X and Y have the same mean (40) and the same variance (16) at the ED level. The autocorrelation of the variables are equal but vary (0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) as measured by Moran's I using Lag 1. This time the initial Pearson correlations are varied (0.1, 0.3, 0.5, 0.7, and 0.9).

The weighted variance is not affected by the initial correlation, regardless of the

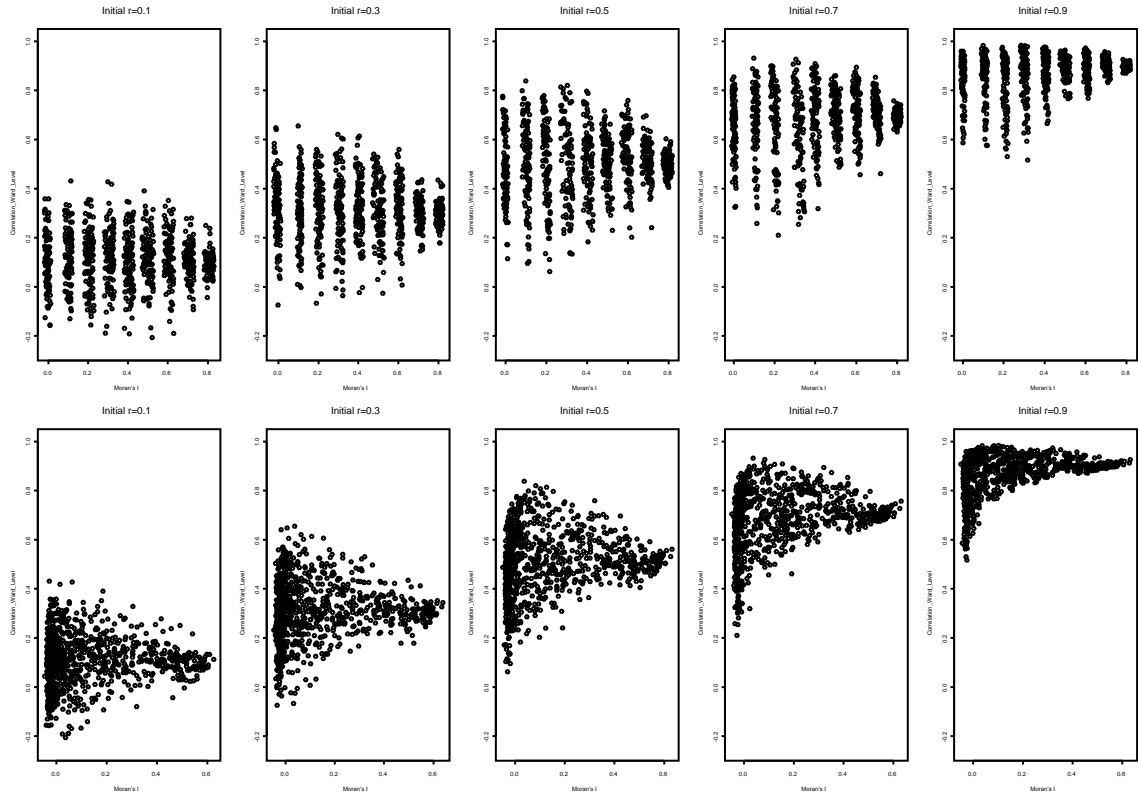


Figure 6.31: Relationship between Correlations and the Moran's I with Lag 1 and Block Proximity at different initial correlations at different degrees of autocorrelations. Note: The labels at the top of the top of the boxes are the initial correlations: from left to right  $r=0.1$ ,  $r=0.3$ ,  $r=0.5$ ,  $r=0.7$ ,  $r=0.9$ ; The vertical axes are the Correlations and the horizontal axes are the Moran's I with values ranges from 0.0 to 0.8 and the upper boxes with Lag1 Proximity and the lower boxes with Block Proximity

initial correlation at ED level, but depends on the degree of autocorrelation. As the degree of autocorrelation increases, the change of the variance decreases but the standard deviation increase with the degree of autocorrelation.

The mean and median of the *direct* correlation for each group fluctuate around their corresponding initial correlation. In terms of the standard deviations of the different categories corresponding to different initial correlations, the patterns are similar, that is the standard deviations tends to decrease when the autocorrelations get larger. When the data have high initial correlation (0.9), the magnitude of the corresponding standard deviations is smaller than the rest of the cases.

The mean of the level 1 *pure* correlation is not affected by the initial correlation of the two variables and the initial degree of autocorrelation as long as the autocorrelations of the two variables are almost equal. The variation of the level 1 *pure* correlations increase as the degree of autocorrelation increases. It can also be observed that as the initial correlation gets bigger, the corresponding variation in each degree of autocorrelation of the level 1 pure correlations decreases.

In terms of the aggregation effect, the standard deviations decrease as the initial correlations becomes bigger but the corresponding magnitude of the standard deviations become bigger as the degree of autocorrelation increases.

## 6.4 Case 2: Variables have different spatial autocorrelation

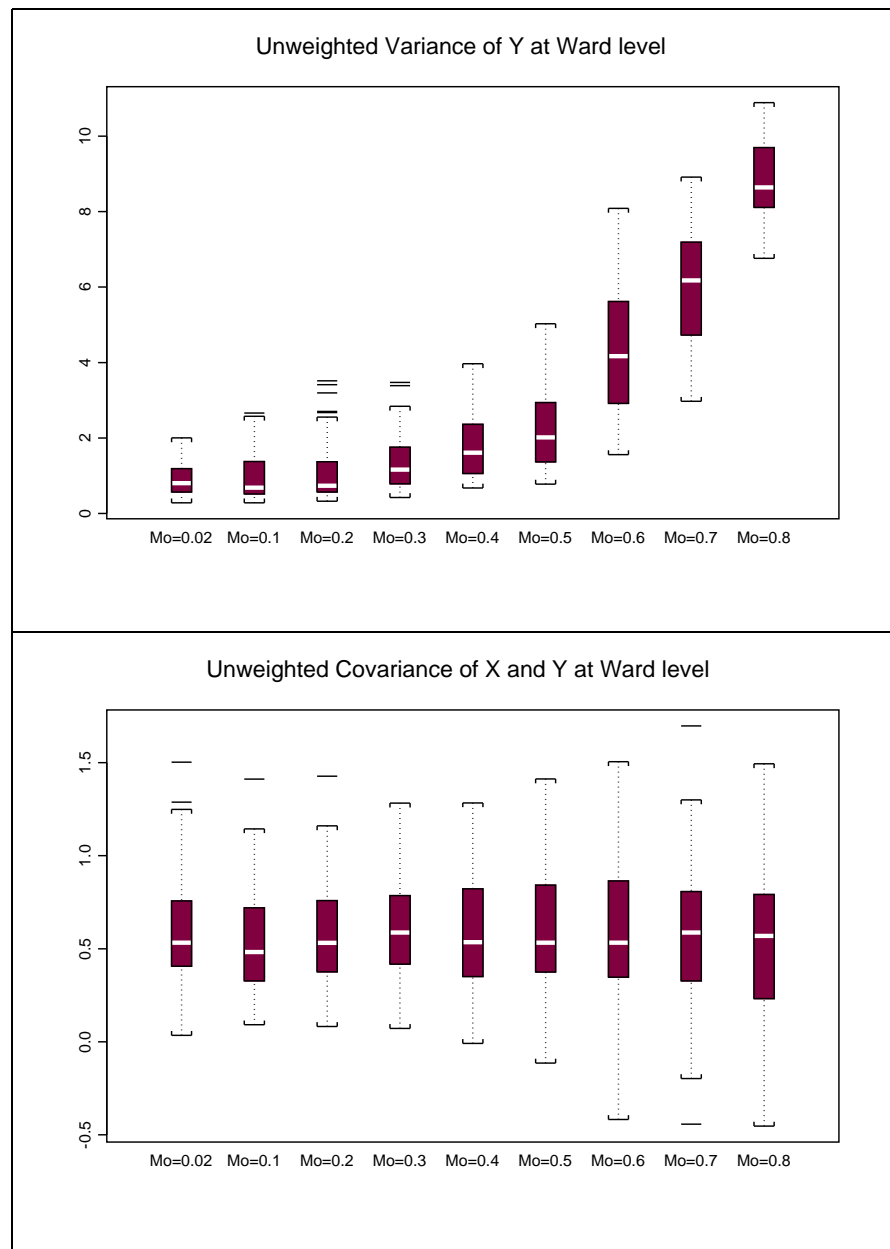
This experiment examines the behavior of statistics when the degree of autocorrelation of one variable is different from the other. The data are generated in such a way that variable X has autocorrelation of 0.4 as measured by Moran's I with *Lag 1* proximity matrix and variable Y with Moran's I at 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. The initial correlation in all cases is 0.3 and both variables have mean 40 and variance 16 in all cases.

Figure 6.32 shows the distribution of the unweighted variance of variable Y and the covariance of variables X and Y at Ward level with Y having different degrees of autocorrelation as described above. The figure support the claim that the change in variance is moderated by the level of autocorrelation (Gotway and Young, 2002). The mean of the covariance of variables X and Y seems to be constant and the standard deviation seems to increase as the degree of autocorrelation increases. These will have some effects on the other statistics. The change of the variance of variable X is not shown but the mean variance of X is 3.8653.

Similar behaviors were observed for the corresponding weighted variance and covariance at Ward level.

Figure 6.33 shows the distribution of the correlation at Ward level with differ-

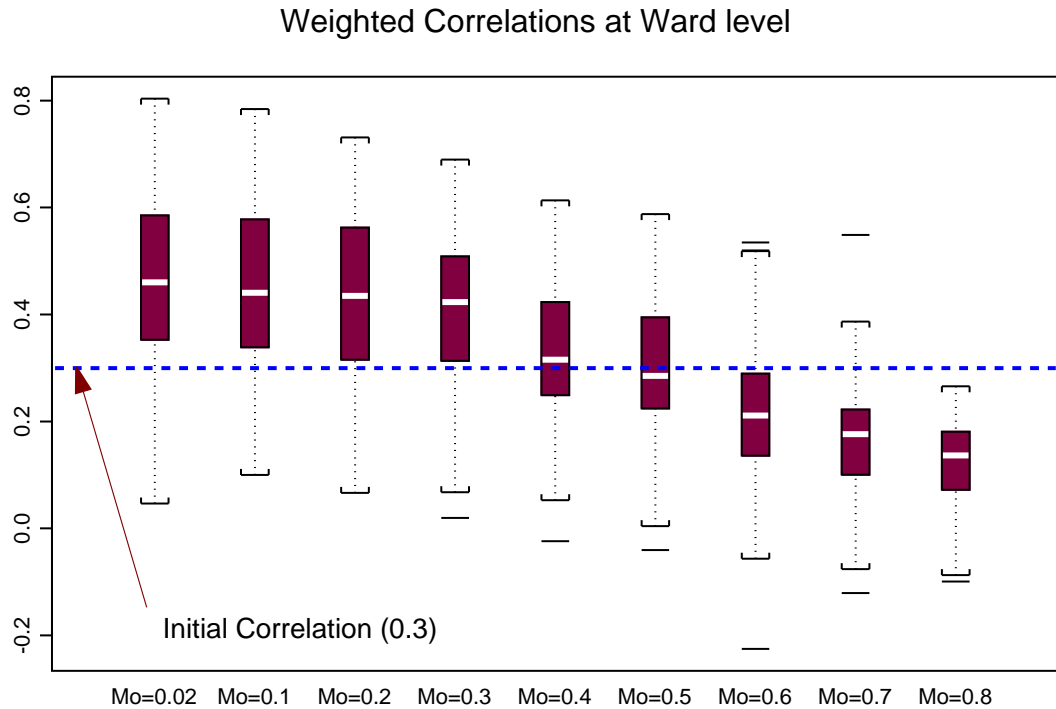




**Figure 6.32: Unweighted Variance of Y and Covariance of X and Y at Ward level**

ent degrees of autocorrelation in variable Y. The mean decreases as the level of autocorrelation of Y increase and the standard deviation in general decrease.

Figure 6.34 shows the distributions of level 1 and level 2 variance components for variable Y. The mean of the level 1 variance component decrease with the degree of autocorrelation. The standard deviation also decrease with the degree of autocorrelation. The mean of the level 2 variance component increases as the degree



**Figure 6.33: Correlation at Ward level with different autocorrelation**

of autocorrelation increases. When the Moran's  $I$  is 0.02, there are values of the variance components at level 2 that are negative including the mean, in fact there are more than 61% of the values that are negative. When the Moran's  $I$  is 0.1, 5% have negative values. This is observed because this case corresponds to effectively zero within Ward correlation. When the degree of autocorrelation is above 0.2, no negative values are observed. The mean and the standard deviation increase with the degree of autocorrelation increase.

Figure 6.35 shows the distribution of level 1 *pure correlation*. The mean of the level 1 *pure correlation* increases with the increase in the autocorrelation of variable  $Y$ . The mean, which is a bit lower than the initial Pearson correlation, approaches 0.3 but has a sudden nonlinear increase when the degree of autocorrelation of variable  $Y$  goes from 0.6 to 0.8. The standard deviation of the level 1 pure coefficient increase with the degree of autocorrelation of the variable  $Y$ .

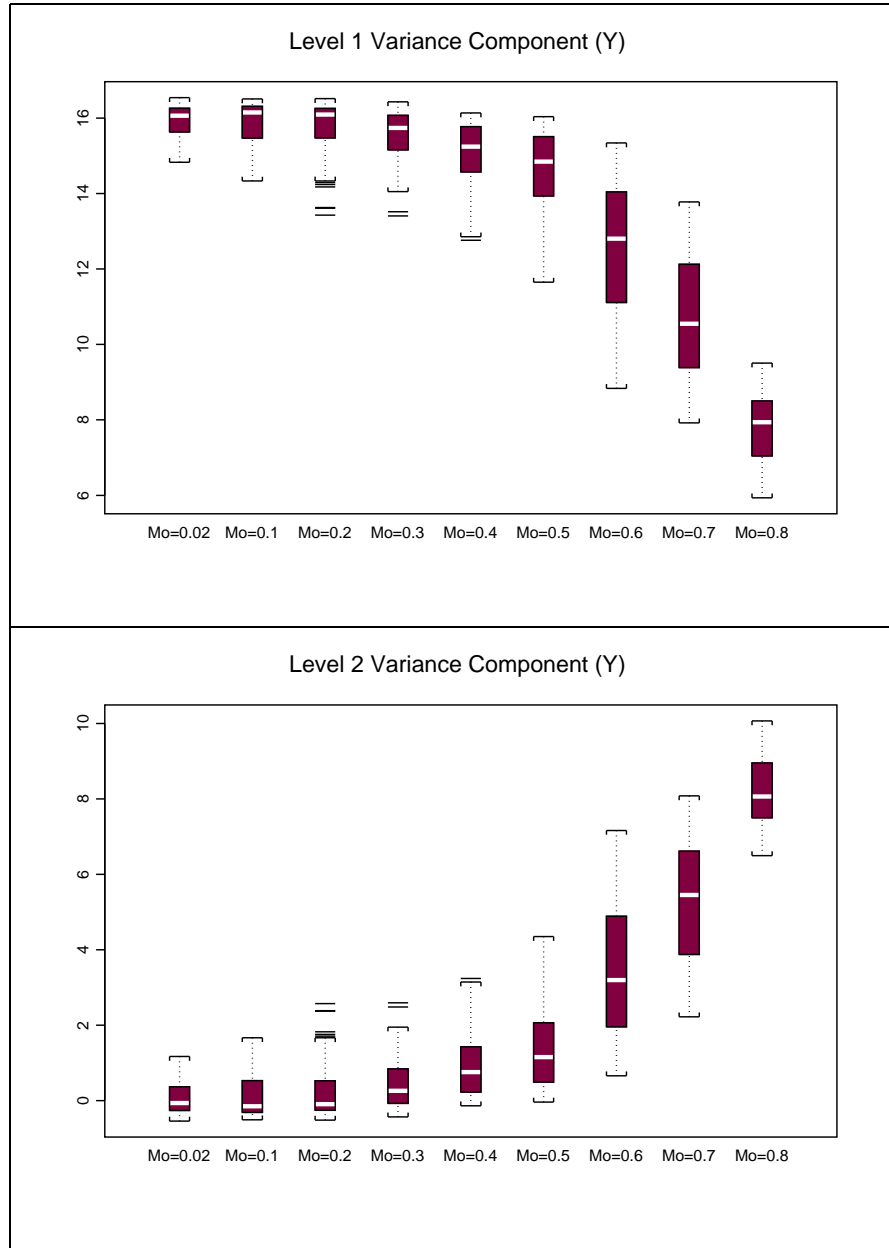


Figure 6.34: Level 1 and Level 2 Variance component of variable Y

Level 2 pure coefficients have unusual values when the degree of autocorrelation of variable Y are  $M_o=0.02$ , 0.1, 0.2. As noted before, when the autocorrelation is low the level 2 variance component is close to zero and negative estimates can be obtained using the moments approach. When the variable Y has  $M_o=0.02$  to  $M_o=0.04$  the mean fluctuates around 0.5 and the values range from -1.369 to 6.748. When variable Y has  $M_o=0.5$  to  $M_o=0.8$  the mean decreases as the degree of au-

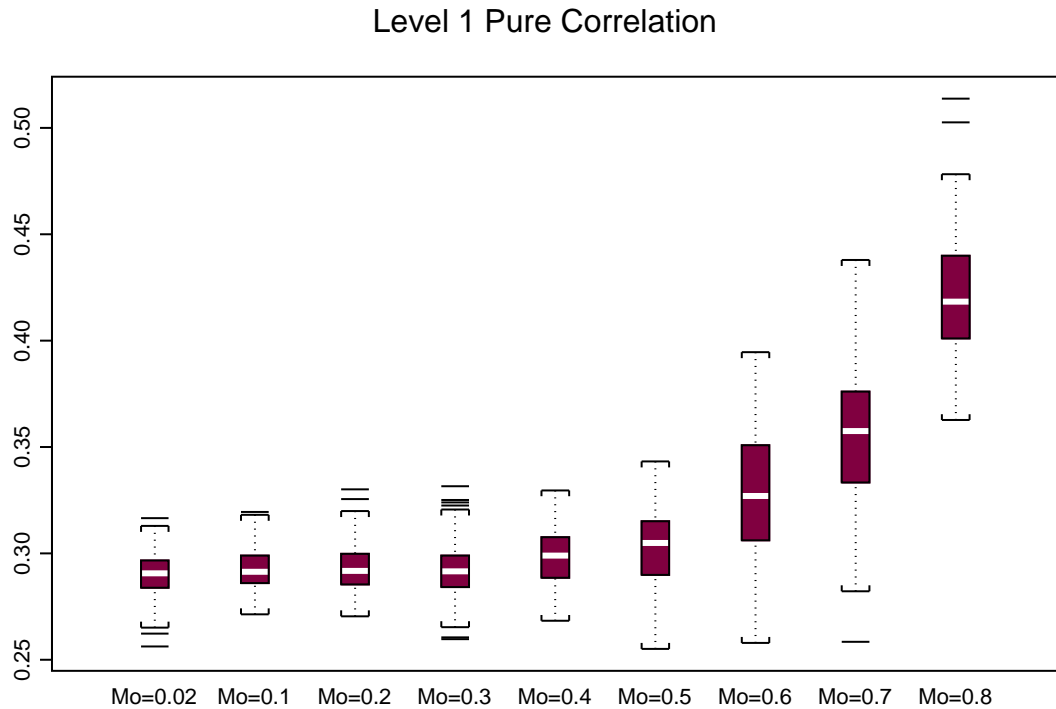


Figure 6.35: Level 1 Pure Correlation at different degrees of autocorrelation in Y

to correlation increase and the values are respectively, 0.32682, 0.19400, 0.13209, 0.07860.

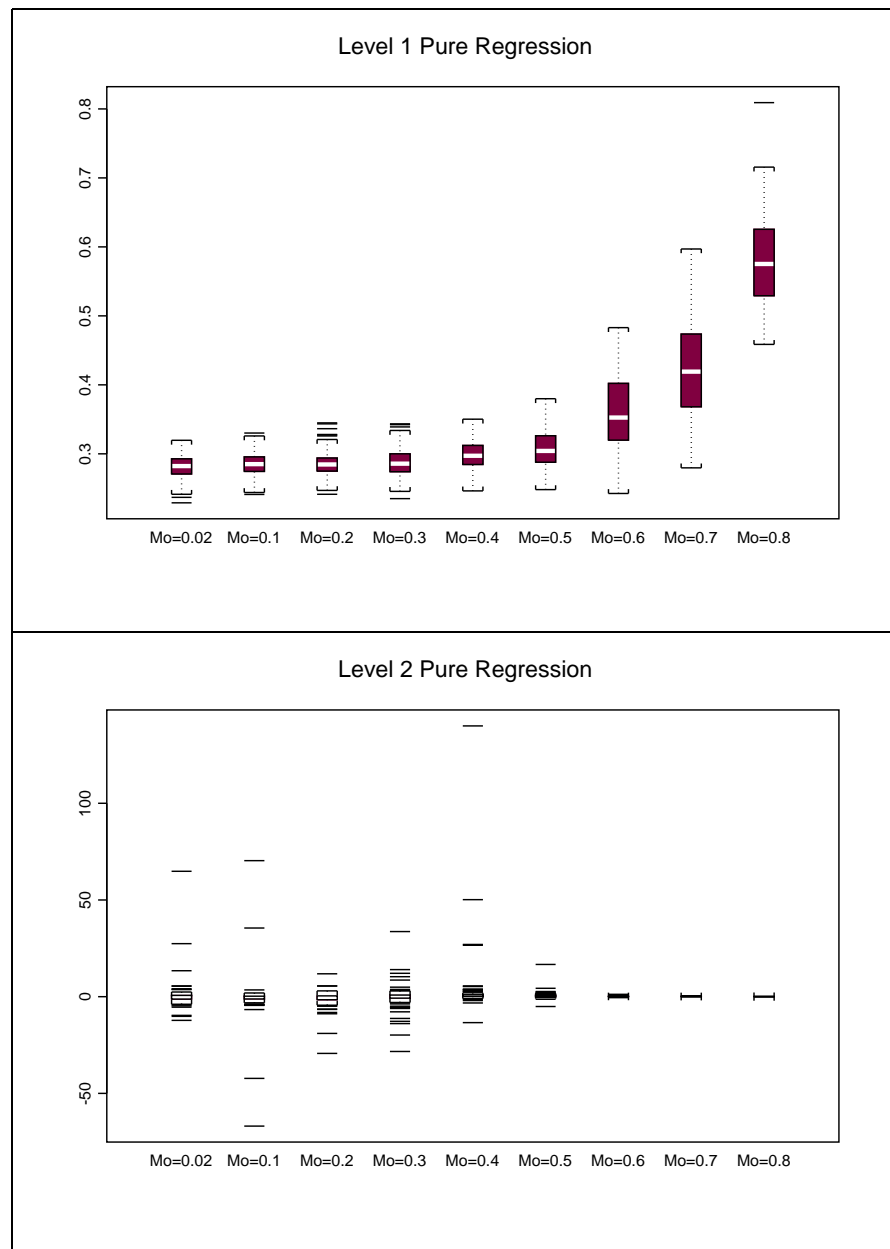
Figure 6.36 shows the distribution of level 1 and level 2 pure regression. The pattern as shown in the figure is similar to the previous figure.

#### Relationship between some statistics and the Moran's I with different proximity matrices:

Figure 6.37 shows the scatter plot between the different Moran's I with 'block' proximity of variable Y and the corresponding intra-area correlation. There is a very strong correlation between the IAC and the Moran's I with 'block' proximity.

#### Summary

This experiment examines the behavior of pertinent statistics when the degree of autocorrelation of one variable is different from the other. The data are generated in such a way that variable X has autocorrelation of 0.4 as measured by Moran's I

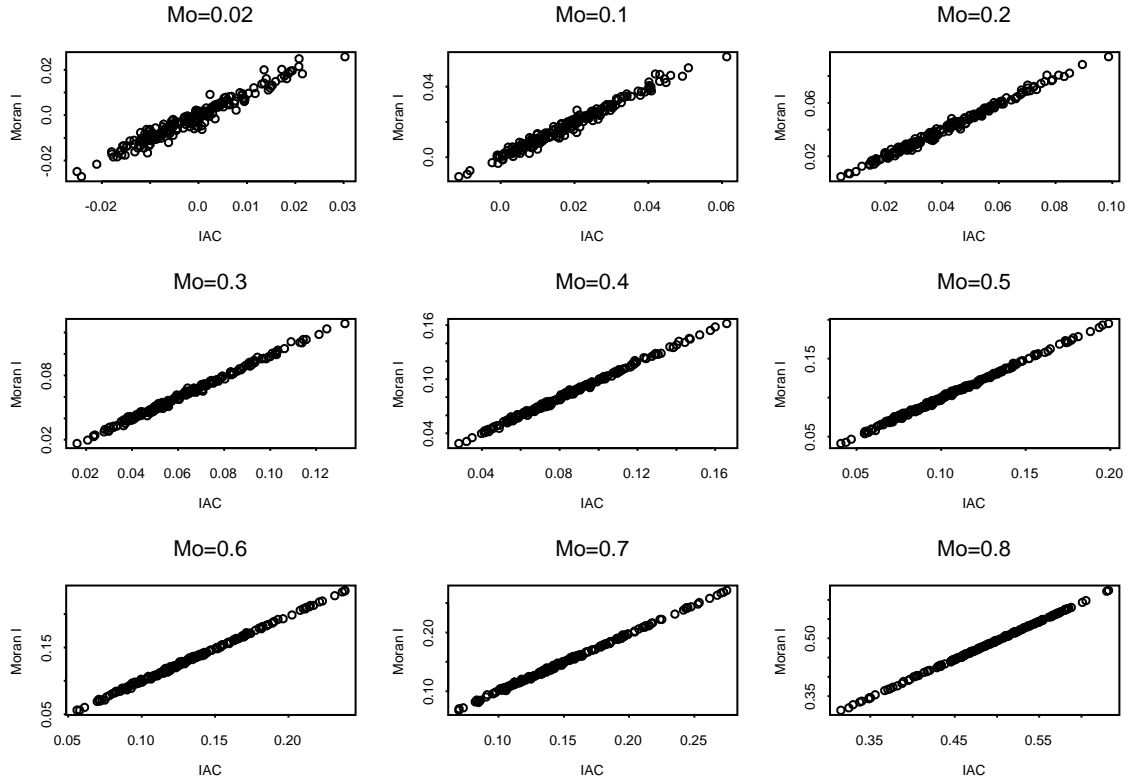


**Figure 6.36: Level 1 and Level Pure Regression**

with *Lag 1* proximity matrix and variable Y with Moran's I at 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. The initial correlation in all cases is 0.3 and both variables have mean 40 and variance 16 in all cases.

The mean of the direct correlations decreases as the level of autocorrelation of Y increases and the standard deviation in general decreases.

The mean of the level 1 *pure* correlation increases with the increasing degree of



**Figure 6.37: Scatter Plot of the Different Moran's I and the Corresponding IAC**

autocorrelation of variable  $Y$ . The mean which is a bit lower than the initial Pearson correlation approaches 0.3 but has a sudden nonlinear increase when the degree of autocorrelation of variable  $Y$  goes from 0.6 to 0.8. The standard deviation of the level 1 pure coefficient increases with the degree of autocorrelation of the variable  $Y$ .

Similar patterns are observed with the level 1 and level 2 *pure* regressions.

## 6.5 Summary

To look deeper into the behavior of the pure statistics and other common statistics, data sets are constructed with pre-determined characteristics using a data set generator based on Reynolds (1999). For more detailed description of the data set generator, refer to Reynolds (1999, pp. 10-17) and Reynolds and Amrhein(1997). The principles behind this data generator are used to generate data based on an actual

region that was divided into enumeration districts (EDs), the lowest geographical level in the 1991 UK population census for which aggregate data are released. These EDs are grouped into larger geographical areas called Wards. The region considered in this study is composed of the districts; Camden, Hackney, Haringey, and Islington. It comprises 1904 EDs nested into 92 Wards.

There were two general cases considered in this study, namely: Case 1, in which the variables have the same spatial autocorrelations; and Case 2 in which the variables have different spatial autocorrelations. For the first case, the data set being generated is composed of two variables X and Y and to make the analysis simple, the variables have the same mean(40) and the same variance (16) at the ED level. The variables also have a fixed Pearson correlation equal to 0.3 but they are generated in such a way that the autocorrelation of the variables are almost equal but varies (0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) as measured by Morans I and uses the weight matrix corresponding to queens case lag 1, ie  $w_{ij}=1$  if ED i and ED j are immediate neighbors and 0 otherwise. The generation of the data is repeated 3000 times for each degree of autocorrelation to generate the distributions of the pure statistics and other statistics. For each degree of autocorrelation described above, 1000 data points were selected at random from those that satisfy the required statistics and were analyzed. All analyses are made with 1000 data points in each level of autocorrelation.

Case 1: The variables have the same spatial autocorrelation

The mean level 1 *pure* correlation at different degrees of autocorrelation is not far from the Pearson correlation at ED level but the standard deviation becomes larger as the degree of autocorrelation increases. This means that the level 1 *pure* correlation has less chance of having a value far from the initial correlation when aggregated for cases when the degree of autocorrelation is low and this chance increases as the degree of autocorrelation increases. In the case of level 2 *pure* correlation, when both variables have low to medium autocorrelations, some cases are observed in which the estimated correlations at Ward level have values greater than 1.0, or less than -1.0 which is not a characteristic of a correlation coefficient. However,

the standard deviation of the level 2 pure correlations decreases as the degree of autocorrelation increases.

Similar patterns are observed with the *pure* regression coefficients. However, the mean level 1 *pure* regression is below the initial regression coefficient of 0.3. Extreme values were observed when the degree of autocorrelations of both variables are low.

There is a very strong positive relationships between the variance aggregation effects defined by  $(S_{XX}^{(2)}/S_{XX}^{(1)})$  and the intra-area correlation with "block" proximity.

There is almost perfect positive correlations between the intra-area cross-correlations and the bivariate Moran's I with low degrees of autocorrelations, and in the cases where the pairs of variables have high degree of autocorrelations, there is a perfect positive linear relationships.

Case 2: The variables have different spatial autocorrelation

The mean of the direct correlations decrease as the level of autocorrelation of Y increases and the standard deviation in general decrease.

The mean of the level 1 *pure* correlation increases with the increase in the degree of autocorrelation of variable Y. The mean is a bit lower than the initial Pearson correlation and approaches 0.3, but has a sudden nonlinear increase when the degree of autocorrelation of variable Y goes from 0.6 to 0.8. The standard deviation of the level 1 pure coefficient increases with the degree of autocorrelation of the variable Y.

A Similar pattern is observed with the level 1 and level 2 *pure* regressions.



# Chapter 7

## Conclusion

Analysis using spatial data is a multi-disciplinary subject attracting the attention of statisticians, geographers, physical and social scientists. The results of statistical analyses on the data available for areal units vary according to the definition of the areal units. This phenomenon is referred to as the modifiable areal unit problem (MAUP). The MAUP reflects the effects of scale and zoning. The scale effect refers to the changes in statistics that occur as the number of areal units into which the region is divided changes, whereas the zoning effects refers to the variation in results as the boundaries of the areal units change for a fixed scale. This study focuses on the scale effect, although some limited exploration of the zoning issue is conducted.

Statistical analysis based on data aggregated over spatial units often produce results that are very different from those obtained from analyzing corresponding individual or household level data (Steel, Holt and Tranmer, 1996). One of the reasons that it is necessary to aggregate data is to reduce the volume of data to be processed. Another reason is to protect the confidentiality of personal data (Openshaw and Albanides, 1996). A further reason is that there may be no interest in purely individual level relationships, but in relationships at some higher level or scale.

A large amount of the research on the MAUP has focused on revealing the problem and has been devoted to assessing the magnitude and impact on standard

statistics such as correlation and regression coefficients. The MAUP is due to the lack of independence or the presence of spatial correlation between different units in the population. Multilevel modeling is now a popular approach to reflect the association between different population units. This thesis examines some statistics derived from a simple multilevel model to clarify the causes of one aspect of the MAUP, the scale or aggregation effect. The thesis also look into the behavior of the the so called "pure statistics", that can be calculated from the results of a simple multilevel model and attempts to provide information about effects at a particular level after removing effects from the other levels.

The thesis is novel in that pure coefficients and variance components are investigated using artificially-generated data as well as real data. Also, pure statistics are compared with direct statistics and connection between the MAUP, multilevel model and spatial autocorrelation are formulated.

## 7.1 Summary and Conclusions

Issues associated with scale effects in a simple multilevel model are considered in this thesis. The thesis focuses on the relationship between a simple multilevel model and a key aspect of the MAUP; the scale effect, and spatial autocorrelation. The flow of the thesis includes an introduction followed by chapters giving a review of the literature on the MAUP and then theoretical aspects of the possible causes of the MAUP from a multilevel perspective. Three empirical chapters then follow.

Several experiments using a number of data sets are conducted in this thesis to analyze the behaviors of the direct statistics and pure statistics. The experiments are based on simulated data, real data and simulations based on real data. The first of the three experiments examines the results for direct statistics and pure statistics derived from a simple multilevel model using artificially-generated data sets in a 100x100 square grid. The next chapter examines the results on direct and pure statistics using real data drawn from two districts of London Boroughs. The sources of data used are the 1991 UK Census and 1991 Sample of Anonimized Records

(SARs). The analytical approach is then repeated using artificially-generated data based on the work of Reynolds (1998). Several situations are considered, involving varying the spatial autocorrelation and initial correlations of variables. These experiments examine the impact of varying scale and the degree of autocorrelation on the distribution of the statistics, including the mean and median of the distribution and the dispersion as reflected standard deviations and the boxplots.

The experiments in this thesis enable us to investigate a number of questions concerning scale effects.

- What is the impact of spatial autocorrelation on the scale effect on standard, or direct, correlation and regression coefficients? In particular we investigate how the level of spatial autocorrelation impacts the distribution of correlation and regression coefficients including the mean and median of the distribution and the dispersion as reflected in the standard deviations and boxplots. These results add to the evidence concerning the MAUP.
- Are the so-called pure correlation and regression coefficients obtained from the simple multilevel model affected by the MAUP?
- What is the impact of spatial autocorrelation on the scale effect of estimated pure correlation and regression coefficients? In particular we investigated how the level of spatial autocorrelation impacts the distribution of correlation and regression coefficients including the impact on the mean and median of the distribution and the dispersion as reflected in the standard deviation and boxplots.
- Is the impact of the MAUP on the pure correlation and regression coefficients less than, or different from, that on the direct coefficients?
- Can we predict the impacts of scale and zoning on features of the distribution of the direct or pure coefficients? This question included finding some indicators may include intra-class correlation.

Here we summarize how the results have provided evidence on the questions.

From Chapter 2 we have the first two aggregation rules (Amrhein, 1995) are:

- The mean does not display any pronounced aggregation effects (scale or zonation) at any level of aggregation used in the study.
- The variance does not display any pronounced scale effect beyond those expected from the decrease in the number of observations. However, it was noted that scale-specific variance values cannot be imputed to other scales without adjusting for the change in the number of reporting units.

Also, from Chapter 2 we have Steel and Holt (1996) rules of random aggregation:

- The expected value of weighted group-level statistics are not affected by aggregation and that any observed change is due to random variation.
- The variance of the weighted group-level statistics are affected mainly by the number of groups in the analysis. The variation will be high when the number of groups is small.

The above rules correspond to situations where there is no spatial autocorrelation present. The following results will extend the above results, by considering situations in which there is spatial autocorrelation present.

### 7.1.1 The Mean and the variance

The population weighted group level mean is identical to the mean calculated at the individual level (see equation 3.13, p23) and is not affected by aggregation and the degree of autocorrelation. This is confirmed empirically. Consequently and population weighted mean and its variance are not affected by aggregation.

The scale effect on the population weighted variance arises because as scale increases the contribution of the level 2 variance component increases from 1 to approximately  $\bar{N}^*$ , whereas the contribution of the level 1 variance component is virtually unchanged (see equation 3.29 and 3.30, p26). For equal size groups the intraclass correlation is equal to the Moran I statistics with spatial proximity weight

equal to 1 if a pair of individuals is within the same group. Thus the intraclass correlation can be regarded as a measure of average spatial autocorrelation within a group. Even when the group sizes vary the relationship between the intraclass correlation and Moran I is still very strong, as shown in figures 6.14, 6.15, and 6.37 (p193, 194 and 214 respectively). As we aggregate we would expect the intra-area correlation to decrease as we are including more units and increasing the average distance between units within a group. Figure 4.41 (p120) shows that the intraclass correlation decrease non-linearly as  $\bar{N}$  increases and, for fixed  $\bar{N}$ , increases with the level of autocorrelation at the individual level. Hence, for a fixed autocorrelation the weighted variance will increase with  $\bar{N}$  but not linearly and the rate of increase slows as the number of groups become small.

The distribution of the weighted group level variance becomes more disperse as the scale increases and the number of groups become small, particularly for 24 and 4 groups. This is due to the small number of degrees of freedom involved in calculating the variance, which is  $M-1$ , where  $M$  is the number of groups. The dispersion is reflected in the standard deviations of the weighted variances, which also increases with the degree of spatial autocorrelation. The change in standard deviation with scale is more than would be expected through the change in  $M$ . If the weighted variance behaved as proportional to a  $\chi^2_{M-1}$  random variable the ratio of its standard deviation to the mean, which is its coefficient of variation (CV), would be  $\sqrt{2/(M-1)}$ , which would be the case for no autocorrelation. Results in table 4.8 (p48), table 4.27 (p68), and table 4.43 (p82) give values a little less than these theoretical values. Further evidence is given in table 6.1 on page 162, although the CV of the weighted variance is a little larger than in the case of no autocorrelation.

The unweighted variances are approximately  $\bar{N}^{-2}$  times the corresponding weighted variance. Consequently they decrease with scale, but increase, for a given scale, with an increase in autocorrelation. If there is no autocorrelation their mean will be the individual level variance divided by  $\bar{N}^2$ . Again the CV is a little less than  $\sqrt{2/(M-1)}$ . Relevant results are given in tables 4.10, 4.28 and 4.44 on pages 50, 68, and 84 respectively.

### 7.1.2 The direct correlation and regression coefficients

When there is no spatial autocorrelation equation 4.2 (p51) suggest that the group level correlation will be close to the individual level correlation, although there will be a tendency to increase in absolute value when the number of groups is quite small. Equation 4.3 (p51) gives a theoretical formula for the standard deviation of the group level correlation when no spatial correlation is present and predicts that it increases as the number of groups decreases, i.e. as scale increases. These results are confirmed empirically in table 4.65 and 4.66 on page 101 and 102 respectively.

Tables 4.72 and 4.79 on pages 111 and 124 respectively summarize the results of the experiments for the direct correlations with varying autocorrelations. Based on the results of these experiments we see that when the autocorrelation is very low the mean of the direct correlation is close to the individual level correlation. When autocorrelation is present the mean of the direct correlation increases with the degree of aggregation, that is as the scale increases. The main increase was evident as the number of groups decreased to 625 and thereafter the increase was minimal.

For a given scale the degree of autocorrelation affects the mean of the direct correlation, initially increasing as the autocorrelation increased, but then it decreases when one or both of the variables has high autocorrelation. As the level of aggregation increases the dispersion of the distribution of the direct correlation increases, which is reflected in the standard deviation. When there are only 25 groups there are some values of the direct correlation less than the individual level correlation. The standard deviation of the direct correlation is only moderately affected by the degree of autocorrelation. In going from very low to low autocorrelation the standard deviation decreases, but then increases as the autocorrelation increases further.

Analysis of real data confirms the general pattern of correlation increasing in absolute value with aggregation, although there are exceptions, see table 5.21 (p154).

Simulations based actual boundaries for varying individual level correlation and autocorrelation show that the Ward level correlation fluctuate around the corresponding unit level correlations, see figure 6.19 on page 197. The lack of any scale

effect may be due to the average number of EDs per ward being only 21. The dispersion tends to decrease as the autocorrelation increases.

The effects for direct correlation are similar to those of the direct correlations.

### 7.1.3 The pure coefficients

One results of the experiments tackle the basic issue of how multilevel model is affected by one aspect of the MAUP, the scale effect. This is an important issue from an applied perspective, since multilevel modeling is sometimes suggested as an approach to handling spatial aggregated data that comprise different levels.

The calculation of estimates of variance components associated with a simple multilevel model enables the calculation of 'pure' correlation and regression coefficient that attempt to separate effects occurring at the individual and group level.

When there is no spatial autocorrelation the mean and median of the estimated level 1 and level 2 variance components are very close to the true values and the standard deviation of the estimates at both levels decreases as aggregation is increased (see table 4.67, p104). As expected, negative estimates of the level 2 variance components occur as they are unbiased estimates of a true parameter that is zero.

When there is no spatial autocorrelation the mean and median of the estimated level 1 variance components are not affected by aggregation (see table 4.67 on page 104). The estimates of the level 2 variance components are very unstable, with very large standard deviations, which affects their mean, but the medians are close to zero. The level 1 variance component estimates have standard deviation close to those of the individual level and are not affected appreciably by aggregation.

When there is no spatial autocorrelation the mean of the level 1 correlation is the same as the individual level correlation and the level 1 regression is smaller than the individual level.

In general the estimated level 1 and level 2 variance components will add to the individual level variance as shown by equation 3.31 (p27). As the scale increases we should expect the estimated level 1 variance component to increase when there is spatial autocorrelation, as more dissimilar units are included in each group and

hence the estimated level 2 variance decreases. This is seen in tables 4.13, 4.30 and 4.46 on pages 55, 72, 87 respectively. The rate of the increase in the level 1 variance component estimates and the decrease in the level 2 estimates depends on the level of the autocorrelation, being greater with higher levels of autocorrelation.

The results of the experiments show little or no scale effect in the mean of the estimated level 2 correlation and regression coefficients, with very small increases as the scale increases, as shown in tables 4.54, 4.55 and figures 4.46 and 4.48 on pages 96, 96, 125 and 126 respectively. However, the means of these coefficients are affected by the degree of autocorrelation, decreasing as the autocorrelation increases.

The standard deviation of the estimated level 2 correlation and regression coefficients increase as scale increases, but is reasonably stable once the autocorrelation is higher than the direct correlation.

The means of the level 1 correlation, when there is spatial autocorrelation, are affected by aggregation, starting below the individual level correlation but approaching it as the number of groups decreases. This is because, as the number of groups becomes smaller, the groups become larger and the individuals within them become more like the whole population. Spatial autocorrelation has little effect on the mean of the level 1 correlations.

The standard deviation of the estimated level 1 correlation is relatively stable and increases only slightly with scale and autocorrelation. The standard deviation tends to be less than that of the individual level correlation, but approaches it as the number of group becomes less. The standard deviation for the level 1 correlation coefficients is much less than for the level 2 coefficients. Similar results are obtained for the regression coefficient.

Decisions based on a study using aggregated data need to be considered carefully. It has been shown in this thesis and other research that relationships between variables in one scale may be different from those found in other scales. Our research suggests the one possible initial step in the investigation is to consider the intra-area correlation which is a measure of the average spatial autocorrelation of the variables. The results of this thesis can be used as a basis for the decision on



which statistics to use in the decision making.

#### 7.1.4 Approach to Aggregation

The results of the analyses in this thesis suggest that there is not necessarily an individual level of aggregation. Direct analyses of aggregated data are affected by relationships operating at the individual level and higher geographic levels. Use of a multilevel modelling approach allows some separation of the effects at each level and attempts to eliminate the contamination of any level on another. Which levels of pure coefficient are relevant depends on the particular application. In some applications the focus may be on individual relationships, having removed the effect of higher levels. However, in many geographical applications the interest will be in relationships that are specifically at the aggregation level remaining the purely individual level effect. There may be interest in both individual and area level effects, but they need to be separated for proper analysis. It is therefore important that a researchers clearly specifies the relevant level or levels for their analysis.

### 7.2 Further Research and Development

Scale effects of pure coefficients derived from a simple multilevel model were investigated in this thesis. Further research can be done on pure coefficients for more complex multilevel models. The experiments in this thesis show that the result of a simple multilevel model are affected by the MAUP but not in the same way as standard correlation and regression coefficients. They also confirm previous studies that demonstrate that the MAUP affects standard analytical statistics. The role of spatial autocorrelation is important and the simple multilevel model will usually be an approximation to a more complex pattern of autocorrelation that applies in a real population. Hence, a further step would be to investigate the MAUP when correlations across areal units are incorporated in the model underpinning the analysis.

In Chapter 4 a number of combinations of variables in terms of the level of

autocorrelation were considered, which give a good picture of the impact of different degrees of autocorrelation. Further research can be carried out to look into the effects of pure coefficients for more combinations of variables in terms of their spatial autocorrelation.

The role of spatial autocorrelation has been demonstrated. More work on the how the values of the individual level correlation and the spatial autocorrelation interact in the MAUP on the direct and pure coefficients is desirable. The methods used in this thesis in the analyses of pure coefficient can be used to analyze multiple regressions and other multivariate techniques. The data generator can be utilized to generate variables with different initial conditions.

# Appendix A

## Dataset Simulation Codes

### A.1 Data Set Generator for a square grid

This program is to generate data sets on a square grid having a dimension of 100x100.

```
#####  
# data generation based Moving average #  
# method                               #  
# Individual level (10000 areal units #  
# in a square grid                     #  
#####
```

```
module(spatial)  
a77<-1  
n<-10000  
r<-100  
c<-100  
#initialisation of matrices
```

The initialization of the variables was omitted to save space

```
p<-1  
repeat  
{ if (p>c0) break  
  set.seed(12+p)  
  AA1<-matrix(rnorm(r*c,0,4),ncol=c)  
  j<-1
```

```

k<-1
AA2[j,k]<-(1/3)*(AA1[j,k+1]+AA1[j+1,k+1]+AA1[j+1,k])
j<-1
k<-r
AA2[j,k]<-(1/3)*( AA1[j+1,k]+AA1[j+1,k-1]+AA1[j,k-1])
j<-c
k<-1
AA2[j,k]<-(1/3)*( AA1[j,k+1]+AA1[j-1,k+1]+AA1[j-1,k])
j<-c
k<-r
AA2[j,k]<-(1/3)*(AA1[j-1,k]+AA1[j-1,k-1]+AA1[j,k-1])
j<-1
for(k in 2:(c-1))
AA2[j,k]<-(1/5)*( AA1[j,k+1]+AA1[j+1,k+1]+AA1[j+1,k]+AA1[j+1,k-1]+AA1[j,k-1])
j<-c
for(k in 2:(c-1))
AA2[j,k]<-(1/5)*( AA1[j,k+1]+AA1[j-1,k+1]+AA1[j-1,k]+AA1[j-1,k-1]+AA1[j,k-1])
k<-1
for(j in 2:(c-1))
AA2[j,k]<-(1/5)*( AA1[j+1,k]+AA1[j+1,k+1]+AA1[j,k+1]+AA1[j-1,k+1]+AA1[j-1,k])
k<-r
for(j in 2:(c-1))
AA2[j,k]<-(1/5)*( AA1[j+1,k]+AA1[j+1,k-1]+AA1[j,k-1]+AA1[j-1,k-1]+AA1[j-1,k-1])
for(j in 2:(r-1))
for(k in 2:(c-1))
AA2[j,k]<-(1/8)*(AA1[j-1,k]+AA1[j+1,k]+AA1[j,k-1]+AA1[j,k+1]+
AA1[j-1,k-1]+AA1[j+1,k-1]+AA1[j+1,k+1]+AA1[j-1,k+1])
ij100<-expand.grid(i=seq(1,100,len=100),j=seq(1,100,len=100))
set.seed(5+p)
r<-100
c<-100
K<-1
EijX<-matrix(rnorm(r*c,0,2), ncol=c)
EeijX<-K*c(EijX)
Xij1<-c(AA2)+c(EeijX)
Xij1.matrix<-matrix(Xij1, ncol=100)

```

```

      A1<-c(AA1)
      A2<-c(AA2)
      ErrX<-c(EeijX)
      X2<-Xij1
test10000X2<-cbind(ij100,A1,A2,ErrX,Xij1)
set.seed(28+p)
      r<-100
      c<-100
      e<- matrix(rnorm(r*c,0,4),ncol=c)
      BB1<-matrix(10+c(AA1+e),ncol=c)

j<-1
k<-1
      BB2[j,k]<-(1/3)*(BB1[j,k+1]+BB1[j+1,k+1]+BB1[j+1,k])
j<-1
k<-r
      BB2[j,k]<-(1/3)*( BB1[j+1,k]+BB1[j+1,k-1]+BB1[j,k-1])
j<-c
k<-1
      BB2[j,k]<-(1/3)*(BB1[j,k+1]+BB1[j-1,k+1]+BB1[j-1,k])
j<-c
k<-r
      BB2[j,k]<-(1/3)*( BB1[j-1,k]+BB1[j-1,k-1]+BB1[j,k-1])
j<-1
for(k in 2:(c-1))
      BB2[j,k]<-(1/5)*( BB1[j,k+1]+BB1[j+1,k+1]+BB1[j+1,k]+BB1[j+1,k-1]+BB1[j,k-1])
j<-c
for(k in 2:(c-1))
      BB2[j,k]<-(1/5)*( BB1[j,k+1]+BB1[j-1,k+1]+BB1[j-1,k]+BB1[j-1,k-1]+BB1[j,k-1])
k<-1
for(j in 2:(c-1))
      BB2[j,k]<-(1/5)*( BB1[j+1,k]+BB1[j+1,k+1]+BB1[j,k+1]+BB1[j-1,k+1]+BB1[j-1,k])
k<-r
for(j in 2:(c-1))
      BB2[j,k]<-(1/5)*( BB1[j+1,k]+BB1[j+1,k-1]+BB1[j,k-1]+BB1[j-1,k-1]+BB1[j-1,k-1])
      for(j in 2:(r-1))
            for(k in 2:(c-1))

```

```

      BB2[j,k]<-(1/8)*(BB1[j-1,k]+BB1[j+1,k]+BB1[j,k-1]+BB1[j,k+1]+
      BB1[j-1,k-1]+BB1[j+1,k-1]+BB1[j+1,k+1]+BB1[j-1,k+1])
ij100<-expand.grid(i=seq(1,100,len=100),j=seq(1,100,len=100))
set.seed(10+p)
K<-1
      EijY<-matrix(rnorm(r*c,0,2),ncol=c)
      EeiY<-K*c(EijY)
      Yij1<-c(BB2)+c(EeiY)
Yij1.matrix<-matrix(Yij1, ncol=100)
      B1<-c(BB1)
      B2<-c(BB2)
ErrY<-c(EeiY)
      Y2<-Yij1
      m<-1:10000
      X1<-X2*((sqrt(6))/ sqrt(var(X2)))
      Y1<-Y2*( (sqrt(8))/ sqrt(var(Y2)))
      X<-X1+(0.005-mean(X1))
      Y<-Y1+(10-mean(Y1))
      Z<-X+Y
test10000XY<-cbind(ij100,m, X, Y,Z)
test10000Y1<-cbind(ij100,A1,A2,ErrY,Yij1)
      nobs<-10000
      nzone<-2500
      zonesize<-ceiling(nobs/nzone)
      zfil2500<-matrix(rep(0,(nobs/(nzone/(sqrt(nobs)/sqrt(zonesize))))
      *(nzone/(sqrt(nobs)/sqrt(zonesize)))),ncol=(nzone/(sqrt(nobs)/sqrt(zonesize))))
zfil2500[,1]<-rep(rep(1:(sqrt(nobs)/sqrt(zonesize)),each=ceiling(sqrt(zonesize)))
      ,ceiling(sqrt(zonesize)))
      for (i in 2:(nzone/max(zfil2500[,1])))
      { zfil2500[,i]<-zfil2500[,1]+(i-1)*max(zfil2500[,1]) }
t10000withzone2500a<-cbind(m,test10000XY,c(zfil2500))
      zonemean <- function(spat)
      {out<-matrix(rep(0,length(spat[,4])),ncol=max(spat[,4]))
      for (i in 1:max(spat[,4]))
      {out[,i]<-spat[spat[,4]==i][1:(length(spat[,4])/max(spat[,4]))] }
      return(colMeans(out)) }

```

---

```

test2500x<-cbind(X,Y,Y,c(zfil2500))
  spat<-test2500x
  z2500x<-zonemean(test2500x)
test2500y<-cbind(Y,X,X,c(zfil2500))
  spat<-test2500y
  z2500y<-zonemean(test2500y)
test2500z<-cbind(Z,X,X,c(zfil2500))
  spat<-test2500z
  z2500z<-zonemean(test2500z)
  x1<-X
  y1<-Y
  vvx<-var(x1)
  vvy<-var(y1)
  cxy<-cov(x1,y1)
  x11<-z2500x
  y11<-z2500y
  wgt<-nobs/nzone
z2500x1<-(sum(wgt*z2500x))/n
z2500y1<-(sum(wgt*z2500y))/n
  m<-2500
  vz2500x<-(sum(wgt*((z2500x-z2500x1)^2)))/(m-1)
  vz2500y<-(sum(wgt*((z2500y-z2500y1)^2)))/(m-1)
cz2500xz2500y<-(sum(wgt*(z2500x-z2500x1)*(z2500y-z2500y1)))/(m-1)
  l2cor11<-cz2500xz2500y/((vz2500x*vz2500y)^.5)
  cvaf<-cov(x11,y11)
  bl1<-cvaf/(var(x11))
  unwr<-cor(x11,y11)
  m1<-mean(x1)
  m2<-mean(y1)
  v1<-var(x1)
  v2<-var(y1)
  cor1<-cor(x1,y1)
  af<-cor1*((v1*v2)^.5)
  b1<-af/v1
  nbar<-n/m
  nobar<-(sum(wgt^2))/n

```

```

    nastbar<-nbar+((nbar-nobar)/(m-1))
    lev2z2500x<-(vz2500x-vvx)/(nastbar-1)
    lev2z2500y<-(vz2500y-vvy)/(nastbar-1)
    lev1z2500x<-vvx-lev2z2500x
    lev1z2500y<-vvy-lev2z2500y
    iacz2500x<-lev2z2500x/(lev2z2500x+lev1z2500x)
    iacz2500y<-lev2z2500y/(lev2z2500y+lev1z2500y)
    lev2cov1<-(cz2500xz2500y-af)/(nastbar-1)
    lev1cov1<-af-lev2cov1
    iaccl11<-(lev1cov1)/((v1*v2)^.5)
    iaccl21<-(lev2cov1)/((v1*v2)^.5)
    lev1rho11<-lev1cov1/((lev1z2500x*lev1z2500y)^.5)
    lev2rho11<-lev2cov1/((lev2z2500x*lev2z2500y)^.5)
    b112<-lev2cov1/lev2z2500y
    b111<-lev1cov1/lev1z2500y
    sxxedlev[p]<-vvx
    syyedlev[p]<-vvy
    sxyedlev[p]<-cxy
    sxxwardlev[p]<-var(x11)
    syywardlev[p]<-var(y11)
    sxywardlev[p]<-cov(y11,x11)
    rwardlev[p]<-cz2500xz2500y/vz2500x
    rUnWtdward[p]<-b11
    cwardlev[p]<-l2cor11
    cUnWtdward[p]<-unwr
    redlev[p]<- cxy/vvx
    cedlev[p]<- cxy/(sqrt(vvx* vvy))
    b11ed<-cz2500xz2500y/vz2500x
    l2cor11<-cz2500xz2500y/((vz2500x*vz2500y)^.5)
    sx2wardlev[p]<-lev2z2500x
    sy2wardlev[p]<-lev2z2500y
    sxy2wardlev[p]<-lev2cov1
    sx1wardlev[p]<-lev1z2500x
    sy1wardlev[p]<-lev1z2500y
    sxy1wardlev[p]<-lev1cov1
    iacwardlevx[p]<-iacz2500x

```



```

    iacwardlevy[p]<-iacz2500y
    iacc2wardlev[p]<-iacc121
    iacc1wardlev[p]<-iacc111
    pc2wardlev[p]<-lev2rho11
    pc1wardlev[p]<-lev1rho11
    pr2wardlev[p]<-bl12
    pr1wardlev[p]<-bl11
    vz2500xwtd[p]<-vz2500x
    vz2500ytd[p]<-vz2500y
    cz2500xz2500ytd[p]<- cz2500xz2500y
    m10000x[p]<-mean(X)
    m10000y[p]<-mean(Y)
    m2500x[p]<-mean(z2500x)
    m2500y[p]<-mean(z2500y)
    datarz2500<-cbind(sxxedlev,syyedlev,sxyedlev,redlev,cedlev,sxxwardlev,
                      syywardlev,sxywardlev,rwardlev,rUnWtdward,cwardlev,
                      cUnWtdward,sx2wardlev,sy2wardlev,sxy2wardlev,sx1wardlev,
                      sy1wardlev,sxy1wardlev,iacwardlevx,iacwardlevy,iacc2wardlev,
                      iacc1wardlev,pc2wardlev,pc1wardlev,pr2wardlev,pr1wardlev,
                      vz2500xwtd,vz2500ytd,cz2500xz2500ytd

    p<-p+1
  }

```

The program will then be run again for  $m=625,400,100,25$ , and 4

## A.2 Data sets for a region

The following R code was designed to create data sets with two variables X and Y with specific mean, variance, initial correlation, initial autocorrelation (Moran's I), and number of iterations. The code applies to the region being defined in the study. Other computations of pertinent statistics are also included in the code.

Required inputs are as follows:

- numiter: The number of iteration
- dmx: The initial mean of X
- dmy: The initial mean of Y
- vvx: The initial variance of X
- vvy: The initial variance of Y
- dcr: The initial correlation of X and Y
- mox: The initial autocorrelation of X
- moy: The initial autocorrelation of Y

Also needed are (1)proximity tables created using GeoDa as a .txt file (WtEDLag01Queen.txt) (2) Number of EDs per Ward (NumEdPerWard.txt), (3)weight to be used in the computation of Moran's I (EDWARD333.txt)

### Data Set Generator

```
iden2<-diag(1,1904,1904)
r<-1904
c<-1904
one1<-matrix(1, ncol=c, nrow=r)
onediv<-one1*(1/r)
onedivbyn<-matrix(onediv, ncol=c)
M<-iden2-onedivbyn
ww<- scan( "WtEDLag01Queen.txt")
```

```

spw<- function(ww)
{w1<-matrix(0, ncol=ww[1], nrow=ww[1])
  nwo<-3
  nw<-ww[3]
  for(i in 1:ww[1])
  {
    if (nw>0)
      { for(k in 1:nw)
        { w1[i,ww[nwo+k]]<-1
        }
      }
    nwo<-nwo+nw+2
    nw<-ww[nwo]
  }
  w1
}

w<- spw(ww)
C<- spw(ww)
one<-matrix(1, ncol=1, nrow=r)
onetrans<-matrix(1, ncol=r, nrow=1)
csn<-onetrans%%one
csd<-onetrans%%C%%one
cs1<-csn/csd
rm(one, onetrans)
cs2<- cs1[1]
Cs<-cs2*C
MCs<-M%%Cs
MCsM<-MCs%%M
ev<-eigen(MCsM)
eval0<-ev$val
eval1<-matrix(eval0, nrow=1)
EVAL<-eval1
evec0<-ev$vec
evec1<-matrix(evec0, nrow=r)
EVEC<-evec1

```

---

```

zz<-read.table("NumEdPerWard.txt")
      xx<-cbind(zz[,1], zz[,2], zz[,3])
      x11<-c(rep(0,92))
      y11<-c(rep(0,92))
xxx<- scan("EDWARD333.txt")
      www<-xxx
      ww1<- spw(www)

c0<-numiter
p<-1
# Initialization of Variables ( Omitted to save space)#
repeat
  {if(p>c0) break
    e1<-EVEC[,1]
    e2<-EVEC[,2]
    e6<-cbind(e1,e2)
    V<-e6
    bb<-cov(e6)
# The Desired variance-covariance #
# let the var(e2) be the diagonal of the matrix #
    dmx<-dmx
    dmy<-dmy
    vvX<-vvX
    vvY<-vvY
    dcr<-dcr
    d11<-vvX
    d12<-dcr*((vvX*vvY)^.5)
    d21<-dcr*((vvX*vvY)^.5)
    d22<-vvY
    d<-c(d11,d21,d12,d22)
    dm<-matrix(d, ncol=2)
    dd<-matrix(d, ncol=2)
    B<-chol(bb)
    D<-chol(dd)
    A<-solve(B,D)
#the desire MC's

```

```

mcd0<-c(mox,moy)
mcd1<-matrix(mcd0, nrow=1)
MC<-mcd1
m1<-MC[,1]
m2<-(((A[1,2]^2)*MC[,1])+((A[2,2]^2)*MC[,2]))/((A[1,2]^2)+(A[2,2]^2))
l1<-m1
l2<-((m2*((A[1,2]^2)+(A[2,2]^2)))-((A[1,2]^2)*l1))/((A[2,2]^2))
rmc0<-c(l1,l2)
RMC<-matrix(rmc0, nrow=1)
set.seed(19+p)
b0<-runif(2,min=0, max=1)
b1<-matrix(b0, nrow=1)
lam11<-as.integer(runif(1,1,440))
lam12<-as.integer(runif(1,1,440))
lam21<-as.integer(runif(1,625,1100))
lam22<-as.integer(runif(1,750,1200))
a1<-(((EVAL[,lam11]-RMC[,1])/(RMC[,1]-EVAL[,lam21]))*(b1[,1]^2))^(.5)
a2<-(((EVAL[,lam12]-RMC[,2])/(RMC[,2]-EVAL[,lam22]))*(b1[,2]^2))^(.5)
v1<-a1*EVEC[,lam21]+b1[,1]*EVEC[,lam11]
v2<-a2*EVEC[,lam22]+b1[,2]*EVEC[,lam12]
Vj<-cbind(v1,v2)
s1<-sqrt(var(v1))
s2<-sqrt(var(v2))
sd1<-sqrt(vvx)
sd2<-sqrt(vvy)
v111<-v1*(sd1/s1)
v222<-v2*(sd2/s2)
VV<- cbind(v111,v222)
XX<-VV%*%A
XXX1<-XX[,1]*(sd1/sqrt(var(XX[,1])))
XXX2<-XX[,2]*(sd2/sqrt(var(XX[,2])))
x11<-XXX1+(dmx-mean(XXX1))
y11<-XXX2+(dmy-mean(XXX2))
#####
# The generated data #
#####

```

```

x1<-x11
y1<-y11
z1<-x1+y1
# Computation of the Moran's I
morani<-
function(x, w, k = 0, rescale = T, rescalepart = T)
{
  n <- length(x)
  s2 <- 0
  if(rescale) {
    for(i in 1:n)
    {
      if(sum(w[i, ]) > 0)
      {
        w[i, ] <- w[i, ]/(sum(w[i, ]))
      }
    }
  }
  for(i in 1:n)
  {
    s2 <- s2 + (sum(w[i, ]) + sum(w[, i]))^2
  }
  sumw <- sum(w)
  mx <- mean(x)
  z <- x - mx
  m <- t(z) %*% w %*% z
  sumz2 <- sum(z^2)
  m <- (n * m)/(sumw * sumz2)
  pm <- (n * pm)/(sumw * sumz2)
  s1 <- 0.5 * sum((w + t(w))^2)
  Ei <- -1/(n - 1)
  Eitwon <- (n^2 * s1 - n * s2 + 3 * sumw^2)/(sumw^2 * (n^2 - 1))
  sdn <- sqrt(Eitwon - Ei^2)
  b2 <- n * (sum(z^4)/(sum(z^2)^2))
  Eitwor <- n * ((n^2 - 3 * n + 3) * s1 - n * s2 + 3 * sumw^2)
  Eitwor <- Eitwor - b2 * ((n^2 - n) * s1 - 2 * n * s2 + 6 * sumw^2)

```

```

Eitwor <- Eitwor/((n - 1) * (n - 2) * (n - 3) * sumw^2)
sdr <- sqrt(Eitwor - Ei^2)
cat("\n UNDER NORMAL APPROXIMATION \n")
cat("\n Moran's I is      = ", round(m, 6))
cat("\n Mean of I is      =", round(Ei, 6))
cat("\n St. Dev of I      = ", round(sdn, 6))
z1 <- (m - Ei)/sdn
cat("\n Z-Value           = ", round(z1, 6))
cat("\n P-Value(2-side) = ", round(2 * (1 - pnorm(abs(z1))), 6)
)
cat("\n\n UNDER RANDOMIZATION ASSUMPTION\n")
cat("\n Moran's I is      = ", round(m, 6))
cat("\n Mean of I is      =", round(-1/(n - 1), 6))
cat("\n St. Dev of I      = ", round(sdr, 6))
z2 <- (m - Ei)/sdr
cat("\n Z-Value           = ", round(z2, 6))
cat("\n P-Value(2-side) = ", round(2 * (1 - pnorm(abs(z2))), 6), "\n")
if(rescalepart)
{
  pm <- pm * n
}
if(k > 0)
{
  cat("\n (Computing Permutation Distribution)\n")
  msim <- rep(0, k)
  for(j in 1:k)
  {
    y <- sample(n)
    x <- x[y]
    z <- x - mx
    msim[j] <- t(z) %*% w %*% z/sumz2
  }
  prob1 <- length(msim[msim > m])/k
  prob2 <- 1 - prob1
  prob <- 2 * min(prob1, prob2)
  cat("\n Results based on ", k, "permutations")

```

```

cat("\n Moran's I          = ", round(m, 6))
cat("\n Mean of I   stat = ", round(mean(msim), 6))
cat("\n Std Deviation    = ", round(sqrt(var(msim)), 6))
cat("\n P-value(2-side) = ", round(prob, 6), "\n")
invisible(msim)
}

else invisible(list(m = m, partial = pm))

mix1<-morani(x1, w, k = 0, rescale = T, rescalepart = T)
mix11<-morani(x1, ww1, k = 0, rescale = T, rescalepart = T)
miy1<-morani(y1, w, k = 0, rescale = T, rescalepart = T)
miy11<-morani(y1, ww1, k = 0, rescale = T, rescalepart = T)
miz1<-morani(z1, w, k = 0, rescale = T, rescalepart = T)
miz11<-morani(z1, ww1, k = 0, rescale = T, rescalepart = T)

wgt<-xx[,1]
n<-sum(wgt)
nc<-1

for ( cn in 0:91)
{
  x11[cn+1]<- mean(x1[nc: (nc+wgt[cn+1]-1)])
  nc<-nc+wgt[cn+1]
}

nc<-1

for ( cn in 0: (91))
{
  y11[cn+1]<- mean(y1[nc: (nc+wgt[cn+1]-1)])
  nc<-nc+wgt[cn+1]
}

age<-x11
ftw<-y11
age1<-(sum(wgt*age))/n
ftw1<-(sum(wgt*ftw))/n
mwtdmeed<-c(age1, ftw1)

m<-92
vage<-(sum(wgt*((age-age1)^2)))/(m-1)
vftw<-(sum(wgt*((ftw-ftw1)^2)))/(m-1)
cageftw<-(sum(wgt*(age-age1)*(ftw-ftw1)))/(m-1)

```



```

l2cor11<-cageftw/((vage*vftw)^.5)
  m1<-mean(x1)
  m2<-mean(y1)
mindivid<-c(m1,m2)
  v1<-var(x1)
  v2<-var(y1)
vindivid<-c(v1,v2)
  cor1<-cor(x1,y1)
  af<-cor1*((v1*v2)^.5)
  vcs1<-c(v1,af)
  b1<-af/v1
  nbar<-n/m
  nobar<-(sum(wgt^2))/n
nastbar<-nbar+((nbar-nobar)/(m-1))
lev2age<-(vage-vvx)/(nastbar-1)
lev2ftw<-(vftw-vvy)/(nastbar-1)
lev1age<-vvx-lev2age
lev1ftw<-vvy-lev2ftw
  iacage<-lev2age/(lev2age+lev1age)
  iacftw<-lev2ftw/(lev2ftw+lev1ftw)
  iac<-c(iacage, iacftw )
lev2cov1<-(cageftw-af)/(nastbar-1)
lev1cov1<-af-lev2cov1
  iaccl11<-(lev1cov1)/((v1*v2)^.5)
  iaccl21<-(lev2cov1)/((v1*v2)^.5)
meindivid<-c(m1,m2)
vaindivid<-c(v1,v2)
mevaindividual<-cbind(meindividual,vaindividual)
  meed<-c(age1,ftw1)
  vaed<-c(vage,vftw)
  mevaed<-cbind(meed,vaed)
lev1varcom<-c(lev1age,lev1ftw)
lev2varcom<-c(lev2age,lev2ftw)
intraareacor<-cbind(lev1varcom,lev2varcom)
  lev1rho11<-lev1cov1/((lev1age*lev1ftw)^.5)
  lev2rho11<-lev2cov1/((lev2age*lev2ftw)^.5)

```

```
bl12<-lev2cov1/lev2ftw
bl11<-lev1cov1/lev1ftw
sxxedlev[p]<-vvx
syyedlev[p]<-vvy
sxyedlev[p]<-af
redlev[p]<-af/vvx
cedlev[p]<-cor1
medlevx[p]<-mix1$m[1]
medlevy[p]<-miy1$m[1]
medlevz[p]<-miz1$m[1]
medlevx1[p]<-mix11$m[1]
medlevy1[p]<-miy11$m[1]
medlevz1[p]<-miz11$m[1]
sxxwardlev[p]<-var(x11)
syywardlev[p]<-var(y11)
sxywardlev[p]<-cov(y11,x11)
rwardlev[p]<-cageftw/vage
rUnWtdward[p]<-bl1
cwardlev[p]<-l2cor11
cUnWtdward[p]<-unwr
sx2wardlev[p]<-lev2age
sy2wardlev[p]<-lev2ftw
sxy2wardlev[p]<-lev2cov1
sx1wardlev[p]<-lev1age
sy1wardlev[p]<-lev1ftw
sxy1wardlev[p]<-lev1cov1
iacwardlevx[p]<-iacage
iacwardlevy[p]<-iacftw
iacc2wardlev[p]<-iacc121
iacc1wardlev[p]<-iacc111
pc2wardlev[p]<-lev2rho11
pc1wardlev[p]<-lev1rho11
pr2wardlev[p]<-bl12
pr1wardlev[p]<-bl11
vagewtd[p]<-vage
vftwtd[p]<-vftw
```

```

cageftwtd[p]<- cageftw
#Computation of the cross-Moran I
cmor[p]<-(((sxxedlev[p]+syyedlev[p]+2*sxyedlev[p])*medlevz[p])-
          (medlevy[p]*syyedlev[p])-(medlevx[p]*sxxedlev[p]))/
          (2*((sxxedlev[p]*syyedlev[p])^.5))

cmor1[p]<-(((sxxedlev[p]+syyedlev[p]+2*sxyedlev[p])*medlevz1[p])-
           (medlevy1[p]*syyedlev[p])-(medlevx1[p]*sxxedlev[p]))/
           (2*((sxxedlev[p]*syyedlev[p])^.5))

p<-p+1
}
data1<-cbind(medlevx,medlevx1,medlevy,medlevy1,medlevz,medlevz1,
             sxxedlev,syyedlev,sxyedlev,redlev,cedlev,sxxwardlev,syywardlev,
             sxywardlev,rwardlev,rUnWtdward,cwardlev,cUnWtdward,sx2wardlev,
             sy2wardlev,sxy2wardlev,sx1wardlev,sy1wardlev,sxy1wardlev,iacwardlevx,
             iacwardlevy,iacc2wardlev,iacc1wardlev,pc2wardlev,pc1wardlev,pr2wardlev,
             pr1wardlev,cmor,cmor1,vagewtd,vftwtd,cageftwtd)
write.table(data1,file="data1.dbf", col.names = NA)

```

### A.3 Proximity Weights (WtEDLag01Queen.txt)

Note:

Only the first 3 and the last 3 wards were described to save space.

The rest of the file is stored as WtEDLag01Queen.txt

```

1904
1 8
513 498 61 9 2 10 19 18
2 7
63 61 6 11 10 3 1
3 7
64 63 52 6 53 7 2
.
.

```

```
.
1904 1606 1903 1897 1896 1608
1903 7
1605 1900 1606 1899 1898 1902 1897
1904 5
1608 1902 1896 1901 1862
```

## A.4 Weight (NumEdPerWard.txt)

Note: Only the first 3 and the last 3 wards were described to save space. The rest of the file is stored as NumEdPerWard.txt

```
28 40.471 40.141 40.141
38 40.166 40.05 40.05
25 41.05 40.315 40.315
.
.
.
21 38.357 39.507 39.507
14 39.701 39.91 39.91
16 39.609 39.883 39.883
25 41.05 40.315 40.315
```

## A.5 Proximity Weights used in the computation of Morans I (Block Proximity)

Note: Only the first 3 and the last 3 wards were described to save space. The rest of the file is stored as EDWARD333.txt

1904

1 27

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28

2 27

1 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28

3 27

1 2 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28

.

.

.

1902 24

1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894

1895 1896 1897 1898 1899 1900 1901 1903 1904

1903 24

1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894

1895 1896 1897 1898 1899 1900 1901 1902 1904

1904 24

1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894

1895 1896 1897 1898 1899 1900 1901 1902 1903

## Glossary of Terms

ED	Enumeration District
GIS	Geographical Information Systems
IAC	Intra-area Correlation Coefficient, also indicated by $\rho$
IACC	Intra-area Cross-Correlation Coefficient, also indicated by $\rho$
ICC	Intra-class Correlation Coefficient, also indicated by $\rho$
MAUP	Modifiable Areal Unit Problem
MLE	Maximum Likelihood Estimation
NA	Not Available, used to indicate missing values
OLS	Ordinary Least Squares
SARs	Samples of Anonymised Records)

# References

- [1] Achen, C. and Shively, W.: *Cross-level Inference* , Chicago, USA. (1996).
- [2] Amrhein, C. G.: Searching for the elusive aggregation effect: evidence from statistical simulation. *Environment and Planning A*, 27, (1995) p. 105-119.
- [3] Amrhein, C. G. and Flowerdew, R. The effect of data aggregation model to Canadian migration. In M. Goodchild and S. Gopal (Eds.), *Accuracy of Spatial Databases*. London, UK: Taylor and Francis Ltd. (1989) p. 229-238.
- [5] Amrhein, C.G. and Reynolds, H.: Using spatial statistic to assess aggregation effect. *Geographical Systems*, Vol 3, (1996) p. 143-158.
- [5] Amrhein, C.G. and Reynolds, H.: Using the Getis statistic to explore effects in metropolitan Toronto census data. *The Canadian Geographer*, 41(2), (1997) p. 137-149.
- [6] Anselin, L.: The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. Fischer, H.J. Scholten and D. Unwin (Eds.) *Spatial Analytical Perspective in GIS*, UK:Taylor and Francis Ltd.(1996) p. 111-125.
- [7] Arbia, G.: *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht, The Netherlands: Kluwer Academic Publishers. (1989).
- [8] Blacke, M and Openshaw, S: Some new classifications of census variables for use in classification research. Working paper, School of Geography, Leeds University (1989).
- [9] Balock, H.: *Causal Inferences in Nonexperimental Research*. Chapel Hill, USA: The University of North Carolina Press. (1964).
- [10] Cliff, A. and K. Ord.:*Spatial Processes, Models and Applications*: London: Pion, 1981.
- [11] Clark, W. A. V. and Avery, K. L.: The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8. (1976) p. 428-438.
- [12] Flowerdew, R. and Green, M.: Areal interpolation and types of data. *Spatial Analysis and GIS*. London: Taylor and Francis. (1994) p. 121-145.

## REFERENCES

## REFERENCES

- [13] Fotheringham, S.: Scale-independent spatial analysis. In Goodchild, M. and Gopal, S. (Eds.), *The Accuracy of Spatial Databases*, London: Taylor and Francis.(1989) p. 221-228.
- [14] Fotheringham, S. and Wong, D. W. S.: The modifiable areal unit problem in multivariate statistical analysis. *Environmental Planning A*, 23, (1991) p. 1025-1044.
- [15] Gehlke, C. E. and Biehl, K.: Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association, Supplement(29)*. (1934) p. 169-170.
- [16] Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M.: *A User's Guide to MLwiN for Windows (MLwiN)* . London, Institute of Education (1998).
- [17] Goodman, L.A.: Some alternatives to ecological correlation. *The American Journal of Sociology*, 64, (1959), p. 610-625.
- [18] Gotway, C. A. and Young, L.J : Combining incompatible spatial data *Journal of the American Statistical Association*, 97, 2002, p. 632-648.
- [19] Green, M. and Flowerdew, R.: New evidence on the modifiable areal unit problem. *Spatial Analysis: Modelling in a GIS Environment*. Cambridge, UK: GeoInformation International. (1996) p. 41-54.
- [20] Holt, D., Steel, D., and Tranmer, M.: Area homogeneity and the modifiable areal unit problem. *Geographical Systems*, 3, (1996) p. 181-200.
- [21] Holt, D., Steel, D., and Tranmer, M.: Targets of inference and methods of analysis in multi-level populations. In *Paper presented at the RSS workshop on Multilevel Modelling*. London. (1997).
- [22] Jones, K. and Duncan, C. People and Places: The multilevel model as a general framework for the quantitative analysis of geographical data. In Longley, P and Batty, M (Eds.),*Spatial Analysis: Modelling in a GIS Environment*: Pearson Professional Ltd., UK,(1996), p. 79-104.
- [23] Moellering, H. and Tobler, W.: Geographical Variances. *Geographical Analysis*, 4, 1972, p. 34-50.
- [24] Openshaw, S.: A geographical solution to scale and aggregation problem in region building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers, New series*, 2. (1977a) p. 459-472.
- [25] Openshaw, S.: Optimal zoning systems for spatial interaction models. *Environmental Planning A*, 9. (1977b) p. 169-184.
- [26] Openshaw, S.: Ecological fallacies and the analysis of areal census data. *Environment and Planning A(16)*, (1984) p. 17-31.



- [27] Openshaw, S.: The modifiable areal unit problem. *Concepts and Techniques in Modern Geography, No. 38*. Norwich, UK: GeoBooks. (1984).
- [28] Openshaw, S. and Taylor, P. J.: A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical Application in the Spatial Sciences*. London, UK: Pion Ltd. (1979) p. 127-144.
- [29] Reynolds, H: The Modifiable Area Unit Problem: Emperical Analysis by Statistical Simulation *PhD Thesis*. Graduate Department of Geography, University of Toronto, 1998 .
- [30] Reynolds, H. and Amrhhein, C.: Using a Spatial Data Set Generator in an Emperical Analysis of Aggregation effects on Univariate Statistics. *Geographical & Environmental Moedelling, Vol 1, No. 2*, 1997, p. 199-219.
- [31] Robinson, A. H.: The necessity of weighting values in correlation analysis of areal data. *Annals, Association of American Geographers, 46*. (1956) p. 233-236.
- [32] Robinson, W.: Ecological correlations and the behaviour of individuals. *American Sociological Review (15)*. (1950) p. 351-357.
- [33] Sawada, M. :ROOKCASE: An Excel 97/Visual Basic (VB) Add-in for exploring global and local spatial autocorrelation. *Bulletin of the Ecological Society of America 80*. (1999) p. 231-234.
- [34] Snijders, T. and Bosker, R.: *Multilevel Analysis: an introduction to basic and advanced multilevel modeling*. London: Sage Publications. (1999).
- [35] Steel, D., Holt, D., and Tranmer, M. (1994). Modelling and adjusting aggregation effects. *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, USA. (1994) p. 382-408.
- [36] Steel, D., Holt, D., and Tranmer, M.: Making unit level inference from aggregate data. *Survey Methodology, 22(1)*. (1996) p. 3-15.
- [37] Steel, D., Holt, D., and Tranmer, M.: Analysis Combining Survey and Geographically Aggregated Data. *Analysis of survey data*. Edited by Chambers, R. L. and Skinner C. J., John Wiley & Sons, Ltd.,(2003) p. 323-343.
- [38] Steel, D. and Holt, D.: Rules for Random aggregation. *Environmental Planning A*. (1996a) p. 957-978.
- [39] Steel, D. and Holt, D.: Analysing and adjusting aggregation effects: the ecological fallacy revisited. *The International Statistical Review, 64(1)*. (1996b) p. 39-60.
- [40] Taylor, P. J.: *Quantitative methods in geography*. Houghton Mifflin Company, Boston, USA.(1977)

## REFERENCES

## REFERENCES

- [41] Tobler, W. R.: A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2). (1970) p. 234-240.
- [42] Tobler, W. R.: Frame independent spatial analysis. In M. Goodchild & S. Gopal (EDs.), *Accuracy of Spatial Databases*. New York, USA: Taylor and Francis. (1989) p. 115-122.
- [43] Tranmer, M. and Steel, D.: Using census data to investigate the causes of ecological fallacy. *Environment and Planning A*, 30. (1998) p. 817-831.
- [44] Tranmer, M & Steel, D.: Using Local Census Data to Investigate Scale Effects. *Modelling Scale in Geographical Information Science*. Tate, N & Atkinson, P., John Wiley & Sons, Ltd., (2001) p. 105-122.
- [45] Wong, D.: Aggregation effects in geo-referenced data. In S. L. Arlinghaus (Ed), *Practical Handbook of Spatial Statistics*, Florida, USA: CRC Press Inc. (1996) p. 83-106.
- [46] Wrigley, N., Holt, D., Steel, D., and Tranmer, M.: Analysing, modelling, and resolving the ecological fallacy. In P. Longley & M. Batty (Eds.), *Spatial Analysis: Modelling in a GIS Environment*. Cambridge, UK: GeoInformation International. (1996) p. 23-40.
- [47] Yule, G. and Kendall, M.: *An introduction to the Theory of Statistics*. London, UK: Charles Griffin & Company Ltd. (1950).