APPLICATIONS OF SPATIAL AUTOCORRELATION

by

CHALANI C. PREMATILAKE, B.Sc., M.S.

A Thesis

in

Statistics

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for the Degree of

Master of Science

Approved

Dr. Petros Hadjicostas
Chair of the committee

Dr. Frits Ruymgaart

Dr. Magdalena D. Toda

Dr. Peggy Gordon Miller
Interim Dean of the Graduate School

August, 2011

©2011, Chalani C. Prematilake

ACKNOWLEDGMENTS

First of all I would like to thank the members of my thesis committee: Dr. Petros Hadjicostas (chair), Dr. Frits Ruymgaart for being nice and for being willing to help whenever I needed something, and Dr. Magdalena D. Toda for consenting to read my thesis and to be a member of my committee.

I owe my deepest gratitude to Dr. Petros Hadjicostas for directing and motivating me throughout the eighteen months I have been at Texas Tech University. In particular, this thesis would not have been possible without his tremendous support. I'm also grateful for his guidance. I wouldn't be able to handle my studies, teaching and research simultaneously without his guidance.

It is a pleasure to thank those who have supported me throughout my academic life to accomplish my goals. I would like to show my gratitude to Dr. Kanthi Perera, who had been my advisor since I was studying for my first Masters Degree. And I would specially thank Dr. Pantaleon Perera for his kind direction and for encouraging me to join Texas Tech University.

I am indebted to all my colleagues for their support. I would also like to express special thanks for Janaki and Chandani for being here with me to share all the happy and sad moments.

I am very thankful to my parents for always being behind me and trying to make me up with a lot of love. I would like to thank my husband Suranga, for his love, caring and encouragement. All my successes are dedicated to him for his commitment. And also I would like to thank my beloved daughter, Tanuli, and my little son, Sandul, for their endless love and for allowing me to stay away from them for hours.

CONTENTS

ACKNOWLEDGMENTS ii
ABSTRACT iv
LIST OF TABLES vi
LIST OF FIGURES vii
I. INTRODUCTION 1
II. SPATIAL AUTOCORRELATION FOR NOMINAL VARIABLES . . . 5
2.1 Introduction 5
2.2 Notation 7
2.3 Moments for the number of joins with the same type 8
2.4 Moments for the number of joins with different types 13
2.5 Covariances among the three different join counts 17
2.6 Some more examples 18
2.7 Connection of join counts to Moran's I autocorrelation coefficient 23
2.8 Connection of join counts to Geary's ratio c 29
2.9 Calculation of join counts and their moments in R 31
2.10 Calculation of spatial autocorrelation statistics in SAS 35
2.11 Implementation of the formulas of the chapter in R 41
III. SPATIAL AUTOCORRELATION FOR NUMERICAL VARIABLES . 44
3.1 Introduction 44
3.2 Moments of each X_i under randomization 47
3.3 Discussion of Moran's I autocorrelation coefficient 49
3.4 Moments of Moran's I 50
3.5 Discussion of Geary's Ratio c 53
3.6 Moments of Geary's ratio c 54
3.7 Calculation of weights when the exact locations of the units are known 56
3.8 Thiessen polygons and Voronoi diagrams 58
IV. CONCLUSIONS AND FUTURE RESEARCH 62
BIBLIOGRAPHY 64

ABSTRACT

As we use time series analysis to study data with respect to their time of occurrence, we can also use spatial statistics to study data with respect to their locations of occurrence in space. Even though the history of spatial probabilistic analysis goes way back to the 18th century (Buffon's needle problem), a serious attempt to study spatial statistics was first made at the beginning of the 20th century by Student.

This study examines measures of overall spatial autocorrelation or association. The word "autocorrelation" means the correlation of a variable with itself (over time or over space, or both). As Griffith said in 1987, the quality and quantity of information contained in spatial data is reflected on spatial autocorrelation. For example, in the case of a numerical variable of interest, if most pairs of neighbouring localities have values of the variable of interest both above the average or both below the average, then spatial autocorrelation tends to be "large" in some way (above a certain number), while if, on the other hand, for most pairs of neighbouring localities, one locality has a value of the variable above the average and the other one has a value below the average, then the autocorrelation measure tends to be "small" (below a certain number).

The study of spatial statistics takes different forms according to the kind of data used. For example, when the data are nominal categorical, we can use join counts as measures of spatial association. For example, we can find the number of neighbouring localities that are of the same "type" (category of the nominal variable) and the number of neighbouring localities of different "types." Moran is one of the first authors who studied join count statistics. He also calculated the moments of join counts in 1948. Similar studies had been carried out by P. V. Krishna Iyer in 1949 and 1950 and by Florence Nightingale David in 1971. They both came out with similar results but in different experimental environments. In the second chapter of the thesis, we review some of the work of Moran on join counts and their moments.

Spatial autocorrelation of numerical data is usually carried out using Moran's I coefficient and Geary's ratio c (introduced in 1950 and 1954, respectively). In the third chapter of this thesis, we review some of the probabilistic properties of these

spatial autocorrelation coefficients that show how a variable is correlated with itself over space.

We use the statistical packages R and SAS to calculate and apply the above statistics to some examples with spatial data. In addition, we show the connection of the join count statistics with Moran's I coefficient and Geary's ratio c, which is probably one of the new contributions in this thesis.

Throughout most of the thesis, we show (using modern notation) the randomization properties of some of the above spatial statistics, that is, we review Moran's and Geary's calculations on the probabilistic behaviour of these statistics by conditioning on the observed data (values of the variable of interest in different localities), but not in the order they appear. In other words, we show how statisticians in spatial statistics derive the distribution of some autocorrelation statistics under the non-free sampling scenario.

LIST OF TABLES

2.1 Adjacency matrix M_W for Example 2 (from [16, p. 203]) 20
2.2 Calculation of some numbers of joins for Example 2 (from [16, p. 204]) 21
2.3 Adjacency matrix M_W for Example 4 (from [11, p. 26]) 33
3.1 The sample space Ω of permutations of the five observed average rainfalls for Example 5 (with $n_1 = 1$, $n_2 = 2$ and $n_3 = 2$) 46
3.2 (a_i, b_i) coordinates and x_i values for 5 rainfall stations 56

LIST OF FIGURES

2.1 Checkerboard and graph for Example 2 (from [16, p. 203]) 19
2.2 Checkerboard and graph for Example 3 27
2.3 "Checkerboard" and connectivity graph for the presence or absence of prostitution among 14 police precincts (B = r = yes, W = s = no) . 32
3.1 Thiessen polygons for the data in Table 3.2. 60

CHAPTER I
INTRODUCTION

As we use time series analysis to study data with respect to their time of occurrence, we can also use spatial statistics to study data with respect to their locations of occurrence in space. Even though the history of spatial probabilistic analysis goes way back to the 18th century (Buffon's needle problem; see, for example, Solomon (1978), pp. 2-3), a serious attempt to study spatial statistics was first made at the beginning of the 20th century (e.g., see Student (1907)).

This study examines measures of overall spatial autocorrelation or association. The word "autocorrelation" means the correlation of a variable with itself (over time or over space, or both). According to Griffith (1987), the quality and quantity of information contained in spatial data is reflected on spatial autocorrelation.

The study of spatial statistics takes different forms according to the kind of data used. For example, when the data are nominal categorical, we can use join counts as measures of spatial association. Moran is the first author who studied join count statistics. He also calculated the moments of join counts in 1948. Similar studies had been carried out by Krishna Iyer in 1949 and 1950 and by David in 1971. They both came out with similar results but in different experimental environments.

In 1907 Student (that is, William Sealy Gosset) studied the distribution of particles in a liquid. He divided the space (liquid) into parcels and considered the number of particles per parcel as the random variable. He encountered that the distribution of the number of particles per parcel is approximately Poisson. Fisher (1935, Chapter IV) worked on the spatial impact on statistical experiments in agricultural fields[1]. He introduced several principles such as randomization,

[1]Fisher worked for several years at the Rothamsted Experimental Station at Harpenden in Hertfordshire, England.

blocking and replication to minimize spatial autocorrelation among data.

Moran (1948) introduced join count statistics (for the categories[2] of a nominal categorical variable) and their moments, which we discuss in detail in Chapter II. Krishna Iyer (1949 and 1950) discussed a problem of assigning randomly k different colors[3] to different counties (localities) and calculated the first two moments of three different join counts under two different sampling scenarios. First he counted the number of joins which connect counties with the same color, then he counted the number of joins which connect counties with two different colours, and finally he counted the number of joins which connect counties with any two different colors. The last one is usually denoted by J_{tot}, but it is not studied in this thesis. David (1971ab) who worked on the spatial diversity of plant species came out with a similar join count as J_{tot}. She observed that low values of J_{tot} are related to low spatial diversity and high values of J_{tot} are related to high spatial diversity.

For nominal categorical data, there are two sampling scenarios: free sampling and nonfree sampling (see Krishna Iyer (1949) and Chapter 3 in Griffith (1987)). In free sampling, the assignments of categories of the nominal variable to the various locations are independent from one other. In addition, under free sampling, it is assumed that each of the categories of the nominal variable has the same probability of being assigned to a particular location. In nonfree sampling, which is also called "randomization sampling," we can condition on the observed categories of the nominal variable (but *not* in the order they appear among the locations), and categories are assigned to locations so that any permutation (with repetition) of the observed data has the same probability of appearing in the various locations as any other permutation of the observed data. In Chapter II, we use the nonfree sampling

[2]In this thesis, the categories are called *types*.

[3]One can think of a "color" as a category (type) of a nominal categorical variable.

technique, and (following Moran (1948) and others) we find the first two moments of join counts statistics under this method of sampling.

For numerical data, there are three sampling scenarios: free sampling, nonfree sampling and sampling under normality (see Chapter 3 in Griffith (1987)). Free sampling and nonfree sampling are almost the same as described in the previous paragraph[4]. Under the normality sampling, instead of considering all possible permutations of the data, we assume that the observations in the sample are distributed normally and independently of one another. For large samples (because of the Central Limit Theorem and its generalizations) the sampling distributions of various autocorrelation statistics are appoximately normal[5] for all three sampling scenarios (under certain regularity assumptions).

With respect to numerical data occuring in space, Moran (1950) suggested a spatial autocorrelation coefficient, known today as Moran's I. Geary in 1954 suggested another measure of spatial autocorrelation, known as Geary's ratio c, in 1954. We discuss these two measures in detail in Chapter III. For more details about the history of spatial statistics, see Chapter 1 in Cliff and Ord (1981), Chapter 1 in Cressie (1991), and Getis (2007).

In this thesis, we study measures of overall spatial association for nominal and numerical data separately. In Chapter II we discuss measures of spatial association for nominal data from a single categorical variable. Even though the categorical variable may have more than two categories, for every two different categories, named as r and s, we follow Moran (1948) and Cliff and Ord (1981) and study three

[4]For the nonfree sampling, which is used in Chapter III of this thesis, we condition on the ordered statistics of the observed data, and as in the case of nominal data, the sample space of the experiment consists of all possible permutations (with repetitions) of these observed data, and all of these permutations are assumed to be equally likely.

[5]See Sections 2.4.1 and 2.4.2 in Cliff and Ord (1981) and Section 11 in Moran (1948).

different join counts: for rr joins, ss joins and rs joins. We show the calculations of the expected values and variances for each join count statistic and find the covariances among those three statistics under randomization (non-free sampling). We discuss the connection of join counts to the two autocorrelation coefficients, Moran's I and Geary's Ratio; this is probably (to the best of our knowledge) one of the new contributions of the thesis. Finally, we use a number examples to illustrate the above derived formulas. In addition, we use the statistical packages R and SAS to implement the formulas and to verify the results.

As mentioned above, in Chapter III we discuss spatial autocorrelation for numerical variables. We discuss Moran's I and Geary's ratio c, and we prove the (well-known) formulas for the expected values of these autocorrelation statistics under the assumption of randomization[6].

[6]In Section 2.10, we mention that in the SAS Variogram procedure the default assumption is that the data are normally and independently distributed, and we provide the formula used by SAS for calculating the variance of the sampling distribution of Geary's ratio c.

CHAPTER II
SPATIAL AUTOCORRELATION FOR NOMINAL VARIABLES

In this chapter we discuss measures of autocorrelation for nominal variables (categorical variables whose categories are not ordered), and in particular, we discuss probabilistic properties of numbers of join counts. We derive some of their randomization moments and discuss how to calculate them and get information about their sampling distributions in practice using the statistical packages R and SAS. In addition, for the case of binary nominal variables (whose categories are coded with the numbers 0 and 1), we discuss the connection between these measures to the famous measures of autocorrelation for numerical data due to Moran (1950) and Geary (1954).

2.1 Introduction

We assume we have n distinct regions (e.g., squares or localities) and we have n objects, n_1 of which are of type 1, n_2 of which are of type 2 and so on, and n_k of which are of type k. Of course

$$n_1 + n_2 + \ldots + n_k = n,$$

and we are interested in distributing the n objects to the n regions so that each region gets exactly one object.

The sample space of interest, denoted by Ω, consists of all possible ways of assigning the n objects to the k regions. In other words, Ω is the set of all permutations of the above n objects such that n_1 of them are identical and of type 1, n_2 of them are identical and of type 2, and so on. We know from combinatorics that the number of elements in Ω is

$$\#\Omega = \frac{n!}{n_1! n_2! \ldots n_k!} = \begin{pmatrix} n \\ n_1, \ldots, n_k \end{pmatrix}.$$

A probability model on the sample space Ω is any function $\mathbb{P} : \Omega \to [0, 1]$ such that

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1.$$

Many times, under null hypothesis, we assume that the n balls are distributed randomly (without replacement) over the n localities or regions. This means that:

$$H_0 : \mathbb{P}(\omega) = \frac{1}{\#\Omega} = \frac{n_1! n_2! n_3! \ldots n_k!}{n!} \quad \text{for all} \quad \omega \in \Omega. \tag{2.1}$$

We know from probability theory that a random variable X is a function from the sample space Ω to the real line. Thus: $\{X = a\} \stackrel{\text{def}}{=} \{\omega \in \Omega |\ X(\omega) = a\}$ for any $a \in \mathbb{R}$.

Example 1. Suppose we have $n = 4$ localities and we have two black balls (B) and two white balls (W), and we want to distribute the four balls to the four localities so that each locality gets exactly one ball. Then

$$\Omega = \{\text{BBWW}, \text{BWBW}, \text{WBWB}, \text{WWBB}, \text{WBBW}, \text{BWWB}\},$$

where it is understood that for allocation BBWW, black balls go to regions 1 and 2, and white balls to regions 3 and 4. Here, $k = 2$ and $\#\Omega = \frac{4!}{2!2!} = 6$. A possible probability model on Ω might be

$$\mathbb{P}(\text{BBWW}) = \mathbb{P}(\text{BWBW}) = \mathbb{P}(\text{WBWB}) = \frac{1}{12};$$
$$\mathbb{P}(\text{WWBB}) = \mathbb{P}(\text{WBBW}) = 0; \quad \text{and} \quad \mathbb{P}(\text{BWWB}) = \frac{3}{4}.$$

We may consider the random variable $(BW)_{ij} : \Omega \to \{0, 1\}$ which equals 1 when locality i is B and j is W or when locality j is B and i is W, and zero otherwise. Then

$$\mathbb{P}[(BW)_{ij} = 1] = \mathbb{P}[\{\omega \in \Omega |\ (BW)_{ij}(\omega) = 1\}].$$

We may also consider other random variables, such as $(BB)_{ij} : \Omega \to \{0, 1\}$, which equals 1 when both localities i and j are B. □

2.2 Notation

Suppose that one of the k types of objects is r and another one is s. For each pair of localities i and j define the random variables $(rr)_{ij} : \Omega \to \{0, 1\}$, $(ss)_{ij} : \Omega \to \{0, 1\}$, and $(rs)_{ij} : \Omega \to \{0, 1\}$ as follows:

$$(rr)_{ij} = \begin{cases} 1 & \text{if localities } i \text{ and } j \text{ are both of type } r, \\ 0 & \text{otherwise;} \end{cases}$$

$$(ss)_{ij} = \begin{cases} 1 & \text{if localities } i \text{ and } j \text{ are both of type } s, \\ 0 & \text{otherwise;} \end{cases}$$

$$(rs)_{ij} = \begin{cases} 1 & \text{if locality } i \text{ is of type } r \text{ and } j \text{ is of type } s, \text{ or vice versa,} \\ 0 & \text{otherwise.} \end{cases}$$

Let w_{ij} be the weight given to join[1] or edge between localities i and j (see Sokal and Oden (1978, p. 226)). It is not necessary for w_{ij} to equal w_{ji}.

We always assume that $w_{ii} = 0$ for all $i \in [1..n]$ because there is *no join* of a region with itself. (Here $[1..n] := \{m \in \mathbb{N} |\ 1 \leq m \leq n\}$.) We denote the $n \times n$ matrix of weights by M_W, i.e.,

$$M_W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{pmatrix}$$

and the sum of all weights by W, i.e.,

$$W = \sum_{i \neq j} w_{ij}, \tag{2.2}$$

where the summation is over all pairs $(i, j) \in [1..n]^2$ with $i \neq j$. As stated above, we

[1]This is a specialized term. The "joins" are defined by the $n \times n$ weight matrix M_W. This is a form of *connections matrix* in which each entry w_{ij} describes whether localities i and j are *joined* or not.

will ignore the diagonal elements of M_W.

Following Sokal and Oden (1978), for localities of the same type, say of type r, we define the **number of rr joins** by

$$N_{rr} := \frac{1}{2}\sum_{i\neq j} w_{ij}(rr)_{ij}. \tag{2.3}$$

For localities of different types, say r and s, we define the **number of rs joins** by

$$N_{rs} := \frac{1}{2}\sum_{i\neq j} w_{ij}(rs)_{ij}. \tag{2.4}$$

This formula appears in p. 225 in Sokal and Oden (1978). Moran (1948) derived various moments of the two join counts, N_{rr} and N_{rs}. In the next three sections we show the calculation of the first two moments (joint or marginal) of these join counts.

2.3 Moments for the number of joins with the same type

We wish to find the expected values and the variances of the random variables $N_{rr} : \Omega \to \mathbb{R}$ and $N_{rs} : \Omega \to \mathbb{R}$ under the null hypothesis (2.1), i.e., under the assumption of sampling without replacement. In this section we find $\mathbb{E}[N_{rr}]$ and $\mathrm{Var}[N_{rr}]$.

We claim that, under H_0,

$$\mathbb{E}[N_{rr}] = \mathbb{E}\left[\frac{1}{2}\sum_{i\neq j} w_{ij}(rr)_{ij}\right] = \frac{W n_r^{(2)}}{2n^{(2)}}, \tag{2.5}$$

where for any nonnegative integer ν and any number m we define

$$m^{(\nu)} := \prod_{l=0}^{\nu-1}(m-l) = m(m-1)(m-2)\dots(m-\nu+1),$$

where an empty product is by definition 1, and thus $m^{(0)} = 1$. For example, $m^{(1)} = m$ and $m^{(2)} = m(m-1)$. Here n_r is the number of objects out of n that are of type r. Formula (2.5) is mentioned, for example, in Sokal and Oden (1978), p. 225.

To prove formula (2.5), note that

$$\begin{aligned}\mathbb{E}[(rr)_{ij}] &= \mathbb{P}[(rr)_{ij}=1] \\ &= \mathbb{P}[\text{object in locality } i \text{ is of type } r] \\ &\quad \times \mathbb{P}[\text{object in locality } j \text{ is of type } r \mid \text{object in locality } i \text{ is of type } r] \\ &= \frac{n_r}{n}\cdot\frac{n_r-1}{n-1}=\frac{n_r^{(2)}}{n^{(2)}}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[N_{rr}] &= \mathbb{E}\left[\frac{1}{2}\sum_{i\neq j} w_{ij}(rr)_{ij}\right] \\ &= \frac{1}{2}\sum_{i\neq j} w_{ij}E\left[(rr)_{ij}\right] \\ &= \frac{1}{2}\sum_{i\neq j} w_{ij}\frac{n_r^{(2)}}{n^{(2)}} \\ &= \frac{Wn_r^{(2)}}{2n^{(2)}}.\end{aligned}$$

To find the variance of the random variable N_{rr} (the number of rr joins), we define

$$L := \{(i,j,k,l)\in[1..n]^4 |\ i\neq j \text{ and } k\neq l\}, \tag{2.6}$$

where $[1..n] := \{m\in N|\ 1\leq m\leq n\}$. Define also

$$\begin{aligned}L_1 : &= \{(i,j,k,l)\in L|\ k\neq i\neq l \text{ and } k\neq j\neq l\}; &(2.7)\\ L_{2a} : &= \{(i,j,k,l)\in L|\ k=i\neq l \text{ and } k\neq j\neq l\}; \\ L_{2b} : &= \{(i,j,k,l)\in L|\ k\neq i=l \text{ and } k\neq j\neq l\}; \\ L_{2c} : &= \{(i,j,k,l)\in L|\ k\neq i\neq l \text{ and } k=j\neq l\}; \\ L_{2d} : &= \{(i,j,k,l)\in L|\ k\neq i\neq l \text{ and } k\neq j=l\}; \\ L_2 : &= L_{2a}\cup L_{2b}\cup L_{2c}\cup L_{2d} &(2.8)\\ &= \{(i,j,k,l)\in L|\ \text{exactly two of the integers } i,j,k,l \text{ are equal}\};\end{aligned}$$

$$\begin{aligned} L_{3a}: &= \{(i,j,k,l) \in L |\ i = k \neq j = l\}; \\ L_{3b}: &= \{(i,j,k,l) \in L |\ i = l \neq k = j\}; \\ L_3 &:= L_{3a} \cup L_{3b}. \end{aligned} \tag{2.9}$$

Then:

$$\mathbb{E}[(rr)_{ij}(rr)_{kl}] = \left\{ \begin{array}{ll} \mathbb{P}[(rr)_{ij} = 1 = (rr)_{kl}] & \text{if } (i,j,k,l) \in L_1 \\ \mathbb{P}[(rr)_{ij} = 1 = (rr)_{il}] & \text{if } (i,j,k,l) \in L_2 \\ \mathbb{P}[(rr)_{ij} = 1] & \text{if } (i,j,k,l) \in L_3 \end{array} \right\}$$

$$= \left\{ \begin{array}{ll} \frac{n_r^{(4)}}{n^{(4)}} & \text{if } (i,j,k,l) \in L_1 \\ \frac{n_r^{(3)}}{n^{(3)}} & \text{if } (i,j,k,l) \in L_2 \\ \frac{n_r^{(2)}}{n^{(2)}} & \text{if } (i,j,k,l) \in L_3 \end{array} \right\} = \frac{n_r^{(5-t)}}{n^{(5-t)}} \text{for } (i,j,k,l) \in L_t, \ \ t = 1,2,3.$$

To understand the last step, consider for example the calculation of

$P[(rr)_{ij} = 1 = (rr)_{il}]$ with $(i,j,k,l) \in L_2$. We have:

$$\begin{aligned} \mathbb{E}\left[(rr)_{ij}(rr)_{kl}\right] &= \mathbb{P}[(rr)_{ij} = 1 = (rr)_{il}] \quad \text{if } (i,j,k,l) \in L_2 \\ &= \mathbb{P}[\text{object in locality } i \text{ is of type } r] \\ &\quad \times \mathbb{P}[\text{object in loc. } j \text{ is of type } r \mid \text{object in loc. } i \text{ is of type } r] \\ &\quad \times \mathbb{P}[\text{object in loc. } l \text{ is of type } r | \text{obj. in loc. } i \text{ and } j \text{ are of type } r] \\ &= \frac{n_r}{n} \cdot \frac{(n_r - 1)}{(n-1)} \cdot \frac{(n_r - 2)}{(n-2)} \\ &= \frac{n_r^{(3)}}{n^{(3)}} \quad \text{for } (i,j,k,l) \in L_2. \end{aligned}$$

Since

$$\text{Var}(N_{rr}) = \mathbb{E}[N_{rr}^2] - (\mathbb{E}[N_{rr}])^2,$$

and we know $\mathbb{E}(N_{rr})$ through formula (2.5), we only need to find $\mathbb{E}[N_{rr}^2]$. We first

define the following quantities:

$$S_1 := \frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})^2 \quad \text{and} \quad S_2 := \sum_{i=1}^{n}(w_{i.}+w_{.i})^2, \tag{2.10}$$

where $w_{i.} := \sum_{j=1}^{n} w_{ij}$ and $w_{.i} := \sum_{j=1}^{n} w_{ji}$, with the understanding that $w_{ii} := 0$.

Since $(rr)_{ij} = (rr)_{ji}$, we have

$$\begin{aligned}
\mathbb{E}[N_{rr}]^2 &= \frac{1}{4}\mathbb{E}\left[\left(\sum_{i\neq j} w_{ij}(rr)_{ij}\right)^2\right] \\
&= \frac{1}{4}\mathbb{E}\left[\left(\frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})(rr)_{ij}\right)^2\right] \\
&= \frac{1}{16}\mathbb{E}\left[\sum_{(i,j,k,l)\in L}(w_{ij}+w_{ji})(w_{kl}+w_{lk})(rr)_{ij}(rr)_{kl}\right] \\
&= \frac{1}{16}\left(E_1+E_2+E_3\right),
\end{aligned}$$

where

$$\begin{aligned}
E_t &:= \sum_{(i,j,k,l)\in L_t}(w_{ij}+w_{ji})(w_{kl}+w_{lk})\mathbb{E}\left[(rr)_{ij}(rr)_{kl}\right] \\
&= \frac{n_r^{(5-t)}}{n^{(5-t)}}e_t
\end{aligned}$$

with

$$e_t := \sum_{(i,j,k,l)\in L_t}(w_{ij}+w_{ji})(w_{kl}+w_{lk}) \tag{2.11}$$

for $t = 1, 2, 3$. In the rest of the section we make intensive use of equations (2.6)-(2.11).

Because of symmetry,

$$\begin{aligned}
e_3 &= \sum_{(i,j,k,l)\in L_3}(w_{ij}+w_{ji})(w_{kl}+w_{lk}) = 2\sum_{(i,j,k,l)\in L_{3a}}(w_{ij}+w_{ji})(w_{kl}+w_{lk}) \\
&= 2\sum_{i\neq j}(w_{ij}+w_{ji})^2 = 4\left[\frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})^2\right] = 4S_1.
\end{aligned} \tag{2.12}$$

Also

$$\begin{aligned} e_2 &= \sum_{(i,j,k,l)\in L_2} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) \\ &= 4 \sum_{(i,j,k,l)\in L_{2a}} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) \\ &= 4 \sum_{(i,j,k,l)\in L_{2a}\cup L_{3a}} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) - 4 \sum_{(i,j,k,l)\in L_{3a}} (w_{ij}+w_{ji})(w_{il}+w_{li}) \\ &= 4\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(w_{ij}+w_{ji})\sum_{\substack{l=1 \\ l\neq i}}^{n}(w_{il}+w_{li}) - 4(2S_1) \\ &= 4\sum_{i=1}^{n}(w_{i.}+w_{.i})^2 - 8S_1 \\ &= 4S_2 - 8S_1, \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} e_1 &= \sum_{(i,j,k,l)\in L_1} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) \\ &= \sum_{(i,j,k,l)\in L} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) - \sum_{(i,j,k,l)\in L_2} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) \\ &\quad - \sum_{(i,j,k,l)\in L_3} (w_{ij}+w_{ji})(w_{kl}+w_{lk}). \end{aligned}$$

Thus

$$\begin{aligned} e_1 &= \sum_{(i\neq j,\ k\neq l)} (w_{ij}+w_{ji})(w_{kl}+w_{lk}) - e_2 - e_3 \\ &= \sum_{i\neq j}(w_{ij}+w_{ji})\sum_{k\neq l}(w_{kl}+w_{lk}) - (4S_2 - 8S_1) - (4S_1) \\ &= (2W)(2W) - 4S_2 + 4S_1 \\ &= 4(W^2 + S_1 - S_2). \end{aligned} \tag{2.14}$$

Finally,

$$\mathbb{E}[N_{rr}^2] = \frac{1}{16}\left(E_1 + E_2 + E_3\right),$$

$$
\begin{aligned}
&= \frac{1}{16}\left(\sum_{t=1}^{3}\frac{n_r^{(5-t)}}{n^{(5-t)}}e_t\right)\\
&= \frac{1}{16}\left[\left(\frac{n_r^{(4)}}{n^{(4)}}e_1\right)+\left(\frac{n_r^{(3)}}{n^{(3)}}e_2\right)+\left(\frac{n_r^{(2)}}{n^{(2)}}e_3\right)\right]\\
&= \frac{1}{16}\left[\left(\frac{n_r^{(4)}}{n^{(4)}}4(W^2+S_1-S_2)\right)+\left(\frac{n_r^{(3)}}{n^{(3)}}4(S_2-2S_1)\right)+\left(\frac{n_r^{(2)}}{n^{(2)}}4S_1\right)\right]\\
&= \frac{1}{4}\left[\frac{S_1 n_r^{(2)}}{n^{(2)}}+\frac{(S_2-2S_1)n_r^{(3)}}{n^{(3)}}+\frac{(W^2+S_1-S_2)n_r^{(4)}}{n^{(4)}}\right],
\end{aligned}
$$

and thus

$$
\begin{aligned}
\mathrm{Var}(N_{rr}) &= \mathbb{E}[N_{rr}^2]-(\mathbb{E}[N_{rr}])^2\\
&= \frac{1}{4}\left[\frac{S_1 n_r^{(2)}}{n^{(2)}}+\frac{(S_2-2S_1)n_r^{(3)}}{n^{(3)}}+\frac{(W^2+S_1-S_2)n_r^{(4)}}{n^{(4)}}\right]-\left(\frac{W n_r^{(2)}}{2n^{(2)}}\right)^2\\
&= \frac{1}{4}\left[\frac{S_1 n_r^{(2)}}{n^{(2)}}+\frac{(S_2-2S_1)n_r^{(3)}}{n^{(3)}}+\frac{(W^2+S_1-S_2)n_r^{(4)}}{n^{(4)}}\right.\\
&\quad \left.-W^2\left(\frac{n_r^{(2)}}{n^{(2)}}\right)^2\right]. \qquad (2.15)
\end{aligned}
$$

The above formula is found in p. 225 in Sokal and Oden (1978).

2.4 Moments for the number of joins with different types

In this section we calculate the expected value and the variance of the number of joins with different types of objects (r and s), i.e., we calculate $\mathbb{E}[N_{rs}]$ and $\mathrm{Var}[N_{rs}]$ (under the H_0 defined by (2.1)). We thus verify two formulas found in p. 225 in Sokal and Oden (1978). For joins between localities with different types (r and s) we have:

$$
\begin{aligned}
\mathbb{E}[(rs)_{ij}] &= \mathbb{P}[(rs)_{ij}=1]\\
&= \mathbb{P}[\text{object in locality } i \text{ is of type } r]\\
&\quad \times\mathbb{P}[\text{object in locality } j \text{ is of type } s \mid \text{object in locality } i \text{ is of type } r]
\end{aligned}
$$

$$
\begin{aligned}
& + P[\text{object in locality } i \text{ is of type } s] \\
& \times \mathbb{P}[\text{object in locality } j \text{ is of type } r \mid \text{object in locality } i \text{ is of type } s] \\
= \; & \frac{n_r}{n} \cdot \frac{n_s}{n-1} + \frac{n_s}{n} \cdot \frac{n_r}{n-1} \\
= \; & \frac{2n_r n_s}{n(n-1)} = \frac{2n_r n_s}{n^{(2)}},
\end{aligned}
$$

where n_s is the number of localities of type s.

Therefore the expected number of rs joins equals

$$
\begin{aligned}
\mathbb{E}[N_{rs}] &= \mathbb{E}\left[\frac{1}{2}\sum_{i\neq j} w_{ij}(rs)_{ij}\right] \\
&= \frac{1}{2}\sum_{i\neq j} w_{ij} E\left[(rs)_{ij}\right] \\
&= \frac{1}{2}\sum_{i\neq j} w_{ij}\frac{2n_r n_s}{n^{(2)}} \\
&= \frac{W n_r n_s}{n^{(2)}}, \qquad (2.16)
\end{aligned}
$$

where W, the sum of all weights, is given by (2.2).

The variance of the number of rs joins is given by

$$\mathrm{Var}(N_{rs}) = \mathbb{E}[N_{rs}^2] - \left(\mathbb{E}[N_{rs}]\right)^2.$$

To find this variance, the set L in (2.6) is split into the sets L_1, L_2 and L_3 as we have done for the case of localities with the same type. We thus have:

$$
\begin{aligned}
\mathbb{E}[N_{rs}^2] &= \frac{1}{4}\mathbb{E}\left[\left(\sum_{i\neq j} w_{ij}(rs)_{ij}\right)^2\right] \\
&= \frac{1}{4}\mathbb{E}\left[\left(\frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})(rs)_{ij}\right)^2\right] \\
&= \frac{1}{16}\mathbb{E}\left[\left(\sum_{(i,j,k,l)\in L}(w_{ij}+w_{ji})(w_{kl}+w_{lk})(rs)_{ij}(rs)_{kl}\right)\right] \\
&= \frac{1}{16}(\tilde{E}_1 + \tilde{E}_2 + \tilde{E}_3),
\end{aligned}
$$

where

$$\begin{aligned}\tilde{E}_t &= \mathbb{E}\left[\left(\sum_{(i,j,k,l)\in L_t}(w_{ij}+w_{ji})(w_{kl}+w_{lk})(rs)_{ij}(rs)_{kl}\right)\right] \\ &= \sum_{(i,j,k,l)\in L_t}(w_{ij}+w_{ji})(w_{kl}+w_{lk})\mathbb{E}\left[(rs)_{ij}(rs)_{kl}\right]. \qquad (2.17)\end{aligned}$$

For each of the cases (a) $(i,j,k,l)\in L_1$; (b) $(i,j,k,l)\in L_2$; and (c) $(i,j,k,l)\in L_3$, the quantity $\mathbb{E}[(rs)_{ij}(rs)_{kl}]$ can be derived as follows.

(a) For $(i,j,k,l)\in L_1$, there exist $2\cdot 2=4$ different ways of placing two different objects (r and s) in the distinct localities i and j and two different objects (again r and s) in the distinct localities k and l. We thus obtain

$$\begin{aligned}\mathbb{E}[(rs)_{ij}(rs)_{kl}] &= \mathbb{P}[(rs)_{ij}=1=(rs)_{kl}] \\ &= \frac{n_r}{n}\cdot\frac{n_s}{n-1}\cdot\frac{n_r-1}{n-2}\cdot\frac{n_s-1}{n-3}+\frac{n_r}{n}\cdot\frac{n_s}{n-1}\frac{n_s-1}{n-2}\cdot\frac{n_r-1}{n-3} \\ &\quad +\frac{n_s}{n}\cdot\frac{n_r}{n-1}\cdot\frac{n_r-1}{n-2}\cdot\frac{n_s-1}{n-3}+\frac{n_s}{n}\cdot\frac{n_r}{n-1}\cdot\frac{n_s-1}{n-2}\cdot\frac{n_r-1}{n-3} \\ &= \frac{4n_r^{(2)}n_s^{(2)}}{n^{(4)}}.\end{aligned}$$

(b) For $(i,j,k,l)\in L_{2a}$ we have $i=k$ and three distinct localities. In this case, either each of the localities $i=k$ and l gets an r object and locality j gets an s object, or each of the localities $i=k$ and l gets an s object and locality j gets an r object. We obtain

$$\begin{aligned}\mathbb{E}[(rs)_{ij}(rs)_{kl}] &= \mathbb{P}[(rs)_{ij}=1=(rs)_{il}] \\ &= \frac{n_r}{n}\cdot\frac{n_s}{n-1}\cdot\frac{n_r-1}{n-2}+\frac{n_s}{n}\cdot\frac{n_r}{n-1}\cdot\frac{n_s-1}{n-2} \\ &= \frac{n_rn_s(n_r+n_s-2)}{n^{(3)}}.\end{aligned}$$

The above formula holds also in the case (i,j,k,l) is in anyone of the sets L_{2b}, L_{2c} and L_{2d}.

(c) For $(i,j,k,l) \in L_{3a}$, we have:

$$\begin{aligned}\mathbb{E}[(rs)_{ij}(rs)_{kl}] &= \mathbb{P}[(rs)_{ij}=1] \\ &= \frac{n_r}{n}\cdot\frac{n_s}{n-1}+\frac{n_s}{n}\cdot\frac{n_r}{n-1} \\ &= \frac{2n_rn_s}{n^{(2)}}.\end{aligned}$$

The above formula holds also in the case $(i,j,k,l) \in L_{3b}$.

We now calculate the quantities $\tilde{E}_t$, $t=1,2,3$, given in (2.17):

$$\begin{aligned}\tilde{E}_1 &= \sum_{(i,j,k,l)\in L_1}(w_{ij}+w_{jk})(w_{kl}+w_{lk})\frac{4n_r^{(2)}n_s^{(2)}}{n^{(4)}} \\ &= \frac{4n_r^{(2)}n_s^{(2)}}{n^{(4)}}e_1 = \frac{4n_r^{(2)}n_s^{(2)}}{n^{(4)}}\cdot 4(W^2+S_1-S_2) \\ &= \frac{16n_r^{(2)}n_s^{(2)}(W^2+S_1-S_2)}{n^{(4)}}.\end{aligned}$$

The value of the quantity e_1 is given in equation (2.14). Also,

$$\begin{aligned}\tilde{E}_2 &= \sum_{(i,j,k,l)\in L_2}(w_{ij}+w_{jk})(w_{kl}+w_{lk})\frac{n_rn_s(n_r+n_s-2)}{n^{(3)}} \\ &= \frac{n_rn_s(n_r+n_s-2)}{n^{(3)}}e_2 = \frac{n_rn_s(n_r+n_s-2)}{n^{(3)}}\cdot 4(S_2-2S_1) \\ &= \frac{4(S_2-2S_1)n_rn_s(n_r+n_s-2)}{n^{(3)}}.\end{aligned}$$

The value of the quantity e_2 is given in equation (2.13). Finally,

$$\begin{aligned}\tilde{E}_3 &= \sum_{(i,j,k,l)\in L_3}(w_{ij}+w_{jk})(w_{kl}+w_{lk})\frac{2n_rn_s}{n^{(2)}} \\ &= \frac{2n_rn_s}{n^{(2)}}\cdot 4S_1 \\ &= \frac{8S_1n_rn_s}{n^{(2)}}.\end{aligned}$$

Now it is clear that

$$\mathbb{E}[N_{rs}^2] = \frac{1}{4}\left[\frac{4n_r^{(2)}n_s^{(2)}(W^2+S_1-S_2)}{n^{(4)}}+\frac{(S_2-2S_1)n_rn_s(n_r+n_s-2)}{n^{(3)}}\right.$$

$$+ \frac{2S_1 n_r n_s}{n^{(2)}}\Bigg].$$

Hence

$$\begin{aligned}\mathrm{Var}(N_{rs}) &= \mathbb{E}[N_{rs}^2] - [\mathbb{E}(N_{rs})]^2 \\ &= \frac{1}{4}\left[\frac{2S_1 n_r n_s}{n^{(2)}} + \frac{(S_2 - 2S_1) n_r n_s (n_r + n_s - 2)}{n^{(3)}}\right. \\ &\quad \left. + \frac{4(W^2 + S_1 - S_2) n_r^{(2)} n_s^{(2)}}{n^{(4)}} - 4W^2 \left(\frac{n_r n_s}{n^{(2)}}\right)^2\right]. \qquad (2.18)\end{aligned}$$

2.5 Covariances among the three different join counts

We discussed the first two moments of the three different join counts N_{rr}, N_{ss} and N_{rs} in Sections 2.3 and 2.4. In this section, we calculate the covariances among these three random variables. That is, we calculate[2] $\mathrm{Cov}(N_{rr}, N_{ss})$, $\mathrm{Cov}(N_{rr}, N_{rs})$ and $\mathrm{Cov}(N_{ss}, N_{rs})$ when $r \neq s$. To achieve that, we consider the allocation of types r and s in four locations, i, j, l and k, and we consider the partition of the set L, defined in equation (2.6), into the sets L_i, $i = 1, 2, 3$, in order to find the covariances. See also equations (2.7)-(2.9).

We first calculate $\mathrm{Cov}(N_{rr}, N_{ss})$, which is the covariance of the numbers of rr joins and ss joins. When considering the N_{rr} and N_{ss} join counts together, we are summing over (i, j, k, l) in L such that locations i and j are of type r while k and l are of type s. Since $r \neq s$, we do not have to sum over the sets L_2 and L_3, but only over the set

$$L_1 := \{(i, j, k, l) \in L |\ k \neq i \neq l \text{ and } k \neq j \neq l\}.$$

We then have

$$\mathrm{Cov}(N_{rr}, N_{ss}) = \mathrm{Cov}\left(\frac{1}{2}\sum_{i \neq j} w_{ij}(rr)_{ij}, \frac{1}{2}\sum_{k \neq l} w_{kl}(ss)_{kl}\right)$$

[2]We remind the reader that $\mathrm{Cov}(N_{rr}, N_{ss}) = \mathbb{E}[(N_{rr} - \mathbb{E}(N_{rr}))(N_{ss} - \mathbb{E}(N_{ss}))] = \mathbb{E}(N_{rr}N_{ss}) - \mathbb{E}(N_{rr})\mathbb{E}(N_{ss})$.

$$
\begin{aligned}
&= \frac{1}{16}\text{Cov}\left(\sum_{i\neq j}(w_{ij}+w_{ji})(rr)_{ij}, \sum_{k\neq l}(w_{kl}+w_{lk})(ss)_{kl}\right)\\
&= \frac{1}{16}\sum_{i\neq j}\sum_{k\neq l}(w_{ij}+w_{ji})(w_{kl}+w_{lk})\text{Cov}[(rr)_{ij},(ss)_{kl}]\\
&= \frac{1}{16}\sum_{i\neq j}\sum_{k\neq l}(w_{ij}+w_{ji})(w_{kl}+w_{lk})P[(rr)_{ij}=1,(ss)_{kl}=1]\\
&= \frac{1}{16}\sum_{(i,j,k,l)\in L_1}(w_{ij}+w_{ji})(w_{kl}+w_{lk})\frac{n_r^{(2)}n_s^{(2)}}{n^{(4)}}.
\end{aligned}
$$

Using equations (2.11) and (2.14), it follows that

$$\text{Cov}(N_{rr},N_{ss}) = \frac{n_r^{(2)}n_s^{(2)}}{16n^{(4)}}\cdot e_1 = \frac{n_r^{(2)}n_s^{(2)}}{4n^{(4)}}(W^2+S_1-S_2). \tag{2.19}$$

Next we find the covariance of the numbers of rr joins and rs joins, which is $\text{Cov}(N_{rr},N_{rs})$. We already know that $N_{rr}+N_{ss}+N_{rs}=\frac{W}{2}$. Therefore we use properties of the covariance of linear functions to find $\text{Cov}(N_{rr},N_{rs})$ as follows:

$$
\begin{aligned}
\text{Cov}(N_{rr},N_{rs}) &= \text{Cov}(N_{rr},\frac{W}{2}-N_{rr}-N_{ss})\\
&= \text{Cov}(N_{rr},\frac{W}{2})-\text{Cov}(N_{rr},N_{rr})-\text{Cov}(N_{rr},N_{ss})\\
&= 0-\text{Var}(N_{rr})-\text{Cov}(N_{rr},N_{ss})\\
&= -[\text{Var}(N_{rr})+\text{Cov}(N_{rr},N_{ss})].
\end{aligned}
$$

Similarly, by replacing r with s and vice versa, we can find the covariance of the numbers of ss joins and rs joins as

$$\text{Cov}(N_{ss},N_{rs}) = -[\text{Var}(N_{ss})+\text{Cov}(N_{ss},N_{rr})].$$

2.6 Some more examples

In this section we provide some more examples to illustrate the previous theory.

Example 2. This example is taken from Sokal and Oden (1978), pp. 203-205. Here we compute spatial autocorrelation for a variable that is in nominal scale when we have regularly-spaced localities, binary weights, and sampling without

replacement. The data consist of the colors in nine squares in a checkboard pattern, as shown in Fig. 2.1 (re-produced from Fig. 2 in Sokal and Oden (1978), p. 203).

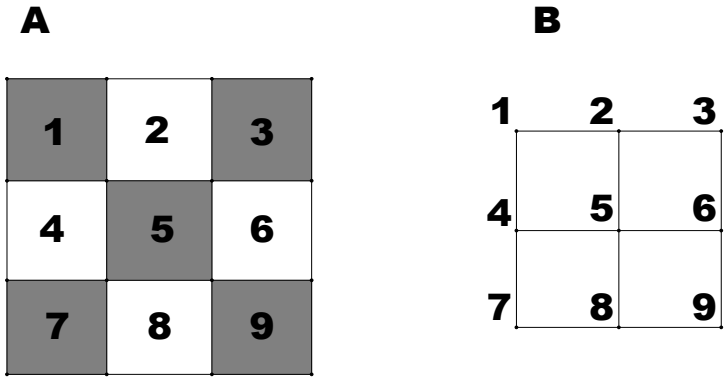

Figure 2.1. Checkerboard and graph for Example 2 (from [16, p. 203])

In this example we first assume that the squares are connected by rook's moves. This means that each square is connected to all squares that are directly below, above, to the left side or to the right side of it. One can construct a graph that illustrates this relationship (connection), and this also shown in Fig. 2.1, where the graph consists of 9 vertices and 12 edges. Each vertex of the graph corresponds to a square, and two vertices are connected by an edge if and only if the corresponding squares are considered adjacent (by rook's moves).

The adjacency matrix (i.e., the matrix of weights M_W) corresponding to this graph is shown in Table 2.1 (also re-produced from Table 1 in Sokal and Oden (1978), p. 203). Next to each locality the type or color (B or W) of the square is presented: We have $n_B = 5$ black squares and $n_W = 4$ white squares.

Even though the numbers of BB, WW and BW joins can be easily read from the checkboard for this example, equations (2.3) and (2.4) from Section 2 of this chapter

Table 2.1. Adjacency matrix M_W for Example 2 (from [16, p. 203])

		Locality Number								
Locality Number	Type	1	2	3	4	5	6	7	8	9
1	B	0	1	0	1	0	0	0	0	0
2	W	1	0	1	0	1	0	0	0	0
3	B	0	1	0	0	0	1	0	0	0
4	W	1	0	0	0	1	0	1	0	0
5	B	0	1	0	1	0	1	0	1	0
6	W	0	0	1	0	1	0	0	0	1
7	B	0	0	0	1	0	0	0	1	0
8	W	0	0	0	0	1	0	1	0	1
9	B	0	0	0	0	0	1	0	1	0

are computed so as to illustrate the numerical calculation of these statistics. In the calculations below neighbors are defined according to rook's moves.

The number of BB joins can be written as $N_{BB} = \frac{1}{2}\sum_{i \neq j} w_{ij}(BB)_{ij}$. In this example, the variable $(BB)_{ij}$ is 1 if both localities i and j are black, and 0 otherwise. Here all the neighboring squares have different colors. Therefore whenever $(BB)_{ij} = 1$, the corresponding w_{ij} is zero. That is, there are no BB joins in this example. Simillarly we can find that there are no WW joins either. (Here the number of WW joins can be found by the equation $N_{WW} = \frac{1}{2}\sum_{i \neq j} w_{ij}(WW)_{ij}$.)

In order to find the number of BW joins, equation $N_{BW} = \frac{1}{2}\sum_{i \neq j} w_{ij}(BW)_{ij}$ can be used. Since all the neighboring squares in this example have different colors, $(BW)_{ij}$ equals to 1 whenever $w_{ij} = 1$. Therefore the number of BW joins is 12, which is the number of all joins in the example. Some of the values for the example are shown in Table 2.2. In summary, the observed values of the join counts are:

$$N_{BB} = 0 = N_{WW} \quad \text{and} \quad N_{BW} = 12.$$

The quantity W is the sum of the weights in the adjacency matrix M_W. Since in

Table 2.2. Calculation of some numbers of joins for Example 2 (from [16, p. 204])

Localities		Weight		
i	j	w_{ij}	$(BB)_{ij}$	$(BW)_{ij}$
1	2	1	0	1
1	3	0	1	0
1	4	1	0	1
1	5	0	1	0
1	6	0	0	1
1	7	0	1	0
1	8	0	0	1
1	9	0	1	0
2	1	1	0	1
2	3	1	0	1
2	4	0	0	0
2	5	1	0	1
2	6	0	0	0
2	7	0	0	1
2	8	0	0	0
2	9	0	0	1
3	1	0	1	0
3	2	1	0	1
3	4	0	0	1
⋮	⋮	⋮	⋮	⋮
9	6	1	0	1
9	7	0	1	0
9	8	1	0	1
			$\sum_{i \neq j} w_{ij}(BB)_{ij} = 0$	$\sum_{i \neq j} w_{ij}(BW)_{ij} = 24$

this example we have binary symmetrical weights, W is 24. Also, recall that the board has a total of $n = 9$ squares: 5 of them are black and the other 4 are white, i.e., $n_B = 5$ and $n_W = 4$.

Equation (2.5) can be used to find the expected numbers of BB and WW joins

under the null hypothesis (2.1):

$$\mathbb{E}[N_{BB}] = \frac{Wn_B^{(2)}}{2n^{(2)}} = \frac{24 \cdot 5 \cdot 4}{2 \cdot 9 \cdot 8} = \frac{10}{3}.$$
$$\mathbb{E}[N_{WW}] = \frac{Wn_W^{(2)}}{2n^{(2)}} = \frac{24 \cdot 4 \cdot 3}{2 \cdot 9 \cdot 8} = 2.$$

Using equation (2.16), the expected number of BW joins under the null hypothesis (2.1) can be calculated:

$$\mathbb{E}[N_{BW}] = \frac{Wn_W n_B}{n^{(2)}} = \frac{24 \cdot 5 \cdot 4}{9 \cdot 8} = \frac{20}{3}.$$

To compute the variances using equations (2.15) and (2.18), the two quantities defined in Section 2.3, $S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2$ and $S_2 = \sum_{i=1}^{n}(w_{i.} + w_{.i})^2$, are needed. For each pair of localities i and j connected by an edge in the connectivity graph B in Fig. 2.1, the contribution to the sum S_1 is $2(1+1)^2/2 = 4$. Therefore

$$S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2 = 4 \times (\text{number of edges in the graph}) = 48.$$

On the other hand, for each locality i, $w_{i.} = w_{.i}$ counts the number of localities to which i is connected. Thus

$$S_2 = \sum_{i=1}^{n}(w_{i.} + w_{.i})^2 = 4(2^2 + 3^2 + 2^2 + 3^2 + 4^2 + 3^2 + 2^2 + 3^2 + 2^2) = 272.$$

The variance of the number of BB joins can be calculated as follows using equation (2.15):

$$\begin{aligned}
\text{Var}[N_{BB}] &= \frac{1}{4}\left[\frac{S_1 n_B^{(2)}}{n^{(2)}} + \frac{(S_2 - 2S_1)n_B^{(3)}}{n^{(3)}} + \frac{(W^2 + S_1 - S_2)n_B^{(4)}}{n^{(4)}} - W^2\left(\frac{n_B^{(2)}}{n^{(2)}}\right)^2\right] \\
&= \frac{1}{4}\left[\frac{48 \cdot 5 \cdot 4}{9 \cdot 8} + \frac{(272 - 2 \cdot 48) \cdot 5 \cdot 4 \cdot 3}{9 \cdot 8 \cdot 7} + \right. \\
&\quad \left. + \frac{(24^2 + 48 - 272) \cdot 5 \cdot 4 \cdot 3 \cdot 2}{9 \cdot 8 \cdot 7 \cdot 6} - 24^2\left(\frac{5 \cdot 4}{9 \cdot 8}\right)^2\right] = 0.95238.
\end{aligned}$$

Similarly, we find that $\text{Var}[N_{WW}] = 0.79365$. To find the variance for BW we use a

different equation (see (2.18)):

$$\begin{aligned}
\text{Var}[N_{BW}] &= \frac{1}{4}\left[\frac{2S_1 n_B n_W}{n^{(2)}} + \frac{(S_2 - 2S_1)n_B n_W (n_B + n_W - 2)}{n^{(3)}}\right. \\
&\quad \left. + \frac{4(W^2 + S_1 - S_2)n_B^{(2)} n_W^{(2)}}{n^{(4)}} - 4W^2\left(\frac{n_B n_W}{n^{(2)}}\right)^2\right] \\
&= \frac{1}{4}\left[\frac{2 \cdot 48 \cdot 5 \cdot 4}{9 \cdot 8} + \frac{(272 - 2 \cdot 48) \cdot 5 \cdot 4 \cdot (5 + 4 - 2)}{9 \cdot 8 \cdot 7}\right. \\
&\quad \left. + \frac{4 \cdot (24^2 + 48 - 272) \cdot 5 \cdot 4 \cdot 4 \cdot 3}{9 \cdot 8 \cdot 7 \cdot 6} - 4 \cdot 24^2 \cdot \left(\frac{5 \cdot 4}{9 \cdot 8}\right)^2\right] \\
&= 2.38095.
\end{aligned}$$

This completes Example 2. □

2.7 Connection of join counts to Moran's I autocorrelation coefficient

In 1950 Moran introduced his famous *spatial autocorrelation coefficient* I that generalizes the time-series autocorrelation coefficient. A more detailed analysis of this coefficient will be given in the next chapter. Here we show the connection between join counts and Moran's I. If the observed data are $x_1, x_2, \ldots, x_n$, where x_i is the value of the variable of interest at location i, then the (modern version of the) spatial autocorrelation coefficient according to Moran (1950) is given by

$$I = \frac{n \sum_{i \neq j} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{W \sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

To examine the connection of Moran's I with the join counts, assume we have only two types of objects, r and s, and define:

$$x_i = \begin{cases} 1 & \text{if locality } i \text{ is of type } r, \\ 0 & \text{if locality } i \text{ is of type } s. \end{cases}$$

It follows that $n_r = \sum_{i=1}^{n} x_i$ and $n_s = \sum_{i=1}^{n}(1 - x_i)$. In addition, sample means can be interpreted as proportions: $\bar{x} = \frac{n_r}{n}$ and $1 - \bar{x} = \frac{n_s}{n}$. We will prove equation (2.24),

which is the main formula of the section:

$$I = \frac{2n}{W}\left[\frac{N_{rr}}{n_r} + \frac{N_{ss}}{n_s}\right] - 1.$$

Of course, we have to assume that at least one weight w_{ij} is not zero and no weight is negative, i.e., $W > 0$.

To prove the above formula, we use the definition of Moran's I to define $U := n\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})$ and $V := W\sum_{i=1}^{n}(x_i - \bar{x})^2$, so that $I = \frac{U}{V}$.

Consider

$$V = W\sum_{i=1}^{n}(x_i - \bar{x})^2 = W\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = W\left[\sum_{i=1}^{n} x_i - n\bar{x}^2\right].$$

It follows that

$$V = W\left[n_r - n\left(\frac{n_r}{n}\right)^2\right] = W\frac{n_r}{n}(n - n_r) = \frac{Wn_rn_s}{n}. \tag{2.20}$$

Consider also

$$\begin{aligned}
\frac{U}{n} &= \sum_{i\neq j} w_{ij}(x_i - \bar{x})(x_j - \bar{x}) \\
&= \frac{1}{2}\sum_{i\neq j}(w_{ij} + w_{ji})(x_i - \bar{x})(x_j - \bar{x}) \\
&= \frac{1}{2}\left[\sum_{i\neq j}(w_{ij} + w_{ji})x_ix_j - \bar{x}\sum_{i\neq j}(w_{ij} + w_{ji})(x_i + x_j) + \bar{x}^2\sum_{i\neq j}(w_{ij} + w_{ji})\right] \\
&= \frac{1}{2}\left[\sum_{\substack{i\neq j \\ \text{s.t. both i and j are r}}}(w_{ij} + w_{ji}) - \bar{x}\sum_{i\neq j} w_{ij}x_i - \bar{x}\sum_{i\neq j} w_{ji}x_i - \bar{x}\sum_{i\neq j} w_{ij}x_j\right. \\
&\quad \left. - \bar{x}\sum_{i\neq j} w_{ji}x_j + \bar{x}^2 \cdot 2W\right] \\
&= \frac{1}{2}\sum_{i\neq j}(w_{ij} + w_{ji})(rr)_{ij} - \frac{1}{2}\bar{x}\left[\sum_{i=1}^{n} w_{i.}x_i + \sum_{i=1}^{n} w_{.i}x_i + \sum_{j=1}^{n} w_{.j}x_j + \sum_{j=1}^{n} w_{j.}x_j\right] \\
&\quad + \left(\frac{n_r}{n}\right)^2 W
\end{aligned}$$

$$= \frac{1}{2}\sum_{i\neq j} w_{ij}(rr)_{ij} + \frac{1}{2}\sum_{i\neq j} w_{ji}(rr)_{ij} - \bar{x}\sum_{i=1}^{n}(w_{i.}+w_{.i})x_i + W\left(\frac{n_r}{n}\right)^2.$$

Therefore

$$\frac{U}{n} = 2N_{rr} - \left(\frac{n_r}{n}\right)\sum_{\substack{i=1\\ \text{i is r}}}^{n}(w_{i.}+w_{.i}) + W\left(\frac{n_r}{n}\right)^2. \tag{2.21}$$

For $i \neq j$ we have $(rr)_{ij} + (rs)_{ij} + (ss)_{ij} = 1$, thus

$$N_{rr} + N_{rs} + N_{ss} = \frac{W}{2}. \tag{2.22}$$

In addition,

$$\sum_{i \text{ is r}}(w_{i.}+w_{.i}) = \sum_{\substack{i \text{ and j}\\ \text{are both r}}}(w_{ij}+w_{ji}) + \sum_{\substack{i \text{ is r}\\ \text{and j is s}}}(w_{ij}+w_{ji}) = 4N_{rr} + 2N_{rs}.$$

Using the above results we can rewrite the equation (2.21) as follows:

$$\begin{aligned}
\frac{U}{n} &= 2N_{rr} - \frac{n_r}{n}\sum_{\substack{i=1\\ i \text{ is r}}}^{n}(w_{i.}+w_{.i}) + W\left(\frac{n_r}{n}\right)^2\\
&= 2N_{rr} - \frac{n_r}{n}(4N_{rr}+2N_{rs}) + W\left(\frac{n_r}{n}\right)^2\\
&= 2\left(\frac{n_r+n_s}{n}\right)N_{rr} - \frac{n_r}{n}(4N_{rr}+2N_{rs}) + W\left(\frac{n_r}{n}\right)^2\\
&= \frac{2n_s}{n}N_{rr} - \frac{2n_r}{n}N_{rs} - \frac{2n_r}{n}N_{rr} + W\left(\frac{n_r}{n}\right)^2.
\end{aligned}$$

Using (2.22), we obtain

$$\begin{aligned}
\frac{U}{n} &= \frac{2n_s}{n}N_{rr} - \frac{2n_r}{n}\left(\frac{W}{2} - N_{rr} - N_{ss}\right) - \frac{2n_r}{n}N_{rr} + W\left(\frac{n_r}{n}\right)^2\\
&= \frac{2n_sN_{rr}}{n} + \frac{2n_rN_{ss}}{n} + W\left(\frac{n_r^2}{n^2} - \frac{n_r}{n}\right)\\
&= \frac{2n_sN_{rr}}{n} + \frac{2n_rN_{ss}}{n} - \frac{Wn_rn_s}{n^2}.
\end{aligned}$$

Therefore

$$U = 2\left[n_sN_{rr} + n_rN_{ss}\right] - \frac{Wn_rn_s}{n}. \tag{2.23}$$

Equations (2.20) and (2.23) give

$$I = \frac{U}{V} = \frac{2[n_sN_{rr} + n_rN_{ss}] - Wn_rn_s/n}{Wn_rn_s/n},$$

Hence

$$I = \frac{2n}{W}\left[\frac{N_{rr}}{n_r} + \frac{N_{ss}}{n_s}\right] - 1. \tag{2.24}$$

This is the main formula of the section. Below we analyze this formula.

Note that if $N_{rr} = 0 = N_{ss}$, then $I = -1$. Conversely, if $I = -1$, we have

$$\frac{2n}{W}\left[\frac{N_{rr}}{n_r} + \frac{N_{ss}}{n_s}\right] = 0.$$

In such a case, since $N_{rr} \geq 0$, $N_{ss} \geq 0$, $n_r > 0$ and $n_s > 0$, we conclude that $N_{rr} = N_{ss} = 0$. Therefore

$$I = -1 \Leftrightarrow N_{rr} = N_{ss} = 0.$$

It is thus clear that $I \geq -1$ with equality iff $N_{rr} = 0 = N_{ss}$.

Next it is natural to ask whether it is always the case that $I \leq 1$. The answer is no. To see this, observe first

$$\begin{aligned} I \leq 1 &\Leftrightarrow \frac{N_{rr}}{n_r} + \frac{N_{ss}}{n_s} \leq \frac{W}{n} \\ &\Leftrightarrow \frac{N_{rr}}{n_r} + \frac{N_{ss}}{n_s} \leq \frac{2(N_{rr} + N_{ss} + N_{rs})}{n_r + n_s} \\ &\Leftrightarrow \frac{n_s}{n_r}N_{rr} + \frac{n_r}{n_s}N_{ss} \leq N_{rr} + N_{ss} + 2N_{rs}. \end{aligned}$$

Therefore

$$I \leq 1 \Leftrightarrow 2N_{rs} \geq \left(\frac{n_s}{n_r} - 1\right) N_{rr} + \left(\frac{n_r}{n_s} - 1\right) N_{ss} \tag{2.25}$$

$$\Leftrightarrow 2N_{rs} \geq (n_s - n_r)\left(\frac{N_{rr}}{n_r} - \frac{N_{ss}}{n_s}\right). \tag{2.26}$$

Inequality (2.25) does not always hold! In other words, it is possible that spatial autocorrelation $I > 1$!

Before we provide an example where $I > 1$, let's examine some practical cases where I is less than 1. Note that inequality (2.25) can be rewritten as

$$2N_{rs} \geq AN_{rr} + BN_{ss},$$

where $A = (n_s/n_r) - 1$ and $B = (n_r/n_s) - 1$.

If $n_r = n_s = n/2$, i.e., the set of localities is even and the number of objects of type r equals the number of objects of type s, then $A = B = 0$. In such a case the right inequality in (2.25) reduces to $N_{rs} \geq 0$, which is of course true, and thus $I \leq 1$.

But if $n_r > n_s$ then $A < 0$ and $B > 0$, while if $n_r < n_s$ then $A > 0$ and $B < 0$. Therefore it is difficult to say whether I is less than or greater than 1 when $n_r \neq n_s$. But clearly

$$I > 1 \quad \Leftrightarrow \quad 2N_{rs} < \left(\frac{n_s}{n_r} - 1\right) N_{rr} + \left(\frac{n_r}{n_s} - 1\right) N_{ss} \tag{2.27}$$

$$\Leftrightarrow \quad 2N_{rs} < (n_s - n_r)\left(\frac{N_{rr}}{n_r} - \frac{N_{ss}}{n_s}\right). \tag{2.28}$$

Example 3. This is an example which gives $I > 1$. The checkerboard and the connectivity graph are given in Fig. 2.2.

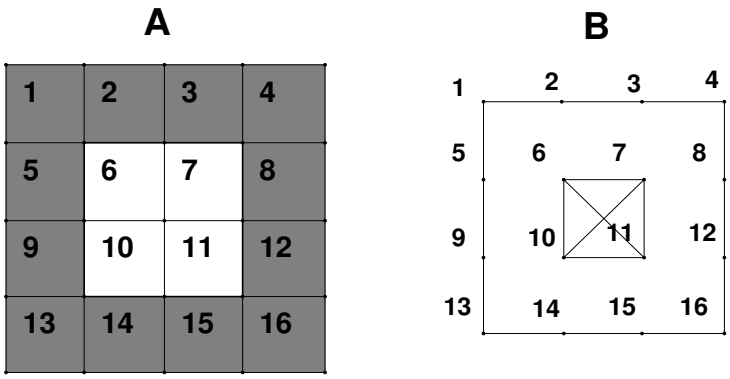

Figure 2.2. Checkerboard and graph for Example 3

Here $n_r = 12$, $n_s = 4$ and $n = 16$. Assume that localities 6, 7, 10 and 11 are disjoint and disconnected from the other localities. Suppose a study is conducted to compare the soil composition. Localities 6, 7, 10 and 11 are on an island inside a lake and it is assumed they have soil of type s, which contains more water. The other 12 localities are located around the lake and it is assumed they have soil of

type r, which contains less water than the two localities inside the lake. The lake is a barrier that practically disconnects the localities 6, 7, 10 and 11 from the other 16 localities. In this case, the number of rs joins is zero, i.e., $N_{rs} = 0$. Assume that around the lake two localities are connected iff they are next to each other (connected by rook's moves), while all four localities on the island are connected to each other (connected by queen's moves).

The weights can be defined as follows:

$$w_{ij} = w_{ji} = 0 \quad \text{for} \quad i \in \{1, 2, 3, 4, 5, 8, 9, 12, 13, 14, 15, 16\} \text{ and } j \in \{6, 7, 10, 11\},$$

while

$$w_{jl} = 1 = w_{lj} \quad \text{for } j \neq l \in \{6, 7, 10, 11\}.$$

In addition, $w_{ik} = 1$ for

$$\begin{aligned}(i,k) \quad \in \quad & \{(1,2),(2,3),(3,4),(4,8),(8,12),(12,16),\\ & (16,15),(15,14),(14,13),(13,9),(9,5),(5,1)\}\end{aligned}$$

and

$$\begin{aligned}(i,k) \quad \in \quad & \{(1,5),(5,9),(9,13),(13,14),(14,15),(15,16),\\ & (16,12),(12,8),(8,4),(4,3),(3,2),(2,1)\}.\end{aligned}$$

All other weights are zero.

The numbers of rr and ss joins are given below:

$$\begin{aligned}N_{rr} &= \frac{1}{2}\sum_{i \neq j} w_{ij}(rr)_{ij} = \frac{1}{2}(24 + 0) = 12;\\ N_{ss} &= \frac{1}{2}(12) = 6.\end{aligned}$$

Substituting the above results in the right inequality in (2.27) we get

$$0 < \left(\frac{4}{12} - 1\right) 12 + \left(\frac{12}{4} - 1\right) 6 = -8 + 12 = 4,$$

which of course holds. Therefore in such a situation, $I > 1$. That is, there exist real situations which have spatial autocorrelation greater than 1. Actually, since in this case, $W = 36$, we obtain from formula (2.24) that $I = 11/9 > 1$. □

Note that if one changes the definition of the binary variable x_i so that

$$x_i = \begin{cases} 0 & \text{if locality } i \text{ is of type } r, \\ 1 & \text{if locality } i \text{ is of type } s, \end{cases}$$

then the derived formula (2.24) for Moran's I stays the same because it is symmetric with respect to r and s – see the SAS program in Example 4 in Section 2.10 below.

2.8 Connection of join counts to Geary's ratio c

In 1954 Geary introduced his famous *spatial contiguity*[3] *ratio* c, which can be considered as an alternative measure of spatial autocorrelation to Moran's I. A more detailed analysis of this coefficient will be given in the next chapter. Here we show the connection between the join counts and the Geary ratio c. If the observed data are $x_1, x_2, \ldots, x_n$, where x_i is the value of the variable of interest at location i, then the contiguity ratio c according to Geary (1954) is given by

$$c = \text{GR} = \frac{(n-1)\sum\limits_{i \neq j} w_{ij}(x_i - x_j)^2}{2W\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}.$$

To examine the connection of Geary's ratio c with the join counts, assume we have only two types of objects, r and s, and (as in the previous section) define:

$$x_i = \begin{cases} 1 & \text{if locality } i \text{ is of type } r, \\ 0 & \text{if locality } i \text{ is of type } s. \end{cases}$$

It follows that $n_r = \sum\limits_{i=1}^{n} x_i$ and $n_s = \sum_{i=1}^{n}(1 - x_i)$. In addition, sample means can be interpreted as proportions: $\bar{x} = \frac{n_r}{n}$ and $1 - \bar{x} = \frac{n_s}{n}$. We will prove the following

[3]The word "contiguity" means "adjacency" or "proximity."

equation, which is the main formula of the section:

$$c = \text{GR} = \frac{n(n-1)N_{rs}}{Wn_rn_s}. \tag{2.29}$$

We have to assume that at least one weight w_{ij} is not zero and no weight is negative, i.e., $W > 0$.

Let $c = \text{GR} = U/V$ such that

$$U = (n-1)\sum_{i\neq j} w_{ij}(x_i - x_j)^2 \text{ and } V = 2W\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

First we consider the denominator:

$$\begin{aligned} V &= 2W\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= 2W\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \\ &= 2W\left(n_r - n\cdot\frac{n_r^2}{n^2}\right) \\ &= \frac{2Wn_r}{n}(n - n_r) \\ &= \frac{2nWn_rn_s}{n}. \end{aligned}$$

Next we consider the numerator:

$$\begin{aligned} U &= (n-1)\sum_{i\neq j} w_{ij}(x_i - x_j)^2 \\ &= (n-1)\sum_{i\neq j} w_{ij}(x_i^2 - 2x_ix_j + x_j^2) \\ &= (n-1)\left(\sum_{i\neq j} w_{ij}x_i^2 - 2\sum_{i\neq j} w_{ij}x_ix_j + \sum_{i\neq j} w_{ij}x_j^2\right) \\ &= (n-1)\left(\sum_{i \text{ is r}} w_{ij} - 2\sum_{\substack{\text{both i and j}\\ \text{are r}}} w_{ij} + \sum_{j \text{ is r}} w_{ij}\right) \\ &= (n-1)\left(\sum_{i\neq j} w_{ij}(rr)_{ij} + \sum_{i \text{ is r}} w_{ij}(rs)_{ij} - 2\sum_{i\neq j} w_{ij}(rr)_{ij}\right. \end{aligned}$$

$$+\sum_{i\neq j} w_{ij}(rr)_{ij} + \sum_{j \text{ is r}} w_{ij}(sr)_{ij}\Bigg).$$

It follows that

$$U = 2(n-1)N_{rs}.$$

Substituting U and V in the equation for c, we obtain

$$c = \frac{U}{V} = \frac{2(n-1)N_{rs}}{2Wn_rn_s/n} = \frac{n(n-1)N_{rs}}{Wn_rn_s}.$$

Note that (under the null hypothesis H_0 in equation (2.1)), we have

$$\mathbb{E}(c) = \frac{n(n-1)\mathbb{E}(N_{rs})}{Wn_rn_s} = 1, \tag{2.30}$$

where we have used equation (2.16).

2.9 Calculation of join counts and their moments in R

In this section we show how one can find the observed values of the three join counts along with their means and variances (under the assumption of sampling without replacement) in the statistical programming language R. These tasks are achieved through the use of an applied example.

Example 4. This example is taken from Chapter 3 in Griffith (1987). It is based on the famous Buffalo, NY, crime data in 1981. Although there are many types of crime data in the data set, we consider only the presence or the absence of prostitution arrests for each police precinct. Thus there are only two types of data (presence or absence of prostitution arrests) distributed among 14 localities (police precincts). The "checkerboard" (graph A) and the connectivity graph (graph B) are given in Fig. 2.3. Note that not all numbers from 1 to 17 are represented in the numbering of the 14 police precincts in Buffalo, NY: numbers 1, 2 and 14 do not appear in the numbering.[4]

[4]The first plot in the figure is not really a checkerboard (since it consists mostly of rectangles of unequal size and other polygons). In addition, the area and shape of each polygon does not

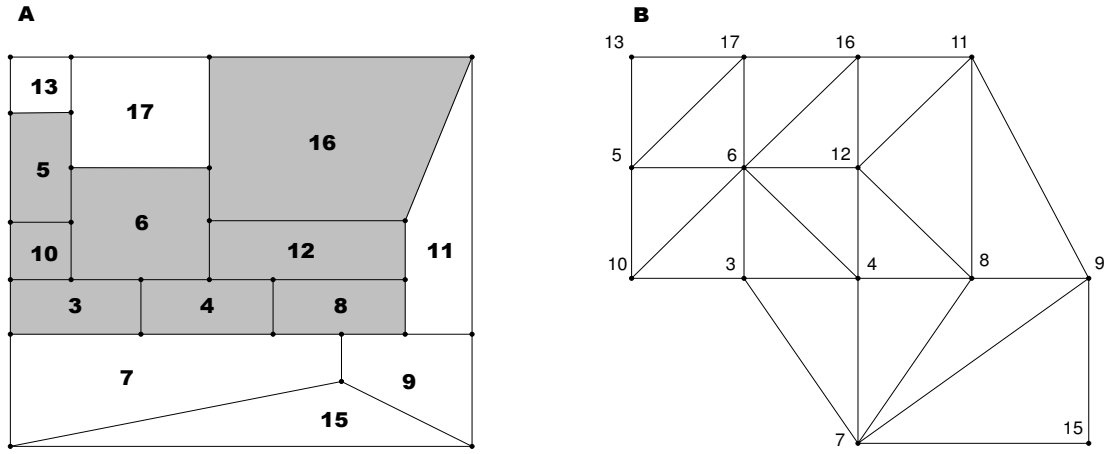

Figure 2.3. "Checkerboard" and connectivity graph for the presence or absence of prostitution among 14 police precincts (B = r = yes, W = s = no)

In Fig. 2.3 the precincts with prostitution arrests are specified in black color and the precincts without prostitution arrests are specified in white color. There are 13 "black-black" joins which we consider as rr joins, 5 "white-white" joins which we consider as ss joins, and 11 "white-black" or "black-white" joins which we consider as sr joins. The adjacency matrix for this example is shown in Table 2.3. In this example, $n_r = 8$ and $n_s = 6$.

The package *spdep*[5] in the statistical package R was used to find the expected values, variances and z-scores for the three different join counts (N_{rr}, N_{ss}, and N_{rs}). This package was written for R by Roger Bivand with contributions from other people. We used version 0.5-19 of this package.

Below one can find a code, written in the statistical programming language R, that can be used to produce expected values, variances and z-scores for the three

correspond to the actual size of the corresponding precinct.

[5]See the website `http://cran.r-project.org/web/packages/spdep/spdep.pdf` (accessed December 31, 2010).

Table 2.3. Adjacency matrix M_W for Example 4 (from [11, p. 26])

		Precinct No.													
Precinct No.	Type	3	4	5	6	7	8	9	10	11	12	13	15	16	17
3	yes (r)	0	1	0	1	1	0	0	1	0	0	0	0	0	0
4	yes (r)	1	0	0	1	1	1	0	0	0	1	0	0	0	0
5	yes (r)	0	0	0	1	0	0	0	1	0	0	1	0	0	1
6	yes (r)	1	1	1	0	0	0	0	1	0	1	0	0	1	1
7	no (s)	1	1	0	0	0	1	1	0	0	0	0	1	0	0
8	yes (r)	0	1	0	0	1	0	1	0	1	1	0	0	0	0
9	no (s)	0	0	0	0	1	1	0	0	1	0	0	1	0	0
10	yes (r)	1	0	1	1	0	0	0	0	0	0	0	0	0	0
11	no (s)	0	0	0	0	0	1	1	0	0	1	0	0	1	0
12	yes (r)	0	1	0	1	0	1	0	0	1	0	0	0	1	0
13	no (s)	0	0	1	0	0	0	0	0	0	0	0	0	0	1
15	no (s)	0	0	0	0	1	0	1	0	0	0	0	0	0	0
16	yes (r)	0	0	0	1	0	0	0	0	1	1	0	0	0	1
17	no (s)	0	0	1	1	0	0	0	0	0	0	1	0	1	0

different join counts. The code makes use of the R package *spdep*. (One should load this package first before running the program.)

```
buffalo <- read.gwt2nb("D:/CHALANI/research/buffalo_spdep.txt",
                  region.id=c(3,4,5,6,7,8,9,10,11,12,13,15,16,17)))
buffcolor <- factor(c(0,0,0,0,1,0,1,0,1,0,1,1,0,1),labels=c("r","s"))
buffalo.w <- nb2listw(buffalo,style="B")
joincts <- joincount.multi(fx=buffcolor,listw=buffalo.w)
```

In the above R program the text file `buffalo_spdep.txt` should contain the number of precincts on the first line, and in each subsequent line, the identifying number of a precinct should be followed by the identifying number of one of its neighbours and then by the corresponding weight w_{ij} between the two precincts (in this case, $w_{ij} = 1$). In this example, if precinct i is neighbour to precinct j, then precinct j is also considered a neighbour to precinct i (with the same weight of 1),

so there must be two lines in the file to denote this symmetric relationship. Here is part of the text file `buffalo_spdep.txt` for the Buffalo, NY, crime data:

```
14
3  4 1
3  6 1
3  7 1
3 10 1
4  3 1
4  6 1
4  7 1
4  8 1
4 12 1
5  6 1
5 10 1
5 13 1
5 17 1
6  3 1
6  4 1
6  5 1
6 10 1
......
```

The file has $1 + 2(29) = 59$ lines. Note that option `region.id` in the R command `read.gwt2nb` allows for the identification of the 14 precincts with the numbers 3,4,...,17 (where the numbers 1, 2 and 14 are excluded). The R command `read.gwt2nb` creates a neighbour nb object that is used extensively in the package *spdep* for many purposes. The R command `factor` allows for the assignment of the levels of the factor (nominal variable) "presence or absence of prostitution" to the 14 precincts: the levels are numbered 0 and 1, and the label "r" corresponds to level 0 (presence of prostitution), while the label "s" goes to level 1 (absence of prostitution). The R command `nb2listw` creates spatial weights for neighbours lists, and the option "B" corresponds to basic binary coding. Finally, the R command `joincount.multi` creates various join counts statistics for a factor with multiple levels. (For best results, it is recommended that this command be used only with symmetric weights, i.e., for weights that satisfy $w_{ij} = w_{ji}$ for all i and j.) The output from the above code in R is given below.

```
-----------------------OUTPUT OF R-SPDEP-----------------------------
       Joincount Expected Variance z-value
r:r    13.0000   8.9231   3.2039  2.2777
s:s     5.0000   4.7802   2.2494  0.1465
s:r    11.0000  15.2967   5.1797 -1.8879
Jtot   11.0000  15.2967   5.1797 -1.8879
```

Based on the previous output, for the observed values of the join counts, we have $N_{rr} = 13$, $N_{ss} = 5$ and $N_{rs} = 11$, while for the expected values we have $\mathbb{E}[N_{rr}] = 8.9231$, $\mathbb{E}[N_{ss}] = 4.7802$ and $\mathbb{E}[N_{rs}] = 15.2967$. As for the variances for the three join counts, we have $\text{Var}[N_{rr}] = 3.2039$, $\text{Var}[N_{ss}] = 2.2494$ and $\text{Var}[N_{rs}] = 5.1797$, respectively. The rs join count has a comparably higher variance than the other two join counts. The z-scores for the three join counts, N_{rr}, N_{ss}, and N_{rs} are given as 2.2777, 0.1465 and -1.8879, respectively. The corresponding (two-tailed) p-values are given as 0.023, 0.883 and 0.059. Even though the p-values for the N_{rr} and N_{rs} join counts are not as large compared to the p-value for the N_{ss} join count, neither one is significant at the $\alpha = .01$ level of significance. The N_{rr} join count is significant at the $\alpha = .05$ level, but one should be cautious in reaching any conclusion at this level because we are checking the significance of three statistics simultaneously.

This completes the implementation of the data from Example 4 in R using the package *spdep*. □

2.10 Calculation of spatial autocorrelation statistics in SAS

In this section, we show how one can find the spatial autocorrelation statistics due to Moran (1950) and Geary (1954) using the procedure *variogram* in the statistical programming language SAS® (Version 9.2)[6]. As we said before, Moran's I and Geary's ratio c would be explained in more detail in the next chapter.

[6]The statistical package SAS (and all of its products) is owned by SAS Institute Inc., Cary, NC, USA.

In the following SAS program, the weights are binary (i.e., w_{ij} is either 0 or 1), and one has to find "artificial" coordinates in the connectivity graph B in Fig. 2.3 for each of the 14 police precincts so that two vertices (precincts) are connected by an edge if and only if their distance is less than the user specified `lagdistance` (set here to be equal to 1). The general solution to this problem (i.e., the determination of these "artificial" coordinates so that the above condition holds) is not known. For the connectivity graph B in Fig. 2.3 such "artificial" coordinates where produced by Dr. Chris Monico using some heuristical procedure. For example, the "distance between precincts 3 and 4 is

$$d_{3,4} = \sqrt{(4.873 - 5.606)^2 + (4.610 - 4.854)^2} = 0.597 < 1,$$

so the two precincts are connected. On the other hand, the distance between precincts 3 and 17 is

$$d_{3,17} = \sqrt{(4.873 - 4.780)^2 + (4.610 - 6.206)^2} = 2.556 > 1,$$

so the two precincts are *not* connected.

The variable `col` denotes the two levels of the factor (nominal variable): 0 means that the precinct has prostitution arrests, while 1 means that the precinct has no prostitution arrests. Recall that it does not matter which level of the binary factor gets the number 0 and which level gets the number 1: the formulas for Moran's I and Geary's c (i.e., equations (2.24) and (2.29)) stay the same.

```
------------------------SAS PROGRAM----------------------------------
DATA buffalo;
INPUT precinct col x y;
CARDS;
 3  0  4.873  4.610
 4  0  5.606  4.854
 5  0  4.228  5.763
 6  0  5.088  5.312
 7  1  5.724  4.358
 8  0  6.139  5.210
 9  1  6.596  4.627
10  0  4.179  5.004
```

```
11  1  6.402  5.468
12  0  5.557  5.352
13  1  4.164  6.608
15  1  6.507  4.007
16  0  5.659  6.120
17  1  4.780  6.206
;
PROC VARIOGRAM DATA=buffalo outacweights=outbf;
COMPUTE  autocorrelation(weights=binary) lagdistance=1 novariogram ;
COORDINATES xc=x yc=y;
VAR col;
RUN;
```

The output of the above SAS program is given below. It gives the expected values, standard deviations, z-scores and the corresponding p-values for Moran's I and Geary's ratio $c = \text{GR}$.

```
------------------------OUTPUT OF SAS PROGRAM-------------------------
The VARIOGRAM Procedure
                          Dependent Variable: col
                   Number of Observations Read          14
                   Number of Observations Used          14
                              Pairs Information
                  Number of Lags                          11
                  Lag Distance                          0.36
                  Maximum Data Distance in x            2.43
                  Maximum Data Distance in y            2.60
                  Maximum Data Distance                 3.56
                         Pairwise Distance Intervals
                                                 Number
            Lag                                     of    Percentage
          Class    ---------Bounds---------     Pairs      of Pairs
              0          0.00          0.18         0        0.00%
              1          0.18          0.53         4        4.40%
              2          0.53          0.89        18       19.78%
              3          0.89          1.25        16       17.58%
              4          1.25          1.60        17       18.68%
              5          1.60          1.96        16       17.58%
              6          1.96          2.31         9        9.89%
              7          2.31          2.67         6        6.59%
              8          2.67          3.03         3        3.30%
              9          3.03          3.38         1        1.10%
             10          3.38          3.74         1        1.10%
                          Autocorrelation Statistics
                                                Std
Assumption  Coefficient Observed Expected    Dev       Z   Pr > |Z|
Normality   Moran's I      0.187  -0.0769  0.148    1.79     0.0738
Normality   Geary's c      0.719   1.0000  0.178   -1.58     0.1138
```

Even though we assume random sampling without replacement (that is, *nonfree sampling*) for this study (see Section 2.1), in the SAS Variogram procedure the default assumption is that the data are normally and independently distributed. Below we verify the SAS output using formulas derived in previous sections and some new formulas.

Geary's ratio $c = \text{GR}$ for join counts is discussed in Section 2.8:

$$c = \frac{n(n-1)N_{rs}}{Wn_rn_s} = \frac{14 \cdot 13 \cdot 11}{58 \cdot 8 \cdot 6} = 0.719.$$

We have already proved in Section 2.8 that the expected value for c (under the H_0 given in equation (2.1)) is always 1. Now let's find the variance of Geary's ratio c for a random data set with sample sizes equal to those of the data set under study:

$$\text{Var}(c) = \text{Var}\left(\frac{n(n-1)}{Wn_rn_s}N_{rs}\right) = \frac{n^2(n-1)^2}{W^2n_r^2n_s^2}\text{Var}(N_{rs}).$$

First we need to find S_1 and S_2 in order to calculate $\text{Var}(N_{rs})$ as given in equation (2.18):

$$S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2 = \frac{2^2W}{2} = 2 \times 58 = 116.$$

In addition,

$$\begin{aligned} S_2 &= \sum_{i=1}^{n}(w_{i.} + w_{.i})^2 \\ &= 8^2 + 10^2 + 8^2 + 14^2 + 10^2 + 10^2 + 8^2 + 6^2 + 8^2 + 10^2 + 4^2 + 4^2 + 8^2 + 8^2 \\ &= 1048. \end{aligned}$$

Thus,

$$\begin{aligned} \text{Var}(N_{rs}) &= \frac{1}{4}\left[\frac{2S_1n_rn_s}{n^{(2)}} + \frac{(S_2 - 2S_1)n_rn_s(n_r + n_s - 2)}{n^{(3)}}\right. \\ &\quad \left. + \frac{4(W^2 + S_1 - S_2)n_r^{(2)}n_s^{(2)}}{n^{(4)}} - 4W^2\left(\frac{n_rn_s}{n^{(2)}}\right)^2\right] \\ &= \frac{1}{4}\left[\frac{2 \cdot 116 \cdot 8 \cdot 6}{14 \cdot 13} + \frac{(1048 - 232) \cdot 8 \cdot 6 \cdot (8 + 6 - 2)}{14 \cdot 13 \cdot 12}\right. \end{aligned}$$

$$+\frac{4\cdot(58^2+116-1048)\cdot 8\cdot 7\cdot 6\cdot 5}{14\cdot 13\cdot 12\cdot 11}-4\left(\frac{58\cdot 8\cdot 6}{14\cdot 13}\right)^2\Bigg]$$
$$= 5.1797.$$

Hence,

$$\begin{aligned}\mathrm{Var}(c) &= \frac{n^2(n-1)^2}{W^2 n_r^2 n_s^2}\mathrm{Var}(N_{rs})\\ &= \frac{14^2\cdot 13^2}{58^2\cdot 8^2\cdot 6^2}\cdot 5.1797\\ &= 0.0221.\end{aligned}$$

Therefore the standard deviation of Geary's Ratio c is found as 0.1487. This value, however, is differenct from the value given in the SAS output for the standard deviation of Geary's ratio c. We assume random sampling without replacement when deriving the formula for the variance of c, but in the SAS Variogram procedure, the formula is derived assuming that the data come from a normal distribution.

The formula used in the SAS Variogram procedure to calculate the variance of the Geary's Ratio, c, is as follows:

$$\mathrm{Var}_N(c)=\frac{(2S_1+S_2)(n-1)-4W^2}{2(n+1)W^2}.$$

Under the normality assumption,

$$\begin{aligned}\mathrm{Var}_N(c) &= \frac{(2S_1+S_2)(n-1)-4W^2}{2(n+1)W^2}\\ &= \frac{(2\cdot 116+1048)(14-1)-4\cdot 58^2}{2\cdot(14+1)\cdot 58^2}\\ &= 0.03155.\end{aligned}$$

We get the same standard deviation, $\mathrm{SD}_N(c)=0.178$, as in the SAS output.

Now let's find the observed value of Moran's I as discussed in Section 2.7:

$$I=\frac{2n}{W}\left[\frac{N_{rr}}{n_r}+\frac{N_{ss}}{n_s}\right]-1=\frac{2\cdot 14}{58}\left[\frac{13}{8}+\frac{5}{6}\right]-1=0.187.$$

The expected value of Moran's I (under the assumption of sampling without replacement with fixed samples sizes n_r, n_s and n) is equal to

$$\begin{aligned}
\mathbb{E}(I) &= \mathbb{E}\left[\frac{2n}{W}\left(\frac{N_{rr}}{n_r}+\frac{N_{ss}}{n_s}\right)\right]-1 \\
&= \frac{2n}{Wn_r}\mathbb{E}(N_{rr})+\frac{2n}{Wn_s}\mathbb{E}(N_{ss})-1 \\
&= \frac{2n}{Wn_r}\cdot\frac{Wn_r^{(2)}}{2n^{(2)}}+\frac{2n}{Wn_s}\cdot\frac{Wn_s^{(2)}}{2n^{(2)}}-1 \\
&= \frac{n_r-1}{n-1}+\frac{n_s-1}{n-1}-1.
\end{aligned}$$

It follows that

$$\mathbb{E}(I)=\frac{n_r+n_s-2}{n-1}-1=-\frac{1}{n-1}. \tag{2.31}$$

Therefore the expected value of the Moran's I for this example is,

$$\mathbb{E}(I)=-1/13=-0.07692.$$

The variance of Moran's I in SAS is also calculated by assuming normality. Below we show this calculation by assuming the data come from a normal distribution (even though in our example, this is obviously not true). First we will find the second moment of I:

$$\begin{aligned}
\mathbb{E}_N(I^2) &= \frac{n^2S_1-nS_2+3W^2}{W^2(n^2-1)} \\
&= \frac{14^2\cdot 116-14\cdot 1048+3\cdot 58^2}{58^2(14^2-1)} \\
&= 0.02768.
\end{aligned}$$

In addition, $\mathbb{E}_N(I)=\mathbb{E}(I)=-0.07692$. Therefore,

$$\begin{aligned}
\mathrm{Var}_N(I) &= \mathbb{E}_N[I^2]-[\mathbb{E}_N(I)]^2 \\
&= 0.02768-(-0.07692)^2 \\
&= 0.02176.
\end{aligned}$$

Then the standard deviation of Moran's I under normality, $\mathrm{SD}_N(I)$, is

found as 0.148.

2.11 Implementation of the formulas of the chapter in R

In this section, we show how one can implement some of the formulas of the chapter in the statistical programming language R. Some of these formulas can also be found in Sokal and Oden (1978).

The results obtained in the previous two sections using R and SAS (without any complicated programming) are obtained again using the R program below. In other words, we obtain the same results for Example 4 that were obtained in Sections 2.9 and 2.10, but we now use the formulas obtained in the chapter *directly* rather than through the use of a package or a procedure. The R code is shown below.

```
buffalo <- function()
{
w <- matrix(c(0,1,0,1,1,0,0,1,0,0,0,0,0,0,1,0,0,1,1,1,0,0,0,1,0,0,0,
0,0,0,0,1,0,0,0,1,0,0,1,0,0,1,1,1,1,0,0,0,0,1,0,1,0,0,1,1,1,1,0,0,0,
1,1,0,0,0,0,1,0,0,0,1,0,0,1,0,1,0,1,1,0,0,0,0,0,0,0,0,1,1,0,0,1,0,0,
1,0,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,1,0,0,1,0,0,1,0,
1,0,1,0,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,1,0,0,
0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,1,0,0,0,1,0,0,1,1,0,0,0,0,0,0,0,1,0,1,0),
nrow=14,ncol=14)
rs <- matrix(c(0,0,0,0,1,0,1,0,1,0,1,1,0,1,0,0,0,0,1,0,1,0,1,0,1,1,
0,1,0,0,0,0,1,0,1,0,1,0,1,1,0,1,0,0,0,0,0,1,0,1,0,1,0,1,1,1,0,1,1,1,1,
1,0,1,0,1,0,1,0,0,1,0,0,0,0,0,0,1,0,1,0,1,0,1,1,1,1,1,1,1,0,1,0,1,
0,1,0,0,1,0,0,0,0,0,0,1,0,1,0,1,0,1,1,0,1,1,1,1,1,1,0,1,0,1,0,1,0,0,1,
0,0,0,0,0,0,1,0,1,0,1,0,1,1,0,1,1,1,1,1,1,0,1,0,1,0,1,0,0,1,0,1,1,1,1,
0,1,0,1,0,1,0,0,1,0,0,0,0,0,1,0,1,0,1,0,1,1,0,1,1,1,1,1,1,0,1,0,1,0,
1,0,0,1,0),nrow=14,ncol=14)
rr <- matrix(c(1,1,1,1,0,1,0,1,0,1,0,0,1,0,1,1,1,1,0,1,0,1,0,1,0,1,0,
0,1,0,1,1,1,1,0,1,0,1,0,1,0,0,1,0,1,1,1,1,0,1,0,1,0,1,0,0,1,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,1,0,1,0,1,0,0,1,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,1,0,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,1,1,1,1,1,0,1,0,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,1,0,1,0,1,0,0,1,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0),nrow=14,ncol=14)
ss <- matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,1,0,1,0,1,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
0,1,0,1,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,1,
0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,1,0,1,1,0,
1,0,0,0,0,1,0,1,0,1,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,0,1,0,1,0,1,1,0,1),nrow=14,ncol=14)
```

```
wsum <- sum(w)
n <- 14
nr <- 8
ns <- n-nr
Nrr <- 0
Nss <- 0
Nrs <- 0
for (a in 1:14)
  {
     for (b in 1:14)
     {
         Nrr <- Nrr+((1/2)*w[a,b]*rr[a,b])
         Nss <- Nss+((1/2)*w[a,b]*ss[a,b])
         Nrs <- Nrs+((1/2)*w[a,b]*rs[a,b])
      }
   }
# Moran's I
MoranI <- ((2*n/wsum)*((Nrr/nr)+(Nss/ns))-1)
s1 <- (1/2)*sum((w+t(w))*(w+t(w)))
s2 <- sum((rowSums(w)+colSums(w))^2)
#Expected values for join counts
ENrr <- wsum*nr*(nr-1)/2/n/(n-1)
ENss <- wsum*ns*(ns-1)/2/n/(n-1)
ENrs <- wsum*nr*ns/n/(n-1)
#Variances for join counts
A1 <- s1*nr*(nr-1)/n/(n-1)
B1 <- (s2-(2*s1))*nr*(nr-1)*(nr-2)/n/(n-1)/(n-2)
C1 <- ((wsum^2)+s1-s2)*nr*(nr-1)*(nr-2)*(nr-3)/n/(n-1)/(n-2)/(n-3)
D1 <- (wsum*nr*(nr-1)/n/(n-1))^2
VNrr <- (1/4)*(A1+B1+C1-D1)
A2 <- s1*ns*(ns-1)/n/(n-1)
B2 <- (s2-(2*s1))*ns*(ns-1)*(ns-2)/n/(n-1)/(n-2)
C2 <- ((wsum^2)+s1-s2)*ns*(ns-1)*(ns-2)*(ns-3)/n/(n-1)/(n-2)/(n-3)
D2 <- (wsum*ns*(ns-1)/n/(n-1))^2
VNss <- (1/4)*(A2+B2+C2-D2)
A3 <- 2*s1*nr*ns/n/(n-1)
B3 <- (s2-(2*s1))*nr*ns*(nr+ns-2)/n/(n-1)/(n-2)
C3 <- 4*((wsum^2)+s1-s2)*nr*(nr-1)*ns*(ns-1)/n/(n-1)/(n-2)/(n-3)
D3 <- 4*((wsum*nr*ns/n/(n-1))^2)
VNrs <- (1/4)*(A3+B3+C3-D3)
# Z-scores
zrr <- (Nrr-ENrr)/sqrt(VNrr)
zss <- (Nss-ENss)/sqrt(VNss)
zrs <- (Nrs-ENrs)/sqrt(VNrs)
list(wsum,rr,ss,rs,"Number of rr joins"=Nrr,
"Number of ss joins" =Nss,
"Number of rs joins"=Nrs,
"Expected values of rr join counts" = ENrr,
"Expected values of ss join counts" = ENss,
"Expected values of rs join counts" = ENrs,
"Variance of rr join counts" = VNrr,
"Variance of ss join counts" = VNss,
```

```
"Variance of rs join counts" = VNrs,
"Z-score for rr join counts" = zrr,
"Z-score for ss join counts" = zss,
"Z-score for rs join counts" = zrs,
"Moran's I"=MoranI)
}
```

Below we have the output from the above R program. It is observed that the results obtained are the same as the results obtained by using the R package *spdep* and the SAS procedure *variogram*.

```
---------------------------OUTPUT-------------------------------------
$`Number of rr joins`
[1] 13
$`Number of ss joins`
[1] 5
$`Number of rs joins`
[1] 11
$`Expected values of rr join counts`
[1] 8.923077
$`Expected values of ss join counts`
[1] 4.78022
$`Expected values of rs join counts`
[1] 15.29670
$`Variance of rr join counts`
[1] 3.203873
$`Variance of ss join counts`
[1] 2.249399
$`Variance of rs join counts`
[1] 5.179699
$`Z-score for rr join counts`
[1] 2.277691
$`Z-score for ss join counts`
[1] 0.1465397
$`Z-score for rs join counts`
[1] -1.887918
$`Moran's I`
[1] 0.1867816
```

CHAPTER III
SPATIAL AUTOCORRELATION FOR NUMERICAL VARIABLES

In this chapter we discuss probabilistic properties of a numerical variable measured at different locations and we derive its first two moments under randomization. We also derive the covariance of the variable with itself at two different locations. Then we discuss two famous measures of overall spatial autocorrelation for numerical data, one due to Moran (1950) and one due to Geary (1954). We derive their expected values and state their variances according to Cliff and Ord (1981). In addition we discuss two different ways of assigning weights to pairs of localities (for which the coordinates are known), one using the SAS Variogram procedure and one using Thiessen polygons. In addition, we expain how to draw Theissen polygons or Voronoi diagrams using MATLAB. Finally we discuss how to calculate two measures of spatial autocorrelation and their moments using the statistical package R.

3.1 Introduction

We assume we have n distinct regions (e.g., squares or localities) and in each locality i we observe the value x_i of a numerical[1] variable of interest X. We assume that the data $x_1, \ldots, x_n$ are such that n_1 of them are equal to y_1, n_2 of them are equal to y_2, and so on, and finally n_m of them are equal to y_m (where $1 \leq m < n$). Before observing the data, assume that for each i, X_i is the random variable that generated observation x_i. In order to find the randomization distribution, we assume that the sample space of the experiment, Ω, consists of all permutations of the list $(x_1, x_2, \ldots, x_n)$. In other words, Ω consists of all permutations with

[1] In the previous chapter, we worked with categorical data. In the case the categorical variable was a binary nominal one, we coded the two categories by 0 and 1, and then applied the theory of this chapter. For more details, see Sections 2.7 and 2.8 (and parts of Section 2.9).

repetitions of the numbers $y_1, \ldots, y_m$, where y_i is repeated exactly n_i times for $i = 1, \ldots, m$. We then have

$$\#\Omega = \frac{n!}{n_1! n_2! \cdots n_m!}.$$

See also Chapter 3 in Griffith (1987).

Note that

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{p=1}^{m} n_p y_p}{n}$$

and

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2 = \sum_{p=1}^{m} \frac{n_p y_p^2}{n} - \frac{\left(\sum_{p=1}^{m} n_p y_p\right)^2}{n}$$

are constants. The same of course holds for the observed values: $\overline{x} = \frac{\sum_{p=1}^{m} n_p y_p}{n}$ and

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{p=1}^{m} \frac{n_p y_p^2}{n} - \frac{\left(\sum_{p=1}^{m} n_p y_p\right)^2}{n}.$$

Example 5. In order to study the annual average rainfall in a big city, five rainfall stations were selected. Observing the distances among the rainfall stations and considering other factors, an experimenter observes three different averages in the five rainfall stations. Let $x_1, x_2, ..., x_5$ be the average values of annual rainfall in the five different stations in a particular year. Let y_1, y_2, y_3 be the three different rainfall averages, where $n_1 = 1$ of the stations have an average rainfall of y_1, $n_2 = 2$ of the stations have an average rainfall of y_2 and $n_3 = 2$ of the stations have an average rainfall of y_3 in a way that $n_1 + n_2 + n_3 = 1 + 2 + 2 = 5$. In this example, $n = 5$, $m = 3$, and

$$\{x_1, \ldots, x_5\} = \{y_1, y_2, y_3\}.$$

Table 3.1. The sample space Ω of permutations of the five observed average rainfalls for Example 5 (with $n_1 = 1$, $n_2 = 2$ and $n_3 = 2$)

Rainfall station				
1	2	3	4	5
y_1	y_2	y_2	y_3	y_3
y_1	y_2	y_3	y_2	y_3
y_1	y_2	y_3	y_3	y_2
y_1	y_3	y_3	y_2	y_2
y_1	y_3	y_2	y_3	y_2
y_1	y_3	y_2	y_2	y_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_3	y_2	y_3	y_2	y_1
y_3	y_2	y_2	y_3	y_1
y_3	y_3	y_2	y_2	y_1

There are $\frac{5!}{1!2!2!} = 30$ ways that the observed values $x_1, \ldots, x_5$ could have occurred among the five locations; see Table 3.1. □

A probability model on the sample space Ω is any function $\mathbb{P} : \Omega \to [0, 1]$ such that

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1.$$

Under the null hypothesis, we may assume that all outcomes in Ω are equally likely:

$$H_0 : \mathbb{P}(\omega) = \frac{1}{\#\Omega} \quad \text{for all } \omega = (\omega_1, \ldots, \omega_n) \in \Omega. \tag{3.1}$$

Note that for any $\omega \in \Omega$, $X_i(\omega) = \omega_i$. Thus, under H_0,

$$\begin{aligned} \mathbb{P}[X_i = y_p] &= \frac{\#\{\omega \in \Omega : X_i(\omega) = y_p\}}{\#\Omega} \\ &= \frac{\frac{(n-1)!}{n_1! \cdots (n_p - 1)! \cdots n_m!}}{\frac{n!}{n_1! \cdots n_m!}} = \frac{n_p}{n}. \end{aligned} \tag{3.2}$$

Also, for $i \neq j$ and $p \neq q$,

$$\begin{aligned}\mathbb{P}[X_i = y_p, X_j = y_q] &= \frac{\frac{(n-2)!}{n_1!\cdots(n_p-1)!\cdots(n_q-1)!\cdots n_m!}}{\frac{n!}{n_1!\cdots n_m!}} \\ &= \frac{n_p n_q}{n(n-1)}. \end{aligned} \tag{3.3}$$

For $i \neq j$ and $p = q$,

$$\begin{aligned}\mathbb{P}[X_i = y_p, X_j = y_q] &= \frac{\frac{(n-2)!}{n_1!\cdots(n_p-2)!\cdots n_m!}}{\frac{n!}{n_1!\cdots n_m!}} \\ &= \frac{n_p(n_p-1)}{n(n-1)}. \end{aligned} \tag{3.4}$$

The above calculated probabilities are needed for the calculation of various moments (joint or marginal) of the X_i's, which in turn are needed for the calculation of some moments of the two measures of autocorrelation we are discussing in this chapter (Moran's I and Geary's ration c).

3.2 Moments of each X_i under randomization

Recall that, for each $i \in [1..n]$, X_i is the random variable of interest in location i that generates the observed value x_i. The marginal distribution[2] of X_i (under the H_0 in (3.1)) was derived in equation (3.2). The expected value of X_i (under the H_0) is

$$\mathbb{E}(X_i) = \sum_{p=1}^{m} y_p \mathbb{P}(X_i = y_p) = \sum_{p=1}^{m} \frac{n_p y_p}{n} = \overline{X} = \overline{x}.$$

The second moment of X_i (again under randomization) is

$$\mathbb{E}(X_i^2) = \sum_{p=1}^{m} y_p^2 \mathbb{P}(X_i = y_p) = \sum_{p=1}^{m} \frac{n_p y_p^2}{n} = \overline{X^2} = \overline{x^2}.$$

(Here $\overline{X^2}$ is the average of the $X_1^2, \ldots, X_n^2$.) Therefore the variance of X_i is

$$\mathrm{Var}(X_i) = \mathbb{E}(X_i^2) - [\mathbb{E}(X_i)]^2$$

[2]Strictly speaking we are finding the conditional marginal distribution of X_i given the set of data $\{x_1, \ldots, x_n\}$. We are conditioning on the data themselves, but not on the order in which they appear.

$$
\begin{aligned}
&= \sum_{p=1}^{m} \frac{n_p y_p^2}{n} - \left(\sum_{p=1}^{m} \frac{n_p y_p}{n} \right)^2 \\
&= \overline{X^2} - (\overline{X})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.
\end{aligned}
$$

The joint randomization distribution of X_i and X_j when $i \neq j$ was derived in equations (3.3) and (3.4). Using this distribution, we will find the covariance of X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j).$$

Now

$$
\begin{aligned}
\mathbb{E}(X_i X_j) &= \sum_{p=1}^{m} \sum_{q=1}^{m} y_p y_q \mathbb{P}[X_i = y_p, X_j = y_q] \\
&= \sum_{p=1}^{m} y_p^2 \mathbb{P}[X_i = y_p, X_j = y_p] + \sum_{p=1}^{m} \sum_{\substack{q=1 \\ q \neq p}}^{m} y_p y_q \mathbb{P}[X_i = y_p, X_j = y_q] \\
&= \sum_{p=1}^{m} y_p^2 \frac{n_p(n_p - 1)}{n(n-1)} + \sum_{p=1}^{m} \sum_{\substack{q=1 \\ q \neq p}}^{m} y_p y_q \frac{n_p n_q}{n(n-1)} \\
&= \frac{1}{n(n-1)} \left[\sum_{p=1}^{m} y_p^2 n_p^2 - \sum_{p=1}^{m} n_p y_p^2 + \sum_{p=1}^{m} \sum_{\substack{q=1 \\ q \neq p}}^{m} n_p y_p n_q y_q \right] \\
&= \frac{1}{n(n-1)} \left[\left(\sum_{p=1}^{m} y_p n_p \right)^2 - \sum_{p=1}^{m} n_p y_q^2 \right] \\
&= \frac{1}{n(n-1)} \left[\left(\sum_{i=1}^{n} X_i \right)^2 - \sum_{i=1}^{n} X_i^2 \right].
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) \\
&= \frac{1}{n(n-1)} \left[\left(n\overline{X} \right)^2 - \sum_{i=1}^{n} X_i^2 \right] - \overline{X}^2
\end{aligned}
$$

$$= \frac{1}{n(n-1)}\left[\left(\overline{X}\right)^2\left(n^2-n^2+n\right)-\sum_{i=1}^{n}X_i^2\right]$$

$$= \frac{-1}{n(n-1)}\left[\sum_{i=1}^{n}X_i^2-n\left(\overline{X}\right)^2\right]=\frac{-\sum_{i=1}^{n}(X_i-\overline{X})^2}{n(n-1)}.$$

The formulas derived in this section will be useful for the calculation of the first two moments of Moran's I and Geary's c under the randomization assumption. See Sections 3.4 and 3.6.

3.3 Discussion of Moran's I autocorrelation coefficient

Recall that we assume we have n distinct regions (e.g., squares or localities) and in each locality i we observe the value x_i of a numerical variable of interest X. As in Section 2.2, we let w_{ij} be the weight given to join or edge between localities i and j (see Sokal and Oden (1978, p. 226)). It is not necessary for w_{ij} to equal w_{ji}.

We always assume that $w_{ii}=0$ for all $i \in [1..n]$ because there is *no join* of a region with itself. (As in Section 2.2, we define $[1..n] := \{m \in \mathbb{N} |\ 1 \leq m \leq n\}$.) We denote the $n \times n$ matrix of weights by M_W, i.e.,

$$M_W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{pmatrix}$$

and the sum of all weights by W, i.e.,

$$W = \sum_{i \neq j} w_{ij}, \tag{3.5}$$

where the summation is over all pairs $(i,j) \in [1..n]^2$ with $i \neq j$. As stated above, we will ignore the diagonal elements of M_W.

Since the observed data are $x_1, x_2, \ldots, x_n$, then Moran's (1950) measure of spatial

autocorrelation coefficient, I, is given by

$$I = \frac{n \sum_{i \neq j} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{W \sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

(See also Section 2.7.)

Usually, Moran's I is between -1 and 1 (but unfortunately there are exceptions to this rule - see equations (2.25) and (2.26) in Section 2.7). If most pairs of neighbouring localities have values of X both above the average or both below the average, then Moran's I tends to be positive. If on the other hand, for most pairs of neighbouring localities, one has a value of X above the average and the other one has a value below the average, then Moran's I tends to be negative.

As discussed in Example 2 in Section 2.6, if the squares are connected by rook's move, this means that each square is connected to all squares that are directly below, above, to the left side or to the right side of it. One can construct a graph that illustrates this relationship (connection), and this also shown in Fig. 2.1, where the graph consists of 9 vertices and 12 edges. Each vertex of the graph corresponds to a square, and two vertices are connected by an edge if and only if the corresponding squares are considered adjacent (by rook's moves). For that example, the adjacency matrix M_W is given in Table 2.1.

3.4 Moments of Moran's I

As we discussed in Chapter II, Moran's I is a measure of spatial autocorrelation. In this section, we calculate the expected value and the variance of Moran's I under the null hypothesis (3.1). In other words, we calculate the randomization mean and variance of I (conditional on the values of the observed data, but not on the order

in which they appear). We have

$$\mathbb{E}(I) = \mathbb{E}\left(\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right]\left[\sum_{l=1}^{n}(X_l - \overline{X})^2\right]}\right).$$

As we stated in Section 3.1, under randomization, $\overline{X}$ and $\sum_{l=1}^{n}(X_l - \overline{X})^2$ are constants. Using the fact that $W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, we have

$$\begin{aligned}
\mathbb{E}(I) &= \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathbb{E}\left[\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)\right]}{W\sum_{l=1}^{n}\left(X_l - \overline{X}\right)^2} \\
&= \frac{2n\sum_{i=1}^{n}\sum_{j=1}^{i} w_{ij}\mathbb{E}\left[(X_i - \overline{X})(X_j - \overline{X})\right]}{W\sum_{l=1}^{n}(X_l - \overline{X})^2}.
\end{aligned}$$

Now let's evaluate $\mathbb{E}\left[(X_i - \overline{X})(X_j - \overline{X})\right]$. We have

$$\begin{aligned}
\mathbb{E}\left[(X_i - \overline{X})(X_j - \overline{X})\right] &= \sum_{p=1}^{m}\sum_{q=1}^{m}(y_p - \overline{x})(y_q - \overline{x})\mathbb{P}[X_i = y_p, X_j = y_q] \\
&= \sum_{p=1}^{m}(y_p - \overline{x})^2\frac{n_p(n_p-1)}{n(n-1)} + \sum_{p=1}^{m}\sum_{\substack{q=1\\q\neq p}}^{m}(y_p - \overline{x})(y_q - \overline{x})\frac{n_p n_q}{n(n-1)} \\
&= \frac{1}{n(n-1)}\left[\sum_{p=1}^{m} n_p^2(y_p - \overline{x})^2 - \sum_{p=1}^{m} n_p(y_p - \overline{x})^2\right. \\
&\quad \left. + \sum_{p=1}^{m}\sum_{\substack{q=1\\q\neq p}}^{m} n_p n_q(y_p - \overline{x})(y_q - \overline{x})\right] \\
&= \frac{1}{n(n-1)}\left[\left(\sum_{p=1}^{m} n_p(y_p - \overline{x})\right)^2 - \sum_{p=1}^{m} n_p(y_p - \overline{x})^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n(n-1)}\left[\left(\sum_{i=1}^{n}(x_i-\overline{x})\right)^2-\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\\
&= \frac{1}{n(n-1)}\left[\left(0-\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\right]\\
&= \frac{-\sum_{i=1}^{n}(x_i-\overline{x})^2}{n(n-1)}.
\end{aligned}
$$

Note that since we condition on the values of the observed data (but not on the order in which they appear), we have $\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}(x_i-\overline{x})^2$. We thus have

$$
\begin{aligned}
\mathbb{E}(I) &= \frac{2n\sum_{i=1}^{n}\sum_{j=1}^{i}w_{ij}\left[\frac{-\sum_{l=1}^{n}(x_l-\overline{x})^2}{n(n-1)}\right]}{W\sum_{l=1}^{n}(x_l-\overline{x})^2}\\
&= \frac{-2n\left[\sum_{l=1}^{n}(x_l-\overline{x})^2\right]\sum_{i=1}^{n}\sum_{j=1}^{i}w_{ij}}{n(n-1)\left[\sum_{l=1}^{n}(x_l-\overline{x})^2\right]W}\\
&= \frac{-2n\left[\sum_{l=1}^{n}(x_l-\overline{x})^2\right](W/2)}{n(n-1)\left[\sum_{l=1}^{n}(x_l-\overline{x})^2\right]W}\\
&= -\frac{1}{n-1}.
\end{aligned}
$$

This formula[3] can be found, for example, in Section 1.5.1 of Cliff and Ord (1981). The idea of the proof was taken from Appendix E in Griffith (1987).

[3]For the special case X is a binary variable taking the values 0 and 1 (that code the two categories of a binary categorical variable), we have proven this formula in Section 2.10. See equation (2.31).

According to Cliff and Ord (1981, Section 1.5.1), the variance of the Moran's I is as follows:

$$\text{Var}(I) = \mathbb{E}(I^2) - [\mathbb{E}(I)]^2,$$

where

$$\mathbb{E}(I^2) = \frac{n[(n^2-3n+3)S_1 - nS_2 + 3W^2] - b_2[(n^2-n)S_1 - 2nS_2 + 6W^2]}{(n-1)^{(3)}W^2}$$

and $\mathbb{E}(I)$ is as given above. In addition, S_1, S_2 and W remain the same as we discussed in Chapter II (see equations (2.10)). Finally, b_2 is the sample coefficient of kurtosis, m_4/m_2^2, where m_j is the jth central moment around the mean:

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^j.$$

3.5 Discussion of Geary's Ratio c

The other measure of spatial autocorrelation we discuss here is Geary's ratio c introduced in 1954. According to Griffith (1987, Chapter 5), it is a paired comparison of juxtaposed map values (that is, values of a variable at localities placed next to each other). The weights, which indicate the connections among the pairs of localities, can be chosen in a way similar to the one discussed in Section 3.3. Since the observed data are $x_1, x_2, \ldots, x_n$, Geary's (1954) measure of spatial autocorrelation coefficient is given by

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} w_{ij}(x_i - x_j)^2}{2W\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Although Moran's I usually takes values between -1 and 1, Geary's ratio c can take positive values only. Geary's ratio is close to zero if many values of the variable X are close to each other for neighbouring locations (i.e., if $x_i = x_j$ for most of the

neighbouring locations i and j for which $w_{ij} > 0$). The maximum value that c can have is around 2, which occurs when x_i and x_j are different for most of the neighbouring locations i and j (for which the weight $w_{ij} > 0$). As proven in the next section, the average value of c both under normality and radomization sampling is equal to 1.

3.6 Moments of Geary's ratio c

Geary's ratio c is the other measure of spatial autocorrelation we discussed in Chapter II. In this section we derive the moments of Geary's ratio for a continuous random variable X under the assumption of randomization. These formulas can be found, for example, in Section 1.5.1 of Cliff and Ord (1981). First we find the mean of c. We have

$$\begin{aligned}\mathbb{E}(c) &= \mathbb{E}\left[\frac{(n-1)\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} w_{ij}(X_i - X_j)^2}{2W\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] \\ &= \frac{(n-1)\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} w_{ij}\mathbb{E}[(X_i - X_j)^2]}{2W\sum_{i=1}^{n}(X_i - \overline{X})^2}.\end{aligned}$$

Now let's evaluate $\mathbb{E}[(X_i - X_j)^2]$. We have

$$\begin{aligned}\mathbb{E}[(X_i - X_j)^2] &= \mathbb{E}(X_i^2) - 2\mathbb{E}(X_iX_j) + \mathbb{E}(X_j^2) \\ &= \sum_{p=1}^{m}\frac{n_p y_p^2}{n} - 2\left[\sum_{p=1}^{m}\frac{y_p^2 n_p(n_p-1)}{n(n-1)} + \sum_{p=1}^{m}\sum_{\substack{q=1\\ q\neq p}}^{m}\frac{n_p n_q y_p y_q}{n(n-1)}\right] + \sum_{q=1}^{m}\frac{n_q y_q^2}{n} \\ &= 2\sum_{p=1}^{m}\frac{n_p y_p^2}{n} - 2\left[\sum_{p=1}^{m}\frac{n_p^2 y_p^2}{n(n-1)} - \sum_{p=1}^{m}\frac{n_p y_p^2}{n(n-1)} + \sum_{p=1}^{m}\sum_{\substack{q=1\\ q\neq p}}^{m}\frac{n_p n_q y_p y_q}{n(n-1)}\right]\end{aligned}$$

$$
\begin{aligned}
&= 2\sum_{p=1}^{m}\frac{n_p y_p^2}{n} - 2\left[\frac{\left(\sum_{p=1}^{m} n_p y_p\right)^2}{n(n-1)} - \sum_{p=1}^{m}\frac{n_p y_p^2}{n(n-1)}\right] \\
&= 2\left[\frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{n\overline{x}^2}{n-1} + \frac{\sum_{i=1}^{n} x_i^2}{n(n-1)}\right] \\
&= 2\frac{\left[n\sum_{i=1}^{n} x_i^2 - n^2\overline{x}^2\right]}{n(n-1)} \\
&= \frac{2\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}{n-1} \\
&= \frac{2\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}.
\end{aligned}
$$

Therefore the randomization mean[4] of Geary's ratio c is

$$\mathbb{E}(c) = \frac{(n-1)W}{2W\sum_{i=1}^{n}(x_i-\overline{x})^2} \cdot \frac{2\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} = 1.$$

The idea of the proof was taken from Appendix E in Griffith (1987).

According to Cliff and Ord (1981), the variance of Geary's ratio c under randomization is given by

$$
\begin{aligned}
\mathrm{Var}(c) &= \frac{(n-1)S_1[n^2-3n+3-(n-1)b_2]}{n(n-2)^{(2)}W^2} \\
&\quad - \frac{(n-1)S_2[n^2+3n-6-(n^2-n+2)b_2]}{4n(n-2)^{(2)}W^2} \\
&\quad + \frac{[n^2-3-(n-1)^2 b_2]}{n(n-2)^{(2)}},
\end{aligned}
$$

[4]For the special case X is a binary variable taking the values 0 and 1 (that code the two categories of a binary categorical variable), we have proven this formula in Section 2.8. See equation (2.30).

where b_2 is the sample coefficient of kurtosis, m_4/m_2^2. Here m_2 and m_4 are 2nd and 4th central moments (around the sample mean). The ith central moment defined as follows:

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^j.$$

3.7 Calculation of weights when the exact locations of the units are known

Suppose that the exact locations of the n units in the sample are (a_i, b_i), $i = 1, \ldots, n$, and let x_i be the observed value of the variable of interest for unit i. In this case, it is reasonable to assign weights between two points (units i and j where $i \neq j$) that are inversely proportional with the distance between the two points (see Section 5.5.2.3 of de Smith *et al.* (2007)):

$$w_{ij} = \frac{1}{d_{ij}} \quad \text{or} \quad w_{ij} = \frac{\epsilon_1}{\epsilon_2 + d_{ij}} \quad \text{(when } d_{ij} \text{ is relatively small)},$$

where ϵ_1 and ϵ_2 are small positive numbers and d_{ij} is the distance between units i and j:

$$d_{ij} = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2}.$$

Table 3.2. (a_i, b_i) coordinates and x_i values for 5 rainfall stations

Station	1	2	3	4	5
a_i	2	3	3	4	8
b_i	4	1	5	2	3
x_i	5.5	2.2	5.3	2.3	7.2

Table 3.2 gives an imaginary example with the coordinates of five rainfall stations (in some rectangular coordinate system) and the corresponding average amount of rain per year. Using the SAS program below, we can calculate the two measures of spatial autocorrelation, Moran's I and Geary's ratio c for the above data. The

program also calculates their expected values, their variances and the corresponding p-values under the randomization assumption.

```
Data rain;
input Station a b rain;
cards;
1 2 4 5.5
2 3 1 2.2
3 3 5 5.3
4 4 2 2.3
5 8 3 7.2
;
proc variogram data=rain outacweights=ow;
  compute  novar autoc (assumption=ran weights=distance);
  coordinates xc=a yc=b;
  var rain;
run;
proc print data=ow;
run;
```

By default SAS uses the following formula for the weights between rainfall stations i and j:

$$w_{ij} = \frac{1}{1 + d_{ij}}.$$

SAS gives the following output:

The VARIOGRAM Procedure
Dependent Variable: rain

Number of Observations Read	5
Number of Observations Used	5

Pairs Information

Number of Lags	11
Lag Distance	0.72
Maximum Data Distance in a	6.00
Maximum Data Distance in b	4.00
Maximum Data Distance	7.21

Pairwise Distance Intervals

Lag Class	Bounds		Number of Pairs	Percentage of Pairs
0	0.00	0.36	0	0.00%
1	0.36	1.08	0	0.00%
2	1.08	1.80	2	20.00%
3	1.80	2.52	0	0.00%
4	2.52	3.24	3	30.00%
5	3.24	3.97	0	0.00%
6	3.97	4.69	2	20.00%
7	4.69	5.41	2	20.00%
8	5.41	6.13	1	10.00%
9	6.13	6.85	0	0.00%
10	6.85	7.57	0	0.00%

Autocorrelation Statistics

Assumption	Coefficient	Observed	Expected	Std Dev	Z	Pr > |Z|
Randomization	Moran's I	-0.0979	-0.25	0.112	1.35	0.1764
Randomization	Geary's c	0.8227	1.00	0.114	-1.56	0.1195

In addition, it prints the distances between every two stations and their corresponding weights:

Obs	VARNAME	X1	Y1	X2	Y2	V1	V2	DISTANCE	ACWGHT12	ID1	ID2
1	rain	2	4	3	1	5.5	2.2	3.16228	0.24025	1	2
2	rain	2	4	3	5	5.5	5.3	1.41421	0.41421	1	3
3	rain	2	4	4	2	5.5	2.3	2.82843	0.26120	1	4
4	rain	2	4	8	3	5.5	7.2	6.08276	0.14119	1	5
5	rain	3	1	3	5	2.2	5.3	4.00000	0.20000	2	3
6	rain	3	1	4	2	2.2	2.3	1.41421	0.41421	2	4
7	rain	3	1	8	3	2.2	7.2	5.38516	0.15661	2	5
8	rain	3	5	4	2	5.3	2.3	3.16228	0.24025	3	4
9	rain	3	5	8	3	5.3	7.2	5.38516	0.15661	3	5
10	rain	4	2	8	3	2.3	7.2	4.12311	0.19519	4	5

SAS allows the user to change the value of ϵ_1 in the formula $w_{ij} = \frac{\epsilon_1}{\epsilon_2 + d_{ij}}$, say from 1 to 2, by changing the following line in the above SAS code:

```
compute  novar autoc (assumption=ran weights=distance (scale=2));
```

The value of ϵ_2, however, must be kept equal to 1.

3.8 Thiessen polygons and Voronoi diagrams

Suppose again that the exact locations of the n units in the sample are (a_i, b_i), $i = 1, \ldots, n$, and let x_i be the observed value of the variable of interest for unit i. The creation of *Thiessen polygons*[5] (also known as *Voronoi diagrams*[6]) around these n points is a method for creating neighbouring localities (polygons) so that each locality contains exactly one unit. This allows us to assign binary weights (that is, points whose corresponding polygons share a common boundary get a weight of 1, and 0 otherwise). According to Griffith (1987), these polygons are constructed so that each point on the boundary of the polygon that contains unit i has a distance

[5]These polygons are named after the meteorologist Alfred H. Thiessen.

[6]These diagrams are named after the mathematician Georgy Feodosevich Voronoy.

to point (a_i, b_i) *less than or equal* to the distance to each other point (a_j, b_j) (where $j \neq i$).

To construct Thiessen polygons around n points, we connect the points so that we create a triangular lattice. (These triangles are non-overlapping and adjacent, and this triangulation is known as *Delaunay triangulation*[7].) Then we draw perpendicular bisectors to each edge of each triangle. These bisectors create the so-called Thiessen polygons, one around each point.

We can produce Voronoi diagrams in Matlab for the data in Table 3.2 using the following set of commands:

```
>> a=[2,3,3,4,8];
>> b=[4,1,5,2,3];
>> voronoi(a,b)
```

The output (edited for clarity) is shown in Figure 3.1.

Based on the graph in Figure 3.1, the weight matrix for the above example is given below:

$$M_W = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

We wrote the following program in the statistical package R to calculate the observed values of Moran's I and Geary's ratio c and their first two moments under the randomization assumption (nonfree sampling).

```
rainfall <- function()
{
w <- matrix(c(0,1,1,1,0,1,0,0,1,0,1,0,0,1,1,1,1,1,0,1,0,0,1,1,0),
     nrow=5,ncol=5)
```

[7]This triangulation is named after the mathematician Boris Nikolaevich Delaunay.

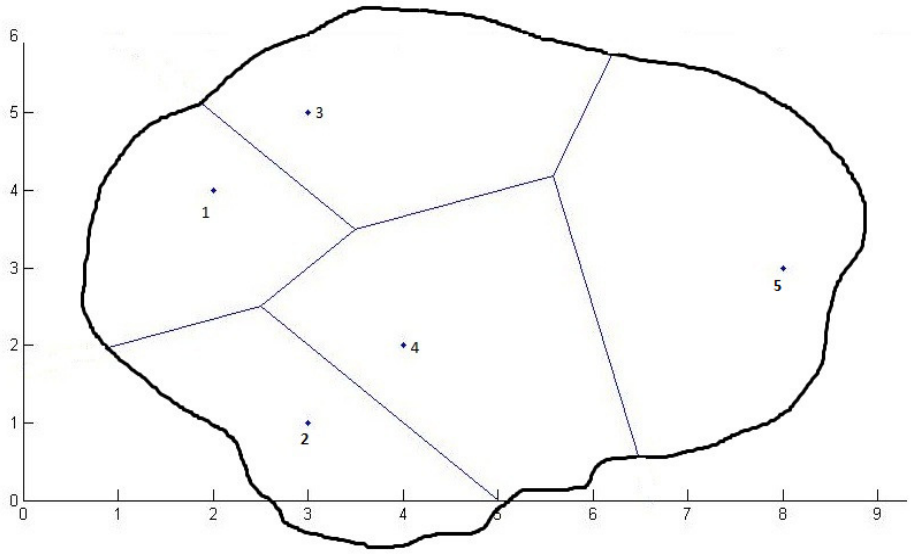

Figure 3.1. Thiessen polygons for the data in Table 3.2.

```
x <- matrix(c(5.5,2.2,5.3,2.3,7.2),nrow=5,ncol=1)
wsum <- sum(w)
n <- 5
xbar <- sum(x)/n

# Moran's I
MoranI <- ((n/wsum)*sum(w*((x-xbar) %*% t(x-xbar))))/sum((x-xbar)^2)
s1 <- (1/2)*sum((w+t(w))*(w+t(w)))
s2 <- sum((rowSums(w)+colSums(w))^2)
EI <- -(1/(n-1))
m2 <- sum((x-xbar)^2)/n
m4 <- sum((x-xbar)^4)/n
b2 <- m4/m2^2
EI2 <- (n*((n^2-(3*n)+3)*s1-n*s2+(3*wsum^2))
       -b2*((n*(n-1)*s1)-(2*n*s2)
       +(6*wsum^2)))/((n-1)*(n-2)*(n-3)*wsum^2)
varI <- EI2-(EI^2)
xij <- matrix(nrow=5,ncol=5,0)
for (i in 1:n)
{
for (j in 1:n)
{
        xij[i,j]<-x[i]-x[j]
}
}
```

```
# Geary ratio c
c <- (n-1)*(sum(w*(xij^2)))/(2*wsum*sum((x-xbar)^2))
Ec <- 1
varc <- (((n-1)*s1*(n^2-(3*n)+3-((n-1)*b2)))
        -((n-1)*s2/4*(n^2+(3*n)-6-((n^2-n+2)*b2)))
        +(wsum^2*(n^2-3-(b2*(n-1)^2))))/(n*(n-2)*(n-3)*wsum^2)
list("Moran's I" = MoranI, "Expected value of Moran's I" = EI,
"Variance of Moran's I" = varI,"Geary's ratio c" = c,
"Expected value of c" = Ec, "Variance of c" = varc )
}
```

We obtained the following output:

```
$`Moran's I`
[1] -0.1566482
$`Expected value of Moran's I`
[1] -0.25
$`Variance of Moran's I`
[1] 0.03918558
$`Geary's ratio c`
[1] 0.8664368
$`Expected value of c`
[1] 1
$`Variance of c`
[1] 0.03730308
```

CHAPTER IV
CONCLUSIONS AND FUTURE RESEARCH

In this research we discussed measures of overall (global) spatial autocorrelation: join count statistics, Moran's I and Geary's ratio c. We discussed join count statistics as a measure of spatial autocorrelation ("auto-association") for nominal categorical data and Moran's I and Geary's ratio c as measures of spatial autocorrelation for numerical data. In the case of nominal data, we derived connections between join count statistics and Moran's I and between join count statistics and Geary's ratio c.

There are *local indicators of spatial association*, called LISAs (Anselin (1995)), which calculate different autocorrelation values for different places (but we have not discussed LISAs in this thesis). According to Anselin (1995), LISAs are useful in two different ways: they may be "interpreted as indicators of local pockets" of spatial nonstationarity and they may be used to "access the influence of individual locations on the magnitude" of a global autocorrelation statistic. The study of these statistics is a possible future research topic.

Another possible future research topic is the use of a spatial autoregressive model similar to the time series autoregressive model, where the spatial autocorrelation parameter replaces the time autocorrelation parameter. The spatial autoregressive model of the numerical variable X is taken from Chapter 4 in Griffith (1987) and is given by

$$x_i = \rho \sum_{j=1}^{n} w_{ij} x_j + u_i,$$

where ρ is the spatial autocorrelation parameter and u_i is the random error term.

Here

$$w_{ij} = \frac{c_{ij}}{\sum_{k=1}^{n} c_{ik}},$$

where $c_{ij} = 1$ if the locations i and j are adjacent, and 0 otherwise. Assuming that the errors u_i's are independently and identically normally distributed with mean μ_u and variance σ_u^2, the MLE method can be applied to estimate the three unknown parameters ρ, μ_u and σ_u^2 (see appendix B in Griffith (1987)).

BIBLIOGRAPHY

[1] Anselin, L. (1995), "Local indicators of spatial association–LISA," *Geographical Analysis*, Vol. 27, pp. 93-115.

[2] Cliff, A. D., and Ord, J. K., *Spatial Processes: Models and Applications*, Pion Limited, London, UK, 1981.

[3] Cressie, N. A. C., *Statistics for Spatial Data*, John Wiley and Sons, New York, 1991.

[4] Dacey, M. F. (1965), "A review of measures of contiguity for two and k-color maps." In: Berry, B. J. L., Marble, D. F., (Eds.), *Spatial Statistics: A Reader in Statistical Geography*, Prentice Hall, Englewood Cliffs, pp. 479-495.

[5] David, F. N. (1971a), "Measurement of diversity," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, California, pp. 631-648.

[6] David, F. N. (1971b), "Measurement of diversity: multiple cell contents," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, University of California Press, Berkeley, California, pp. 109-136.

[7] de Smith, M. J., Goodchild, M. F., and Longley, P. A., *Geospatial analysis: A comprehensive guide to principles, techniques and software tools*, 2nd ed., Matador, Leicester, UK, 2007.

[8] Fisher, R. A., *The Design of Experiments*, Oliver and Boyd, Edinburgh, UK, 1935.

[9] Geary, R. C. (1954), "The contiguity ratio and statistical mapping," *The Incorporated Statistician,* Vol. 5, pp. 115-145.

[10] Getis, A. (2007), "Reflections on spatial autocorrelation," *Regional Science and Urban Economics,* Vol. 37, pp. 491-496.

[11] Griffith, D. A., *Spatial Autocorrelation: A Primer*, Resource Publications in Geography, State College, PA, USA, 1987.

[12] Krishna Iyer, P. V. (1949), "The first and second moments of some probability distributions arising from points on a lattice and their application," *Biometrika*, Vol. 36, pp. 135-141.

[13] Krishna Iyer, P. V. (1950), "The theory of probability distributions of points on a lattice," *Annals of Mathematical Statistics*, Vol. 21, pp. 198-217. Correction note (1961), Vol. 32, p. 619.

[14] Moran, P. A. P. (1948), "The interpretation of statistical maps," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 10, pp. 243-251.

[15] Moran, P. A. P. (1950), "Notes on continuous stochastic phenomena," *Biometrika*, Vol. 37, pp. 17-23.

[16] Sokal, R. R., and N. L. Oden (1978), "Spatial autocorrelation in biology I: Methodology," *Biological Journal of the Linnean Society,* Vol. 10, pp. 199-228.

[17] Solomon, H., *Geometric Probability*, Society for Industrial and Applied Mathematics, California, USA, 1978.

[18] Student[1] (1907), "On the error of counting with a haemacytometer," *Biometrika*, Vol. 5, pp. 351-360.

[1]This is a well-known pseudonym of the statistician William Sealy Gosset, who introduced the t-distribution.