Chapter 1
Introduction

Abstract In this chapter we give an introduction to spatial data analysis, and distinguish it from other forms of data analysis. By spatial data we mean data that contain locational as well as attribute information. We focus on two broad types of spatial data: area data and origin–destination flow data. Area data relate to a situation where the variable of interest—at least as our book is concerned—does not vary continuously, but has values only within a fixed set of areas or zones covering the study area. These fixed sites may either constitute a regular lattice (such as pixels in remote sensing) or they may consist of irregular areal units (such as, for example, census tracts). Origin–destination flow (also called spatial interaction) data are related instead to pairs of points, or pairs of areas in geographic space. Such data—that represent flows of people, commodities, capital, information or knowledge, from a set of origins to a set of destinations—are relevant in studies of transport planning, population migration, journey-to work, shopping behaviour, freight flows, and the transmission of information and knowledge across space. We consider the issue of spatial autocorrelation in the data, rendering conventional statistical analysis unsafe and requiring spatial analytical tools. This issue refers to situations where the observations are non-independent over space. And we conclude with a brief discussion of some practical problems which confront the spatial analyst.

Keywords Spatial data · Types of spatial data · Spatial data matrix · Area data · (Origin–destination) Flow data · Spatial autocorrelation · Tyranny of spatial data

1.1 Data and Spatial Data Analysis

Data consist of numbers, or symbols that are in some sense neutral and—in contrast to information—almost context-free. Raw geographical facts, such as the temperature at a specific time and location, are examples of data. Following

M. M. Fischer and J. Wang, *Spatial Data Analysis*,
SpringerBriefs in Regional Science, DOI: 10.1007/978-3-642-21720-3_1,
© Manfred M. Fischer 2011

Longley et al. (2001, p. 64) we can view spatial data as built up from atomic elements or facts about the geographic world. At its most primitive, an atom of spatial data (strictly, a datum) links a geographic location (place), often a time, and some descriptive property or attribute of the entity with each other. For example, consider the statement "The temperature at 2 pm on December 24, 2010 at latitude 48°15′ North, longitude 16°21′ 28 s East, was 6.7°C". It ties location and time to the property or attribute of atmospheric temperature. Hence, we can say that spatial (geographic) data link place (location), time and an attribute (here: temperature).

Attributes come in many forms. Some are physical or environmental in nature, while others are social or economic. Some simply identify a location such as postal addresses or parcel identifiers used for recording land ownership. Other attributes measure something at a location (examples include atmospheric temperature and income), while others classify into categories such as, for example, land use classes that differentiate between agriculture, residential land and industry.

While time is optional in spatial data analysis, geographic location is essential and distinguishes *spatial data analysis* from other forms of data analysis that are said to be non-spatial or aspatial. If we would deal with attributes alone, ignoring the spatial relationships between sample locations, we could not claim of doing spatial data analysis, even though the observational units may themselves be spatially defined. Even if attribute data would be of fundamental importance, divorced from their spatial context, they lose value and meaning (Bailey and Gatrell 1995, p. 20). In order to undertake spatial data analysis, we require—as a minimum—information for both locations and attributes, regardless, of how the attributes are measured.

Spatial data analysis requires an underlying spatial framework on which to locate the spatial phenomena under study. Longley et al. (2001) and others have drawn a distinction between two fundamental ways of representing geography: a *discrete* and a *continuous* view of spatial phenomena. In other words, a distinction is made between a conception of space as something filled with "discrete objects", and a view of space as covered with essentially "continuous surfaces". The former has been labelled an *object* or *entity view of space*, the latter a *field view*.

In the object view the sorts of spatial phenomena being analysed are identified by their dimensionality. Objects that occupy area are called two-dimensional, and are generally referred to as areas. Other objects are more like one-dimensional lines, including rivers, railways, or roads, and are represented as one-dimensional objects and generally referred to as lines. Other objects are more like zero-dimensional points, such as individual plants, people, buildings, the epicentres of earthquakes, and so on, and are referred to as points (Longley et al. 2001, pp. 67–68; Haining 2003, pp. 44–46). Note that surface or volume objects—not considered in this book—have length, breadth, and depth, and thus are three-dimensional. They are used to represent natural objects such as river basins or artificial phenomena such as the population potential of shopping centres.

Of course, how appropriate this is depends upon the spatial scale (level of detail at which we seek to represent "reality") of study. If we are looking at the distribution of urban settlements at a national scale, it is reasonable to treat them as a

distribution of points. At the scale of a smaller region, for example, it becomes less sensitive. Phenomena such as roads can be treated as lines as mentioned above. But there is again scale dependence. On large scale maps of urban areas roads have a width, and this may be important when interest is on car navigation issues, for example. Lines also mark the boundaries of areas. By areas we generally understand those entities which are administratively or legally defined, such as countries, districts, census zones, and so on, but also "natural areas" such as soil or vegetation zones on a map.

In a field view the emphasis is on the continuity of spatial phenomena, and the geographic world is described by a finite number of variables, each measurable at any point of the earth's surface, and changing in value across the surface (Haining 2003, pp. 44–45). If we think of phenomena in the natural environment such as temperature, soil characteristics, and so on, then such variables can be observed anywhere on the earth's surface (Longley et al. 2001, pp. 68–71). Of course, in practice such variables are discretised. Temperature, for example, is sampled at a set of sites and represented as a collection of lines (so-called isotherms). Soil characteristics might be also sampled at a set of discrete locations and represented as a continuously varying field. In all such cases, an attempt is made to represent underlying continuity from discrete sampling (Bailey and Gatrell 1995, p. 19).

1.2 Types of Spatial Data

In describing the nature of spatial data it is important to distinguish between the discreteness or continuity of the space on which the variables are measured, and the discreteness or continuity of the variable values (measurements) themselves. If the space is continuous (a field view), variable values must be continuous valued since continuity of the field could not be preserved under discrete valued variables. If the space is discrete (an object view) or if a continuous space has been made discrete, variable values may be continuously valued or discrete valued (nominal or ordinal valued) (see Haining 2003, p. 57).

The classification of spatial data by type of conception of space and level of measurement is a necessary first step in specifying the appropriate statistical technique to use to answer a question. But the classification is not sufficient because the same spatial object may be representing quite different geographical spaces. For example, points (so-called centroids) are also used to represent areas. Table 1.1 provides a typology that distinguishes four types of spatial data:

(i) *point pattern data*, that is, a data set consisting of a series of point locations in some study region, at which events of interest (in a general sense) have occurred, such as cases of a disease or incidence of a type of crime,
(ii) *field data* (also termed *geostatistical data*) that relate to variables which are conceptually continuous (the field view) and whose observations have been sampled at a predefined and fixed set of point locations,

Table 1.1 Types of spatial data: conceptual schemes and examples

Type of spatial data	Conceptual scheme		Example	
	Variable scheme	Spatial index	Variable	Space
Point pattern data	Variable (discrete or continuous) is a random variable	Point objects to which the variable is attached are fixed	Trees: diseased or not	Two-dimensional discrete space
			Hill forts: classified by type	Two-dimensional discrete space
Spatially continuous (geostatistical) data	Variable is a continuous valued function of location	Variable is defined everywhere in the (two-dimensional) space	Temperature	Two-dimensional continuous space
			Atmospheric pollution	Two-dimensional continuous space
Area (object) data	Variable (discrete or continuous) is a random variable	Area objects to which the variable is attached are fixed	Gross regional product	Two-dimensional discrete space
			Crime rates	Two-dimensional discrete space
Spatial interaction (flow) data	Variable representing mean interaction frequencies is a random variable	Pairs of locations (points or areas) to which the flow variable is attached	International trade	Two-dimensional discrete space
			Population migration	Two-dimensional discrete space

(iii) *area data* where data values are observations associated with a fixed number of areal units (area objects) that may form a regular lattice, as with remotely sensed images, or be a set of irregular areas or zones, such as counties, districts, census zones, and even countries,
(iv) *spatial interaction data* (also termed *origin–destination flow* or *link data*), consisting of measurements each of which is associated with a pair of point locations, or pair of areas.

In this book, we do neither consider point pattern data nor field (geostatistical) data. The focus is rather on the analysis of object data where the observations relate to areal units (see Part I) and on the analysis of origin–destination flow (spatial interaction) data (see Part II). The analysis of spatial interaction data has a long and distinguished history in the study of human activities, such as transportation movements, migration, and the transmission of information and knowledge. And area data provide an important perspective for spatial data analysis applications, in particular in the social sciences.

1.3 The Spatial Data Matrix

All the analytical techniques in this book use a data matrix that captures the spatial data needed for the conduct of analysis. Spatial data are classified by the type of spatial object (point object, area object) to which variables refer and the level of measurement of these variables.

Let $Z_1, Z_2, \ldots, Z_K$ refer to K random variables and S to the location of the point or area. Then the spatial data matrix (see Haining 2003, pp. 54–57) can be generally represented as

$$\begin{array}{cccccc} \multicolumn{4}{c}{\text{Data on the } K \text{ variables}} & \multicolumn{2}{c}{\text{Location}} & \\ Z_1 & Z_2 & \ldots & Z_K & S & \\ \left[\begin{array}{cccccc} z_1(1) & z_2(1) & \ldots & z_K(1) & s(1) \\ z_1(2) & z_2(2) & \ldots & z_K(2) & s(2) \\ \vdots & \vdots & & \vdots & \vdots \\ z_1(n) & z_2(n) & \ldots & z_K(n) & s(n) \end{array}\right] & \begin{array}{c} \text{Case 1} \\ \text{Case 2} \\ \vdots \\ \text{Case } n \end{array} \end{array}$$

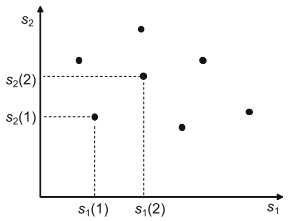
which may be shortened to

$$\left\{ z_1(i), z_2(i), \ldots, z_K(i) \,|\, s(i) \right\}_{i=1,\ldots,n} \tag{1.1}$$

where the lower case symbol z_k denotes an realisation (actual data value) of variable Z_k $(k = 1, \ldots, K)$ while the symbol i inside the brackets references the particular case. Attached to each case $i = 1, \ldots, n$ is a location $s(i)$ that represents the location of the spatial object (point or area). Since we are only interested in

(a) Assigning locations to point objects

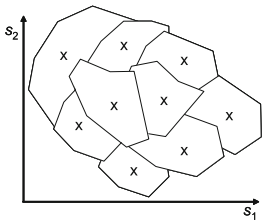
Case i	$s(i)$		Variables			
	s_1	s_2	Z_1	Z_2	…	Z_K
1	$s_1(1)$	$s_2(1)$	$z_1(1)$	$z_2(1)$	…	$z_K(1)$
2	$s_1(2)$	$s_2(2)$	$z_1(2)$	$z_2(2)$	…	$z_K(2)$
⋮	⋮		⋮	⋮		⋮
n	$s_1(n)$	$s_2(n)$	$z_1(n)$	$z_2(n)$	…	$z_K(n)$



(b) Assigning locations to irregularly shaped area objects

Case i	$s(i)$	Variables			
		Z_1	Z_2	…	Z_K
1	1	$z_1(1)$	$z_2(1)$	…	$z_K(1)$
2	2	$z_1(2)$	$z_2(2)$	…	$z_K(2)$
⋮	⋮	⋮	⋮		⋮
n	n	$z_1(n)$	$z_2(n)$	…	$z_K(n)$

1, 2, …, n look up table



(c) Assigning locations to regularly shaped area objects

Case i	$s(i)$		Variables			
	p	q	Z_1	Z_2	…	Z_K
1	$s_1(1)$	$s_2(1)$	$z_1(1)$	$z_2(1)$	…	$z_K(1)$
2	$s_1(2)$	$s_2(2)$	$z_1(2)$	$z_2(2)$	…	$z_K(2)$
⋮	⋮		⋮	⋮		⋮
n	$s_1(n)$	$s_2(n)$	$z_1(n)$	$z_2(n)$	…	$z_K(n)$

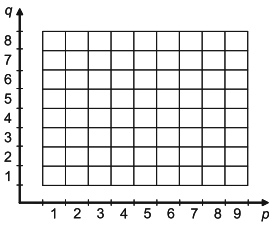


Fig. 1.1 Assigning locations to spatial objects (points, areas) (adapted from Haining 2003, p. 55)

two-dimensional space, referencing will involve two geographic coordinates s_1 and s_2. Thus, $s(i) = (s_1(i), s_2(i))'$ where $(s_1(i), s_2(i))'$ is the transposed vector of $(s_1(i), s_2(i))$. It is important to note that in this book we generally consider methods that treat locations as fixed and do not consider problems where there is a randomness associated with the location of the cases.

In the case of data referring to point objects in two-dimensional space the location of the ith point may be given by a pair of (orthogonal) Cartesian coordinates as illustrated in Fig. 1.1a. The axes of the coordinate system will usually have been constructed for the particular data set, but a national or global referencing system may be used. In the case of data referring to irregularly shaped area objects one option is to select a representative point such as the centroid and then use the same procedure as for a point object to identify $s(i) = (s_1(i), s_2(i))'$ for $i = 1, \ldots, n$. Alternatively, each area i is labelled and a look-up table provided

so that rows of the data matrix can be matched to areas on the map (see Fig. 1.1b). If the areas are regularly shaped as in the case of a remotely sensed image they may be labelled as in Fig. 1.1c.

There are situations where the georeferencing information provided by $\{s(i)\}$ in expression (1.1) has to be supplemented with neighbourhood information that defines not only which pairs of areas are adjacent to each other but may also quantify the closeness of that adjacency. This information is needed for the specification of many spatial statistical models such as spatial regression models.

It is worth noting that on various occasions throughout the book, the variables $Z_1, \ldots, Z_K$ will be divided into groups and labelled differently. In the case of data modelling, the variable whose variation is to be modelled will be denoted Y and the variables used to explain the variations in the dependent variable are called explanatory or independent variables, labelled differently such as $X_1, \ldots, X_Q$.

Spatial interaction data record flows between locations (points, areas) or between nodes (intersection points) of a network. The situation, we are considering in this book is one of a series of observations $y_{ij}(i,j = 1, \ldots, n)$, on random variables Y_{ij}, each of which corresponds to movements of people, goods, capital, information, knowledge, and so on between spatial locations i and j, where these locations may be point locations or alternatively areas or zones. These data are captured in the form of an origin–destination or spatial interaction matrix

$$\begin{array}{c} \\ \text{Origin location} \end{array} \begin{array}{c} \text{Destination location} \\ \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{n'1} & y_{n'2} & \cdots & y_{n'n} \end{bmatrix} \end{array} \tag{1.2}$$

where the number of rows and columns correspond to the number of origin and destination locations, respectively, and the entry on row i and column j, y_{ij}, records the observed total flow from origin location i to destination location j. In the special case where each location is both origin and destination $n' = n$. Georeferencing of the origin and destination locations follows the same procedures as described in the above case of object data.

1.4 Spatial Autocorrelation

The basic tenet underlying the analysis of spatial data is the proposition that values of a variable in near-by locations are more similar or related than values in locations that are far apart. This inverse relation between value association and distance is summarised by Tobler's first law stating that "*everything is related to everything else, but near things are more related than distant things*" (Tobler 1970, p. 234).

If near-by observations (i.e. similar in location) are also similar in variable values then the pattern as a whole is said to exhibit *positive* spatial autocorrelation

(self-correlation). Conversely, *negative* spatial autocorrelation is said to exist when observations that are near-by in space tend to be more dissimilar in variable values than observations that are further apart (in contradiction to Tobler's law). Zero autocorrelation occurs when variable values are independent of location. It is important to note that spatial autocorrelation renders conventional statistical analysis invalid and makes spatial data analysis different from other forms of data analysis.

A crucial aspect of defining spatial autocorrelation is the determination of near-by locations, that is, those locations surrounding a given data point that could be considered to influence the observation at that data point. Unfortunately, the determination of that neighbourhood is not without some degree of arbitrariness.

Formally, the membership of observations in the neighbourhood set for each location may be expressed by means of an n-by-n spatial contiguity or weights matrix W

$$W = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ \vdots & \vdots & & \vdots \\ W_{n1} & W_{n2} & \dots & W_{nn} \end{bmatrix} \tag{1.3}$$

where n represents the number of locations (observations). The entry on row i ($i = 1, \dots, n$) and column j ($j = 1, \dots, n$), denoted as W_{ij}, corresponds to the pair (i, j) of locations. The diagonal elements of the matrix are set to zero, by convention, while the non-diagonal elements W_{ij} ($i \neq j$) take on non-zero values (one, for a binary matrix) when locations i and j are considered to be neighbours, otherwise zero.

For areal objects, such as the simple nine-zone system shown in Fig. 1.2a, (first order) spatial contiguity (or adjacency) is often used to specify neighbouring locations in the sense of sharing a common border. On this basis, Fig. 1.2a may be re-expressed as the graph shown in Fig. 1.2b. Coding $W_{ij} = 1$ if zones i and j are contiguous, and $W_{ij} = 0$ otherwise, we may derive a weights matrix W shown in Table 1.2. This matrix provides an example of the simplest way of specifying W.

In the classical case of a regular square grid layout the options of contiguity are referred to as the *rook* contiguity case (only common boundaries), the *bishop* contiguity case (only common vertices), and the *queen* contiguity case (both boundaries and vertices). Depending on the chosen criterion, an area will have four (rook, bishop) or eight (queen) neighbours on average. This implies quite different neighbour structures. Even in the case of irregularly shaped areal units, a decision has to be made whether areas that only share a common vertex should be considered to be neighbours (queen criterion) or not (rook criterion).

Contiguity may and is often defined as a function of the distance between locations (areas, points). In this sense, two objects are considered to be contiguous if the distance between them falls within a chosen range. In essence, the spatial weights matrix summarises the topology of the data set in graph-theoretic terms (nodes and links).

Higher order contiguity is defined in a recursive manner, in the sense that an object (point, area) is considered to be contiguous of a higher order to a given object if it is

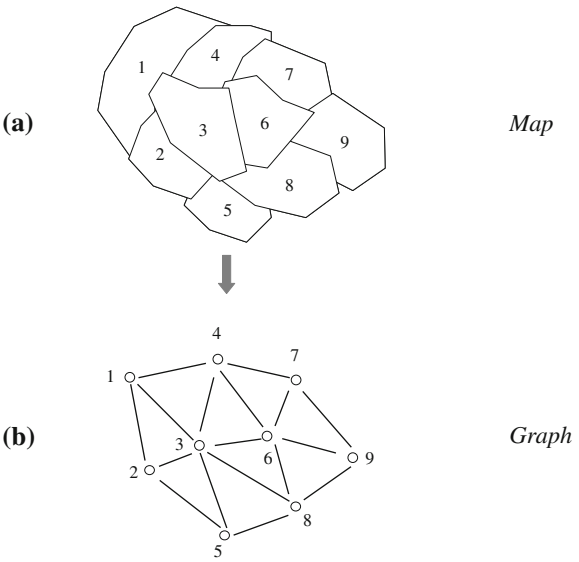


Fig. 1.2 A zoning system: **a** a simple mosaic of discrete zones, **b** re-expressed as a graph

Table 1.2 A spatial weights matrix *W* derived from the zoning system in Fig. 1.2: the case of a binary first order contiguity matrix

	1	2	3	4	5	6	7	8	9
1	0	1	1	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	1	1	0	1	1	1	0	1	0
4	1	0	1	0	0	1	1	0	0
5	0	1	1	0	0	0	0	1	0
6	0	0	1	1	0	0	1	1	1
7	0	0	0	1	0	1	0	0	1
8	0	0	1	0	1	1	0	0	1
9	0	0	0	0	0	1	1	1	0

first order contiguous to an object that is contiguous to an object that is contiguous of the next lower order. For example, objects that are viewed to be second order contiguous to an object are first order contiguous to the first order contiguous ones. In Fig. 1.2a, for example, areas 1 and 2 are first order contiguous to area 3, and area 3 is first order contiguous to area 6. Hence, areas 1 and 2 are second order contiguous to area 6. Thus, higher order contiguity yields bands of observations around a given location being included in the neighbourhood set, at increasing instances.

Clearly, a large number of spatial weights matrices may be derived for a given spatial layout such as that one shown in Fig. 1.2a. In particular, the spatial weights matrix does not have to be binary, but can take on any value that reflects the interaction between spatial units *i* and *j*, for example, based on inverse distances or inverse distances raised to some power.

The type of matrix shown in Table 1.2 allows us to develop measures of spatial autocorrelation. Many tests and indicators of spatial autocorrelation are available.

Chief among these is Moran's spatial autocorrelation statistic (see Cliff and Ord 1973, 1981). At the local scale Getis and Ord's statistics (see Getis and Ord 1992; Ord and Getis 1995) and Anselin's LISA statistics (see Anselin 1995) enable analysts to evaluate spatial autocorrelation at particular sites. We will say more about this in the next chapter.

1.5 The Tyranny of Spatial Data

Spatial data analysis crucially depends on data quality. Good data are reliable, contain few or no mistakes, and can be used with confidence. Unfortunately, nearly all spatial data are flawed to some degree. Errors may arise in measuring both the location (points, lines, areas) and attribute properties of spatial objects. In the case of measurements of location (position), for example, it is possible for every coordinate to be subject to error. In the two-dimensional case, a measured location would be subject to error in both coordinates.

Attribute errors can arise as a result of collecting, storing, manipulating, editing or retrieving attribute values. They can also arise from inherent uncertainties associated with the measurement process and definitional problems, including the point or area location a measurement refers to (Haining 2003, pp. 59–63; see also Wang et al. 2010). The solution to the data quality problem is to take the necessary steps to avoid having faulty data determining research results.

The particular form (i.e. size, shape and configuration) of the spatial aggregates can affect the results of the analysis to a varying—usually unknown—degree as evidenced in various types of analysis (see, for example, Openshaw and Taylor 1979; Baumann et al. 1983). This problem generally has become recognised as the *modifiable areal unit problem* (MAUP), the term stemming from the fact that areal units are not 'natural' but usually arbitrary constructs.

Confidentiality restrictions usually dictate that data (for example, census data) may not be released for the primary units of observation (individuals, households or firms), but only for a set of rather arbitrary areal aggregations (enumeration districts or census tracts). The problem arises whenever area data are analysed or modelled and involves two effects: One derives from selecting different areal boundaries while holding the overall size and the number of areal units constant (the zoning effect). The other derives from reducing the number but increasing the size of the areal units (the scale effect). There is no analytical solution to the MAUP (Openshaw 1981), but the modifiable areal unit problem can be investigated through simulation of large numbers of alternative systems of areal units (Longley et al. 2001, p. 139). Such systems can obviously take many different forms, both in relation to the level of spatial resolution and also in relation to the shape of the areas.

An issue that has been receiving increasing attention in recent years relates to the data suitability problem. It is not unusual to find published work where the researcher uses data available at one spatial scale to come to conclusions about a

relationship or process at a finer scale. This *ecological fallacy*, as it is known, leads us into a false sense of the power of our techniques and usefulness of our conclusions (Getis 1995). The ecological fallacy and the modifiable areal unit problem have long been recognised as problems in applied spatial data analysis, and, through the concept of spatial autocorrelation, they are understood as related problems.

Part I
The Analysis of Area Data

In Part I we consider the analysis of area data. Area data are observations associated with a fixed number of areal units (areas). The areas may form a regular lattice, as with remotely sensed images, or be a set of irregular areas or zones, such as countries, districts and census zones.

We draw a distinction between methods that are essentially exploratory in nature, concerned with mapping and geovisualisation, summarising and analysing patterns, and those which rely on the specification of a statistical model and the estimation of the model parameters. The distinction is useful, but not clear cut. In particular, there is usually a close interplay of the two, with data being visualised and interesting aspects being explored, which possibly lead to some modelling.

Chapter 2 will be devoted to methods and techniques of exploratory data analysis that concentrates on the spatial aspects of the data, that is, exploratory spatial data analysis [ESDA] (see Haining 2003; Bivand 2010). The focus is on univariate techniques that elicit information about spatial patterns of a variable, and identify atypical observations (outliers).

Exploratory spatial data analysis is often only a preliminary step towards more formal modelling approaches that seek to establish relationships between the observations of a variable and the observations of other variables, recorded for each area. Chapter 3 provides a concise overview of some of the central methodological issues related to spatial regression analysis in a simple cross-sectional setting.

Keywords Area data · Spatial weights matrix · Moran's I statistic · Geary's c statistic · G statistics · LISA statistics · Spatial regression models · Spatial Durbin model · Tests for spatial dependence · Maximum likelihood estimation · Model parameter interpretation

Chapter 2
Exploring Area Data

Abstract Here in this chapter, we first consider the visualisation of area data before examining a number of exploratory techniques. The focus is on spatial dependence (spatial association). In other words, the techniques we consider aim to describe spatial distributions, discover patterns of spatial clustering, and identify atypical observations (outliers). Techniques and measures of spatial autocorrelation discussed in this chapter are available in a variety of software packages. Perhaps the most comprehensive is GeoDa, a free software program (downloadable from http://www.geoda.uiuc.edu). This software makes a number of exploratory spatial data analysis (ESDA) procedures available that enable the user to elicit information about spatial patterns in the data given. Graphical and mapping procedures allow for detailed analysis of global and local spatial autocorrelation results. Another valuable open software is the spdep package of the R project (downloadable from http://cran.r-project.org). This package contains a collection of useful functions to create spatial weights matrix objects from polygon contiguities, and various tests for global and spatial autocorrelation (see Bivand et al. 2008).

Keywords Area data · Spatial weights matrix · Contiguity-based specifications of the spatial weights matrix · Distance-based specifications of the spatial weights matrix · k-nearest neighbours · Global measures of spatial autocorrelation · Moran's I statistic · Geary's c statistic · Local measures of spatial autocorrelation · G statistics · LISA statistics

2.1 Mapping and Geovisualisation

In exploratory spatial data analysis the map has an important role to play. The map is the most established and conventional means of displaying areal data. There is a variety of ways ascribing continuous variable data to given areal units that are predefined. In practice, however, none is unproblematic. Perhaps the most commonly

M. M. Fischer and J. Wang, *Spatial Data Analysis*,
SpringerBriefs in Regional Science, DOI: 10.1007/978-3-642-21720-3_2,
© Manfred M. Fischer 2011

used form of display is the standard *choropleth* map (Longley et al. 2001, pp. 251–252; Bailey and Gatrell 1995, pp. 255–260; Demšar 2009, pp. 48–55). This is a map where each of the areas is coloured or shaded according to a discrete scale based on the value of the variable (attribute) of interest within that area. The number of classes (categories) and the corresponding class (category) intervals can be based on several different criteria.

There are no hard rules about numbers of classes. Clearly, this is a function of how many observations we have. For example, if we have only a sample of 20 or 30 areas it hardly makes sense to use seven or eight classes. But perhaps with some hundreds of measurements a set of seven or eight classes is likely to prove informative. As a general rule of thumb some statisticians recommend a number of classes of $(1 + 3.3 \ln n)$, where n is the number of areas and 'ln' stands for the logarithm naturalis (Bailey and Gatrell 1995, p. 153).

As for class interval selection, four basic classification schemes may be used to divide interval and ratio areal data into categories (Longley et al. 2001, p. 259):

(i) *Natural breaks* by which classes are defined according to some natural groupings of the data values. The breaks may be imposed on the basis of break points which are known to be relevant in a particular application context, such as fractions and multiples of mean income levels, or rainfall thresholds of vegetation ('arid', 'semi-arid', 'temperate' etc.). This is a deductive assignment of breaks, while inductive classifications of data values may be carried out by using GISystem software tools to look for relatively large jumps in data values, as shown in Fig. 2.1a.

(ii) *Quantile breaks*, where each of a predetermined number of classes (categories) contains an equal number of observations (see Fig. 2.1b). Quartile (four category) and quintile (five category) classifications are commonly used in practice. The numeric size of each class is rigidly imposed. Note that the placing of the class boundaries may assign almost identical observations to adjacent classes, and observations with quite widely different values to the same class. The resulting visual distortion can be minimised by increasing the number of classes.

(iii) *Equal interval breaks* are self-explanatory (see Fig. 2.1c). They are valuable where observations are reasonably uniformly distributed over their range. But if the data are markedly skewed they will give large numbers of observations in just a few classes. This is not necessarily a problem, since unusually high (low) values are easily picked out on the map. An extension of this scheme is to use "trimmed equal" intervals where the top and bottom of the frequency distribution (for example, the top and bottom ten percent) are each treated as separate classes and the remainder of the observations are divided into equal classes.

(iv) *Standard deviation classifications* are based on intervals distributed around the mean in units of standard deviation (see Fig. 2.1d). They show the distance of an observation from the mean. One calculates the mean value and then generates class breaks in standard deviation measures above and below it.

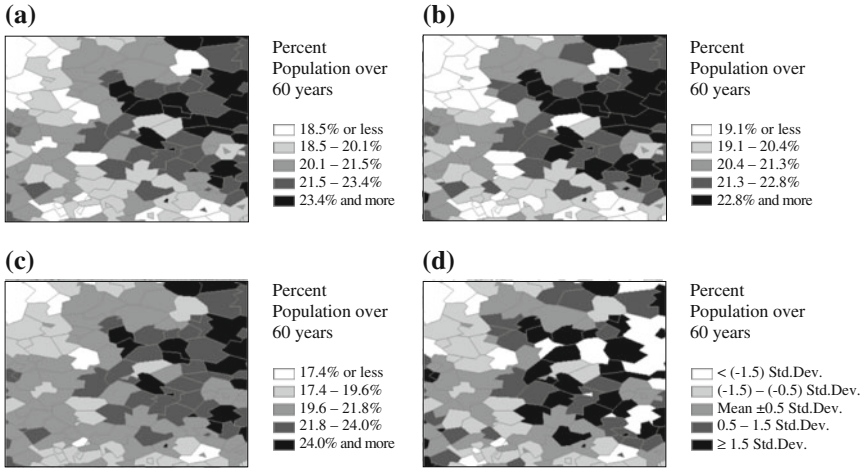


Fig. 2.1 Class definition using **a** natural breaks, **b** quantile breaks, **c** equal interval breaks, and **d** standard deviation breaks

We can obtain a large variety of maps simply by varying the class intervals. It is important to try out some of the above possibilities to get some initial sense of spatial variation in the data.

There are also other problems associated with the use of choropleth maps. *First*, choropleth maps bring the (dubious) visual implication of within-area uniformity of variable values. Moreover, conventional choropleth mapping allows any physically large area to dominate the display, in a way which may be quite inappropriate for the type of data being mapped. For example, in mapping socioeconomic data, large and sparsely populated rural areas may dominate the choropleth map because of the visual 'intrusiveness' of the large areas. But the real interest may be in physically smaller areas, such as the more densely populated urban areas.

A variant of the conventional choropleth map is the dot density map that uses dots as a more aesthetically pleasing means of representing the relative density of zonally averaged data, but not as a means of depicting the precise location of point events. Proportional circles provide one way around this problem, since the circle can be centred on any convenient point within an areal unit. But there is a tension between using circles that are of sufficient size to convey the variability in the data and the problems of overlapping circles (Longley et al. 2001, p. 259).

Second, the variable of interest has arisen from the aggregation of individual data to the areas. It has to be taken into account that these areas may have been designed rather arbitrarily on the basis of administrative convenience or ease of enumeration. Hence, any pattern that is observed across the areas may be as much a function of the area boundaries chosen, as it is of the underlying spatial distribution of variable values. This has become known as the *modifiable areal unit problem*. It can be a particularly significant problem in the analysis of socioeconomic and demographic data, where the enumeration areas have rarely been arrived at any basis that relates to the data under study (see also Sect. 1.5).

A solution to the problem of modifiable areal units is difficult. The ideal solution is to avoid using area aggregated data altogether if at all possible. In some applications of spatial data analysis, such as epidemiology and crime, one could perform valuable analyses on point data, without aggregating the data to a set of inherently arbitrary areal units. But of course in many cases such an approach will be not viable, and one has to live with areal units for which data are available.

Third, it is important to realise that the statistical results of any analysis of patterns and relationships will inevitably depend on the particular areal configuration which is being used. In general, data should be analysed on the basis of the smallest areal units for which they are available and aggregation to arbitrary larger areas should be avoided, unless there are good reasons to doing so. It is also important to check any inferences drawn from the data by using different areal configurations of the same data, if possible.

One approach to the problem of the dominance of physically large areas is to geometrically transform each of the areal units in such a way as to make its area proportional to the corresponding variable value, whilst at the same time maintaining the spatial contiguity of the areal units. The resulting map is often termed *cartogram* (Bailey and Gatrell 1995, p. 258). Cartograms lack planimetric correctness, and distort area or distance in the interest of some specific objective. The usual objective is to reveal patterns that might not be readily apparent from a conventional map. Thus, the integrity of the spatial object (area), in terms of areal extent, location, contiguity, geometry, and topology is made subserviant to an emphasis upon variable values or particular aspects of spatial relations.

An example of a cartogram is given in Fig. 2.2 that shows a country's size as the proportion of global gross domestic product (*gdp*) found there in 2005, measured in terms of constant US dollars. The map reveals that global *gdp* is concentrated in a few world regions, in North America, Western Europe and North-East Asia. This global concentration matters greatly for the development prospects of today's lagging world regions, especially Africa which shows up as a slender peninsula in this cartogram.

Mapping and geovisualisation is an important step to provoke questions, but exploratory data analysis requires highly interactive, dynamic data displays. Recent developments in spatial data analysis software provide an interactive environment that combines maps with statistical graphs, using the technology of dynamically linked windows. Perhaps, the most comprehensive software with such capabilities is GeoDa. GeoDa includes functionality from conventional mapping to exploratory data analytic tools, and the visualisation of global and local autocorrelation. The software adheres to ESRI's (Environmental Systems Research Institute's) shape file as the standard for storing spatial information, and uses ESRI's Map-Objects LT2 technology for spatial data access, mapping and querying.

All graphic windows are based on Microsoft Foundation Classes and hence limited to Microsoft Windows platforms. In contrast, the computational engine (including statistical operations) is pure C++ code and largely cross platform. The bulk of the graphical interface implements five basic classes of windows:

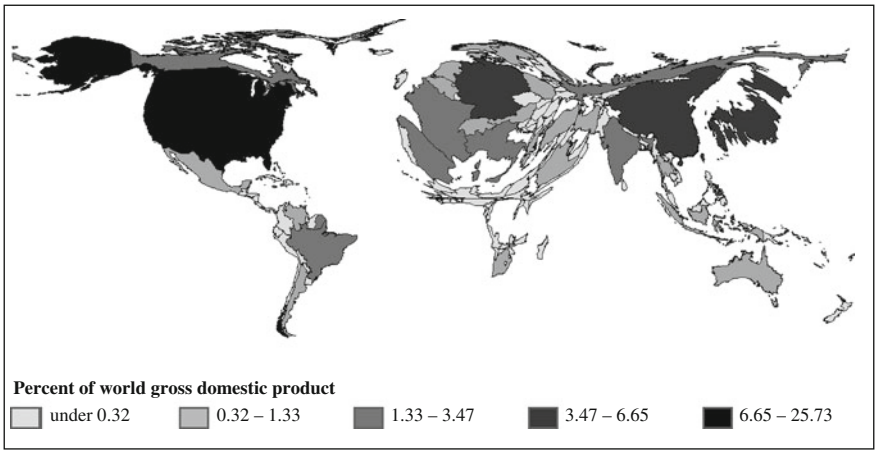


Fig. 2.2 A cartogram illustrating the concentration of global gross domestic product in a few world regions. A country's size shows the proportion of global gross domestic product found there. *Data source:* GISCO-Eurostat (European Commission); *Copyright:* EuroGeographics for the European administrative boundaries; *Copyright:* UN-FAO for the world administrative boundaries (except EuroGeographics members)

map, histogram, box plot, scatter plot (including the Moran scatter plot, see Anselin 1996), and grid (for the table selection and calculations). The choropleth map, including cluster maps for the local indicators of spatial autocorrelation (see Sect. 2.4), are derived from MapObjects classes. For an outline of the design and review of the overall functionality of GeoDa see Anselin et al. (2010).

2.2 The Spatial Weights Matrix

The focus of exploratory spatial data analysis is on measuring and displaying global and local patterns of spatial association, indicating local non-stationarity, discovering islands of spatial heterogeneity, and so on. A crucial aspect of defining spatial association (autocorrelation) is the determination of the relevant "neighbourhood" of a given area, that is, those areal units surrounding a given data point (area) that would be considered to influence the observation at that data point. In other words, neighbouring areas are spatial units that interact in a meaningful way. This interaction could relate, for example, to spatial spillovers and externalities.

The neighbourhood structure of a data set is most conveniently formalised in form of a spatial weights matrix W, of dimension equal to the a priori given number n of areal units considered. Each area is identified with a point (centroid) where Cartesian coordinates are known. In this matrix each row and matching column corresponds to an observation pair. The elements W_{ij} of this matrix take on a non-zero value (one for a binary matrix) when areas (observations) i and j are

considered to be neighbours, and a zero value otherwise. By convention, an observation is not a neighbour to itself, so that the main diagonal elements W_{ii} $(i = 1, \ldots, n)$ are zero.

The spatial weights matrix is often row-standardised, that is, each row sum in the matrix is made equal to one, the individual values W_{ij} are proportionally represented. Row-standardisation of W is desirable so that each neighbour of an area is given equal weight and the sum of all W_{ij} (over j) is equal to one. If the observations are represented as an n-by-1 vector X, the product, WX, of such a row-standardised weights matrix W with X has an intuitive interpretation. Since for each element i WX equals $\Sigma_j W_{ij}X_j$, WX is in fact a vector of weighted averages of neighbouring values. This operation and the associated variable are typically referred to as (first order) spatial lag of X, similar to the terminology used in time series analysis. Note that the space in which the observations are located need not to be geographic, any type of space is acceptable as long as the analyst can specify the spatial interactions between the areas given.

One way to represent the spatial relationships with areal data is through the concept of contiguity. First order contiguous neighbours are defined as areas that have a common boundary. Formally,

$$W_{ij} = \begin{cases} 1 & \text{if area } j \text{ shares a common boundary with area } i \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Alternatively, two areas i and j may be defined as neighbours when the distance d_{ij} between their centroids is less than a given critical value, say d, yielding distance-based spatial weights

$$W_{ij} = \begin{cases} 1 & \text{if } d_{ij} < d \; (d > 0) \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where distances are calculated from information on latitude and longitude of the centroid locations. Examples include straight-line distances, great circle distances, travel distances or times, and other spatial separation measures.

Straight-line distances determine the shortest distance between any two point locations in a flat plane, treating longitude and latitude of a location as if they were equivalent to plane coordinates. In contrast, great circle distances determine distances between any two points on a spherical surface such as the earth as the length of the arc of the great circle between them (see Longley et al. 2001, pp. 86–92, for more details). In many applications the simple measures—straight-line distances and great circle distances—are not sufficiently accurate estimates of actual travel distances, and one is forced to resort to summing the actual lengths of travel routes, using a GISystem. This normally means summing the length of links in a network representation of a transportation system.

The distance-based specification (2.2) of the weights matrix depends on a given critical distance value, d. When there is a high degree of heterogeneity in the size of the areal units, however, it can be difficult to find a satisfactory critical distance. In such circumstances, a small distance will tend to lead to a lot of islands

(i.e. unconnected observations), while a distance chosen to guarantee that each areal unit (observation) has at least one neighbour may yield an unacceptably large number of neighbours for the smaller areal units (Anselin 2003a).

In empirical applications, this problem is encountered when building distance-based spatial weights, for example, for NUTS-2 regions in Europe, where such areal units in sparsely populated parts of Europe are much larger in physical size than in more populated parts such as in Central Europe. A common solution to this problem is to constrain the neighbour structure to the *k-nearest* neighbours, and thereby precluding islands and forcing each areal unit to have the same number k of neighbours. Formally,

$$W_{ij} = \begin{cases} 1 & \text{if centroid of } j \text{ is one of the } k \text{ nearest centroids to that of } i \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

If the number of nearest neighbours, for example, is set to six, then the non-normalised weights matrix will have six ones in each row, indicating the six closest observations to $i = 1, \ldots, n$. The number of neighbours, k, is the parameter of this weighting scheme. The choice of k remains an empirical matter (see LeSage and Fischer 2008).

The above specifications of the spatial weights matrix share the property that their elements are fixed. It is straightforward to extend this notion by changing the weighting on the neighbours so that more distant neighbours get less weight by introducing a parameter θ that allows to indicate the rate of decline of the weights. A commonly used continuous weighting scheme is based on the inverse distance function so that the weights are inversely related to the distance separating area i and area j

$$W_{ij} = \begin{cases} d_{ij}^{-\theta} & \text{if intercentroid distance } d_{ij} < d \; (d > 0, \; \theta > 0) \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

where the parameter θ is either estimated or set a priori. Common choices are the integers one and two, the latter following from the Newtonian gravity model. Another continuous weighting scheme is derived from the negative exponential function yielding

$$W_{ij} = \begin{cases} \exp(-\theta d_{ij}) & \text{if intercentroid distance } d_{ij} < d \; (d > 0, \; \theta > 0) \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where θ is a parameter that may be estimated, but is usually a priori chosen by the researcher. A popular choice is $\theta = 2$.

Evidently, a large number of spatial weights matrices can be derived for the same spatial layout. It is important to always keep in mind that the results of any spatial statistical analysis are conditional on the spatial weights matrix chosen. It is often good practice to check the sensitivity of the conclusions to the choice of the spatial weights matrix, unless there is a compelling reason on theoretical grounds to consider just a single one.

2.3 Global Measures and Tests for Spatial Autocorrelation

Spatial autocorrelation (association) is the correlation among observations of a single variable (*auto* meaning self) strictly attributable to the proximity of those observations in geographic space. This notion is best summarised by Tobler's first law which states that "*everything is related to everything else, but near things are more related than distant things*" (Tobler 1970, p. 234). Today, a number of measures of spatial autocorrelation are available (see Getis 2010 for a review).

Spatial autocorrelation measures deal with covariation or correlation between neighbouring observations of a variable. And thus compare two types of information: similarity of observations (value similarity) and similarity among locations (Griffith 2003). To simplify things, we will use the following notation

n number of areas in the sample,
i, j any two of the areal units,
z_i the value (observation) of the variable of interest for region i,
W_{ij} the similarity of i's and j's locations, with $W_{ii} = 0$ for all i,
M_{ij} the similarity of i's and j's observations of the variable.

Spatial autocorrelation (association) measures and tests may be differentiated by the scope or scale of analysis. Generally one distinguishes between global and local measures. Global implies that all elements in the W matrix are brought to bear on an assessment of spatial autocorrelation. That is, all spatial associations of areas are included in the calculation of spatial autocorrelation. This yields one value for spatial autocorrelation for any one spatial weights matrix. In contrast, local measures are focused. That is, they assess the spatial autocorrelation associated with one or a few particular areal units.

Global measures of spatial autocorrelation compare the set of value (observation) similarity M_{ij} with the set of spatial similarity W_{ij}, combining them into a single index of a cross-product, that is

$$\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} W_{ij}. \tag{2.6}$$

In other words, the total obtained by multiplying every cell in the W matrix with its corresponding entry in the M matrix, and summing. Adjustments are made to each index to make it easy to interpret (see below).

Various ways have been suggested for measuring value similarity (association) M_{ij}, dependent upon the scaling of the variable. For nominal variables, the approach is to set M_{ij} to one if i and j take the same variable value, and zero otherwise. For ordinal variables, value similarity is generally based on comparing the ranks of i and j. For interval variables both the squared difference $(z_i - z_j)^2$ and the product $(z_i - \bar{z})\,(z_j - \bar{z})$ are commonly used, where $\bar{z}$ denotes the average of the z-values.

The two measures that have been most widely used for the case of areal units and interval scale variables are Moran's I and Geary's c statistics. Both indicate the degree of spatial association as summarised for the whole data set. Moran's I uses cross-products to measure value association, and Geary's c squared differences. Formally, Moran's I is given by the expression (see Cliff and Ord 1981, p. 17)

$$I = \frac{n}{W_o} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \tag{2.7}$$

with the normalising factor

$$W_o = \sum_{i=1}^{n} \sum_{j \neq i}^{n} W_{ij}. \tag{2.8}$$

For ease of interpretation the spatial weights W_{ij} may be in row-standardised form, though this is not necessary, and by convention $W_{ii} = 0$ for all i. Note that for a row-standardised W, $W_o = n$.

Geary's c is estimated as (Cliff and Ord 1981, p. 17)

$$c = \frac{(n-1)}{2\,W_o} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \tag{2.9}$$

where W_o is given by Eq. (2.8). Neither of these statistics is constrained to lie in the $(-1, 1)$ range as in the case of conventional non-spatial product moment correlation. This is unlikely to present a practical problem for most real world data sets and reasonable W matrices (Bailey and Gatrell 1995, p. 270).

Spatial autocorrelation tests are decision rules based on statistics such as Moran's I and Geary's c to assess the extent to which the observed spatial arrangement of data values departs from the null hypothesis that space does not matter. This hypothesis implies that near-by areas do not affect one another such that there is independence and spatial randomness.

In contrast, under the alternative hypothesis of spatial autocorrelation (spatial association, spatial dependence), the interest renders on cases where large values are surrounded by other large values in near-by areas, or small values are surrounded by large values and vice versa. The former is referred to as *positive* spatial autocorrelation, and the latter as *negative* spatial autocorrelation. Positive spatial autocorrelation implies a spatial clustering of similar values (see Fig. 2.3a), while negative spatial autocorrelation implies a checkerboard pattern of values (see Fig. 2.3b).

Spatial autocorrelation is considered to be present when the spatial autocorrelation statistic computed for a particular pattern takes on a larger value, compared to what would be expected under the null hypothesis of no spatial

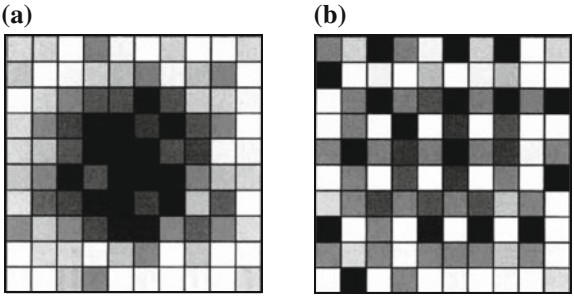


Fig. 2.3 Patterns of spatial autocorrelation on a regular grid: **a** positive spatial autocorrelation where cells with similar values (*gray tones*) are near-by; and **b** negative spatial autocorrelation where near-by cells have dissimilar values

association. What is viewed to be significantly larger depends on the distribution of the test statistic. We consider this question for the case of Moran's I statistic next.

In principle, there are two main approaches to testing observed I-values for significant departure from the hypothesis of zero spatial autocorrelation (Cliff and Ord 1981, p. 21). The first is the *random permutation test*. Under the randomisation assumption the observed value of I is assessed relative to the set of all possible values that could be obtained by randomly permuting the observations over the locations in the data set. Suppose we have n observations, z_i, relating to the a priori given areal units $i = 1, \ldots, n$.

Then $n!$ permutations are possible, each corresponds to a different arrangement of the n observations, z_i, over the areal units. One of these relates to the observed arrangement. The Moran I statistic can be computed for any of these $n!$ permutations. The resulting empirical distribution function provides the basis for a statement about the extremeness (or lack of extremeness) of the observed statistic, relative to the values computed under the null hypothesis (the randomly permuted values).

But computation of as many as $n!$ arrangements will be infeasible, even in the case of smaller n, since, for example, for $n = 10$ already 3,628,000 I-values would have to be calculated. But a close approximation to the permutation distribution can be obtained by using a Monte Carlo approach and simply sampling randomly from a reasonable number of the $n!$ possible permutations. Note that permutation re-orders the original data, whereas a Monte Carlo procedure generates "new" data of similar structure.

The other approach to testing observed I-values for significant departure from the hypothesis of zero spatial autocorrelation is based on an *approximate sampling distribution of I*. If there is a moderate number of areal units then an approximate result for the sampling distribution of I under certain assumptions may be utilised to develop a test. If it is assumed that the z_i are observations on random variables Z_i whose distribution is normal, then I has a sampling distribution that is appropriately normal with the moments

$$E(I) = -\frac{1}{(n-1)} \tag{2.10}$$

$$var\,(I) = \frac{n^2(n-1)\,W_1 - n\,(n-1)\,W_2 - 2\,W_o^2}{(n+1)\,(n-1)^2\,W_o^2} \tag{2.11}$$

where

$$W_o = \sum_{i=1}^{n}\sum_{j\neq i}^{n} W_{ij} \tag{2.8}$$

$$W_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(W_{ij} + W_{ji}\right)^2 \tag{2.12}$$

$$W_2 = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} W_{kj} + \sum_{i=1}^{n} W_{ik}\right)^2. \tag{2.13}$$

Hence, we can test the observed value of I against the percentage points of the appropriate sampling distribution. An "extreme" observed value of I indicates significant spatial autocorrelation. A value of Moran's I that exceeds its expected value of $-1/(n-1)$ points to *positive* spatial autocorrelation, while a value of Moran's I that is below the expectation indicates *negative* spatial autocorrelation (Bailey and Gatrell 1995, pp. 281–282).

Note that the hypothesis involved in each of the above two tests is somewhat different. The randomisation test embodies the assumption that no values of z_i other than those observed are realisable. In other words the data are treated as a population and the question analysed is how the data values are arranged spatially. Hence, the test is a test of patterns in the observations relative to the set of all possible patterns in the *given* observations. The approximate sampling distribution test makes the assumption that the observations z_i are observations on (normal) random variables Z_i. That is, they are one realisation of a random process and other possible realisations can occur. The test is, thus, one of spatial autocorrelation, providing the distribution of the random variables Z_i can be assumed to be normal (Bailey and Gatrell 1995, pp. 280–282; Fortin and Dale 2009).

Care is necessary in applying the above formal tests of spatial autocorrelation when I has been computed from residuals that arise from a regression (see Chap. 3). The problem arises because if Q parameters (regression coefficients β_q, $q = 1, \ldots, Q$) have been estimated in the regression, then the observed residuals are subject to Q linear constraints. That is, the observed residuals will be automatically spatially autocorrelated to some extent, and consequently the above testing procedure for Moran's I will not be valid. If $Q \ll n$, however, then it might be justified in ignoring this. If not, then strictly one should use adjustments to the mean and variance of the approximate sampling distribution of I. We do not go into details

here, but refer the reader to Chap. 3 and to the literature cited therein which also covers tests for spatial autocorrelation at spatial lags.

2.4 Local Measures and Tests for Spatial Autocorrelation

With the advent of large data sets characteristic of GISystems, it has become clear that the need to assess spatial autocorrelation globally may be of only marginal interest. During the past two decades, a number of statistics, called local statistics, have been developed. These provide for each observation of a variable an indication of the extent of significant spatial clustering of similar values around that observation. Hence, they are well suited to identify the existence of *hot spots* (local clusters of high values) or *cold spots* (local clusters of low values), and are appropriate to identify distances beyond which no discernible association exists.

Let us assume that each area $i\,(i = 1, \ldots, n)$ has associated with it a value z_i that represents an observation upon the random variable Z_i. Typically, it is assumed that the Z_i have identical marginal distributions. If they are independent, we say that there is no spatial structure. Independence implies the absence of spatial autocorrelation. But the converse is not necessarily true. Nevertheless, tests for spatial autocorrelation are characteristically viewed as appropriate assessment of spatial dependence (association). Usually, if spatial autocorrelation exists, it will be exhibited by similarities among neighbouring areas, although negative patterns of spatial association are also possible (Ord and Getis 1995, p. 287).

The basis for local tests for and measures of spatial autocorrelation comes from the cross-product statistic

$$\sum_{j=1}^{n} M_{ij}\, W_{ij} \tag{2.14}$$

that allows for spatial autocorrelative comparisons for a given observation (areal unit) $i = 1, \ldots, n$ where M_{ij} and W_{ij} are defined as in the previous section. We briefly describe four local statistics: the Getis and Ord local statistics G_i and G_i^*, and the local versions of Moran's I and Geary's c. Let us begin with the local statistics suggested by Getis and Ord (1992). The statistics are computed by defining a set of neighbours for each area i as those observations that fall within a critical distance d from i where each $i = 1, \ldots, n$ is identified with a point (centroid). This can be formally expressed in a set of symmetric binary weights matrices $W(d)$, with elements $W_{ij}(d)$ indexed by distance d. For each distance d, the elements $W_{ij}(d)$ of the corresponding weights matrix $W(d)$ equal one if i and j are within a distance from each other, and zero otherwise. Clearly, for different distance measures, a different set of neighbours will be found.

The G_i and G_i^* statistics measure the degree of local association for each observation i in a data set containing n observations. They consist of the ratio of

the sum of values in neighbouring areas, defined by a given distance band, to the sum over all observations (excluding the value in area i for the G_i statistic, but including it for the G_i^* statistic). These statistics may be computed for many different distance bands. Formally, the G_i measure for observation (area) i can be expressed as

$$G_i(d) = \frac{\sum_{j \neq i}^{n} W_{ij}(d)\, z_j}{\sum_{j \neq i}^{n} z_j} \tag{2.15}$$

with the summation in j exclusive of i. The G_i^* measure is given by

$$G_i^*(d) = \frac{\sum_{j=1}^{n} W_{ij}(d)\, z_j}{\sum_{j=1}^{n} z_j} \tag{2.16}$$

except that the summation in j is now inclusive of i. The G_i statistic can be interpreted as a measure of clustering of like values around a particular observation i, irrespective of the value in that area, while the G_i^* statistic includes the value within the measure of clustering. A positive value indicates clustering of high values and a negative value indicates a cluster of low values. It is interesting to note that G_i^* is mathematically associated with global Moran's $I(d)$ so that Moran's I may be interpreted as a weighted average of local statistics (Getis and Ord 1992). A slightly different form of the G-statistic was suggested by Ord and Getis (1995) where the distributional characteristics are discussed in detail (see also Getis 2010).

Getis and Ord (1992), and Ord and Getis (1995) provide the expected values and variances of the two statistics. Their distribution is normal if the underlying distribution of the observations is normal. But if the distribution is skewed, the test only approaches normality as the critical distance d increases, and does so more slowly for boundary areas where there are fewer neighbours. In other words, under these circumstances normality of the test statistic can only be guaranteed when the number of j neighbouring areas is large. When n is relatively small, as few as eight neighbours could be used without serious inferential errors unless the underlying distribution is very skewed (Getis and Ord 1996). Hot spots identified by these statistics can be interpreted as clusters or indications of spatial non-stationarity.

Local indicators of spatial association (LISA) *statistics* were derived by Anselin (1995), with the motivation to decompose global spatial autocorrelation statistics, such as Moran's I and Geary's c, into the contribution of each individual observation $i = 1, \ldots, n$. The local Moran statistic I_i for observation (area) $i = 1, \ldots, n$ is defined (Anselin 1995) as

$$I_i = (z_i - \bar{z}) \sum_{j \in J_i}^{n} W_{ij}\left(z_j - \bar{z}\right)^2 \tag{2.17}$$

where J_i denotes the neighbourhood set of area i, and the summation in j runs only over those areas belonging to J_i, $\bar{z}$ denotes the average of these neighbouring observations.

It is evident that the sum of I_i for all observations i

$$\sum_{i=1}^{n} I_i = \sum_{i=1}^{n} (z_i - \bar{z}) \sum_{j=J_i}^{n} W_{ij}(z_j - \bar{z}) \tag{2.18}$$

is proportional to the global Moran statistic I given by Eqs. (2.7) and (2.8).

The moments for I_i under the null hypothesis of no spatial association can be derived using the principles outlined in Cliff and Ord (1981, pp. 42–46). For example, for a randomisation hypothesis, the expected value is found as

$$E[I_i] = -\frac{1}{(n-1)} \tilde{W}_i \tag{2.19}$$

and the variance turns out to be

$$var\,[I_i] = \frac{1}{(n-1)} W_{i(2)}(n - b_2) + \frac{2}{(n-1)(n-2)} W_{i(kh)}(2b_2 - n) - \frac{1}{(n-1)^2} \tilde{W}_i^2 \tag{2.20}$$

where

$$W_{i(2)} = \sum_{j \neq i}^{n} W_{ij}^2 \tag{2.21}$$

$$2W_{i(kh)} = \sum_{k \neq i}^{n} \sum_{h \neq i}^{n} W_{ik} W_{ih} \tag{2.22}$$

$$\tilde{W}_i = \sum_{j=1}^{n} W_{ij} \tag{2.23}$$

with $b_2 = m_4 m_2^{-2}$, $m_2 = \Sigma_i (z_i - z)^2 n^{-1}$ as the second moment, and $m_4 = \Sigma_i (z_i - z)^4 n^{-1}$ as the fourth moment. A test for significant local spatial association may be based on these moments, although the exact distribution of such a statistic is still unknown (Anselin 1995, p. 99).

Alternatively, a conditional random permutation test can be used to yield so-called *pseudo significance levels*. The randomisation is conditional in the sense that the value z_i associated with area i is hold fixed in the permutation, and the remaining values are randomly permuted over the areas. For each of these resampled data sets, the value of the local Moran I_i can be computed. The resulting empirical distribution function provides the basis for a statement about the extremeness or lack of extremeness of the observed statistic I_i, relative—and conditional on—the I_i-values computed under the null hypothesis.

A complicating factor in the assessment of significance is that the statistics for individual locations (areas) will tend to be correlated whenever the neighbourhood sets J_i and J_k of two areas i and k contain common elements. Due to this

correlation, and the associated problem of multiple comparisons, the usual interpretation of significance will be flawed. Furthermore, it is impossible to derive the exact marginal distribution of each statistic, and the significance levels have to be approximated by Bonferroni inequalities or following the approach suggested by Sidák (1967). This means—as pointed out by Anselin (1995, p. 96)—that when the overall significance associated with the multiple comparisons (correlated tests) is set to α, and there are m comparisons, then the individual significance α_i should be set to either α/m (Bonferroni) or $1-(1-\alpha)^{1/m}$ (Sidák). Note that the use of Bonferroni bounds may be too conservative for local indicators of association. If, for example, $m = n$, then an overall significance of $\alpha = 0.05$ would imply individual levels of $\alpha_i = 0.0005$ in a data set with one hundred observations, possibly revealing only very few if any significant areas. But since the correlation between individual statistics is due to the common elements in the neighbourhood sets, only for a small number of areas k will the statistics actually be correlated with an individual I_i (Anselin 1995, p. 96).

Using the same notation as before, a local Geary statistic c_i for each observation $i\,(i = 1, \ldots, n)$ may be defined as

$$c_i = \sum_{j \in J_i}^{n} W_{ij}\left(z_i - z_j\right)^2 \tag{2.24}$$

where J_i denotes the neighbourhood set of area i. The c_i statistic is interpreted in the same way as the local Moran. The summation of the c_i over all observations yields

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \sum_{j \in J_i}^{n} W_{ij}\left(z_i - z_j\right)^2 \tag{2.25}$$

that is evidently proportional to the global Geary c statistic given by Eq. (2.9).

These LISA statistics, I_i and c_i, serve two purposes. On the one hand, they may be viewed as indicators of local pockets of non-stationarity, or hot spots, similar to the G_i and G_i^* statistics. On the other hand, they may be used to assess the influence of individual locations (observations) on the magnitude of the corresponding global spatial autocorrelation statistic, Moran's I and Geary's c.