Centre of Full Employment and Equity

Working Paper No. 01-13

Introduction to Spatial Econometric Modelling

William Mitchell[1]

April 2013

Centre of Full Employment and Equity
The University of Newcastle, Callaghan NSW 2308, Australia
Home Page: http://e1.newcastle.edu.au/coffee
Email: coffee@newcastle.edu.au

1. Introduction

Typically social and economic data have some spatial dimension. Unemployment is recorded by local unemployment agencies and tracked by government, accumulations of hazardous waste occur in proximity to specific human populations, crimes are committed at a location, consumers purchase goods at stores located in certain places and social inequality is spatially-situated. While human geography and urban planning have a long tradition of observing the spatial patterns of various phenomena and using these to develop and test explanatory models of social interaction and urbanisation (Frank, 2003: 147), conventional procedures of social data analysis, particularly in economics, often do not make use of this important locational information. The urban sociologists (and later the criminologists) at the Chicago school (Park *et al.*, 1925; Park, 1936 and Hawley, 1950), from which theories of human ecology evolved, stressed that social facts are located facts, situated in time and place and that social life cannot be fully understood "without understanding the arrangement of actors at particular social times and places" (Abbott, 1997: 1152). In sociology and economics there has been a renewed interest in models of social interaction and dependence among economic agents (Wilson, 1987; Durlauf, 2003; Akerlof, 1997), spatial spillovers (Topa, 2001), knowledge externalities and agglomeration economies (Banerjee, 1992; Krugman, 1991). In such models, information about the location of economic agents is essential to correctly predict the nature and magnitude of outcomes generated (a summary of these developments can be found in Goodchild *et al.*, 2000: 141).

Statistically, it is traditional for regionally based cross-sectional data to be viewed as conceptually identical to cross-sectional data on individuals or businesses at a single location. However, spatially adjacent observations are likely to exhibit spatial interdependence, owing to dynamics (knowledge flows, trade flows, commuting and social spill-overs), which accompany proximity. This emerging consensus begins with Tobler's (1970) maxim that 'everything is related to everything else but near things are more related than distant things'.

Spatial autocorrelation refers to the formal measure of the extent near and distant things are related. There are three types of spatial autocorrelation:

1. Positive spatial autocorrelation occurs when features that are similar in location are also similar in attributes;
2. Negative spatial autocorrelation occurs when features that are close together in space are dissimilar in attributes; and
3. Zero autocorrelation occurs when attributes are independent of location (that is, the observations are equivalent to a standard cross-sectional dataset).

Ignoring dependence between neighbouring regions will lead to biased regression results (Anselin, 1988). The problem is that in many cases direct analysis of the interactions between regions is not possible, due to the scarcity of data, and this "requires us to apply a method that allows us to analyse the effects of spatial interaction without quantitative information on the different linkages between labour markets" Niebuhr (2002: 5). In the last twenty years, a range of spatial regression techniques have been developed to measure latent forces of interaction and handle data that violate standard statistical assumptions of independence (Cliff and Ord, 1981; Anselin, 1988). These are estimated using maximum likelihood techniques and include: the general, mixed (SAR), spatial Durbin and first-order (FAR) auto-regressive models.

This Chapter is based on a series of lectures given during the annual ARCRNSISS Summer School between 2006 and 2008, which introduced spatial econometric techniques to students who were assumed to have no background in regression analysis. The material used in the lectures and reproduced here owes a great debt to the work of Luc Anselin. Much of it is drawn from his 1988 book and related materials. The demonstration dataset is Anselin's famous data covering 49 neighbourhoods in Columbus, Ohio data. We use a combination of the open source software GeoDa, produced by Anselin's team at the University of Arizona and the **spdep** package in R (Bivand, 2006) to generate the maps and the statistical results. The lectures also were significantly aided by the work of James Le Sage, who produced an open source 1999 technical manual with supporting MatLab scripts that allowed many younger scholars to enter the field of spatial econometrics.

The aim of the Chapter is to acquaint students with the concept of spatial autocorrelation in regionally based data (such as Census data); to introduce spatial econometric techniques and their theoretical underpinnings, special cases and regression diagnostics. We will consider the issues surrounding the creation of a weight matrix - which captures the degree of inter-relationship between regions in the system.

2. Hypothesis development

In applied analysis, we typically will be motivated by some problem to be solved for which we form a conjecture or hypothesis. The steps in the analysis are as follows:

- Construct a problem and consider the extant literature to demarcate where hypothesis development might lead to new knowledge in the field of interest.
- Make the enquiry operational by forming some conjectures.
- Gather relevant data, which might become an iterative process with the hypothesis development given that we frequently encounter the situation that what we might be interested in examining cannot be studied in a statistical manner because there is no available data of a sufficient standard.
- Determine the statistical and econometric techniques that are most appropriate given the nature of the problem and the data available.
- Conduct exploratory data analysis to determine the basic characteristics of the data. In the spatial context, this involved examining for spatial correlation.
- Conduct the spatial modelling, which involves several steps:
  - Choosing the specification (the form of the regression to be estimated);
  - Choosing the appropriate estimation techniques;
  - Confronting the regression estimates with a battery of tests to ensure they have desirable properties, which permit inference (hypothesis testing); and
  - Interpreting the estimates, if statistically acceptable.
- Reconsider hypotheses in the light of the results. This may involve a modification to reflect the information that has been generated by the modelling.

To aid our understanding of hypothesis development, we now consider the demonstration dataset.

3. The Columbus Dataset

The Columbus, Ohio dataset covers 49 contiguous Planning Neighbourhoods in 1980 (Anselin, 1988). Anselin (1988: 187) says, "these neighbourhoods correspond to census tracts or aggregates of a small number of census tracts, and are representative of the type of data used in many empirical urban analyses."

Three variables of interest in the dataset include:

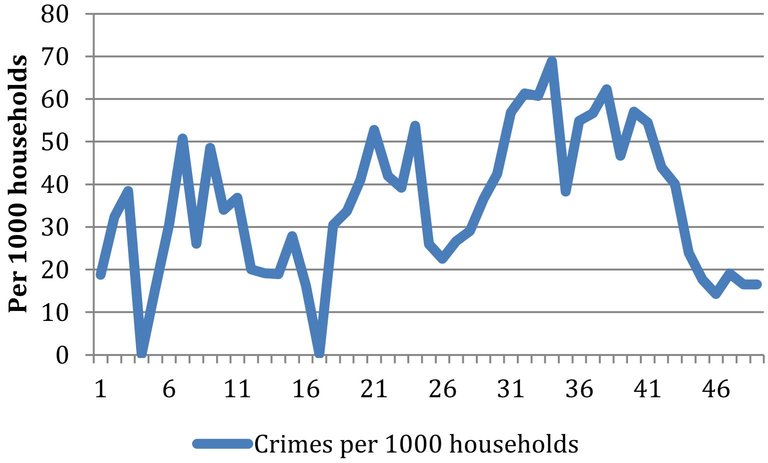
- Crime - residential burglaries and vehicle thefts per thousand households in the neighbourhood (CRIME)
- Household income in units of $1,000 (INC)
- House values in units of $1,000 (HOVAL)

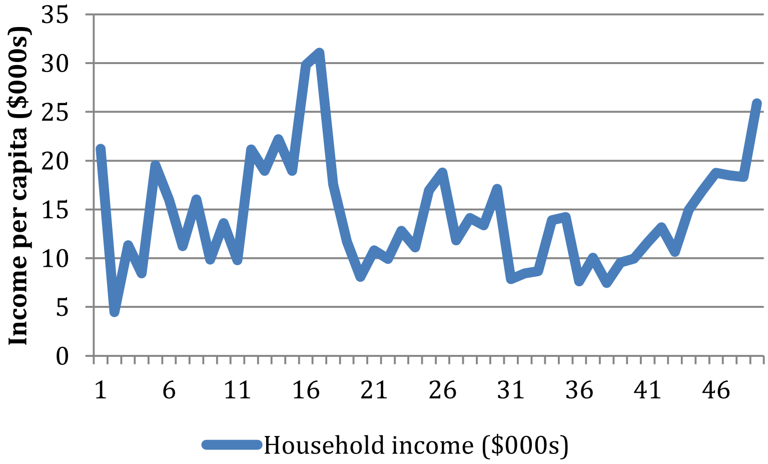
All 49 observations are also geographically located by their latitude coordinate (LAT) and the corresponding longitude coordinate (LONG). We use the newer version of this dataset where the order of the neighbourhoods was changed.

It is always advisable to visualise the data prior to more formal analysis. Any outliers that you see in the graphs should make sense (for example, they might be due to a known change in behaviour, or a policy change). Graphs also allow you to quickly appreciate some of the key characteristics of the data.

Figure1 plots the crime rate and household income for the 49 neighbourhoods, which are arranged in the order they appear in the database on the horizontal axis. A quick visual examination would suggest that where income levels are high, the crime rate is low and vice versa.

Figure 1 Residential burglaries and vehicle thefts per thousand households and Income Per Capita ($000s), Columbus, Ohio, 1980

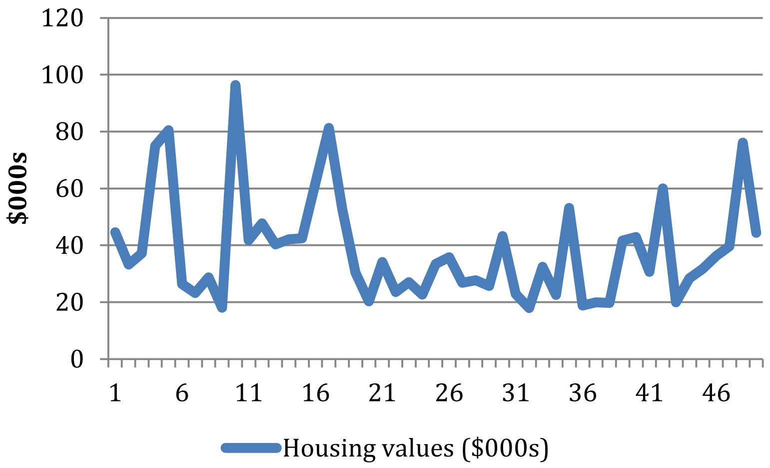




Source: Anselin (1988).

Figure 2 shows the housing values for each neighbourhood. When juxtaposed with the information in Figure 1, it appears that in neighbourhoods where the housing values are high, there is a lower crime rate and higher household incomes. This makes intuitive sense, in that higher income individuals will tend to reside in higher quality housing and more expensive neighbourhoods, have less need to steal, and will devote more resources to home security and theft protection.

Figure 2 Housing values, Columbus, Ohio, 1980, $000s



Source: Anselin (1988).

Figure 3 crime against income (Panel A) and housing values (Panel B). The lines are simple linear regression with the dependent variable being crime and a constant term included. The graphs support the view that the relationship between crime and income is likely to be strong and negative. The relationship between housing values and crime is also negative but less clustered around the regression line than we see in Panel A.

On this basis, we might thus form the following hypothesis:

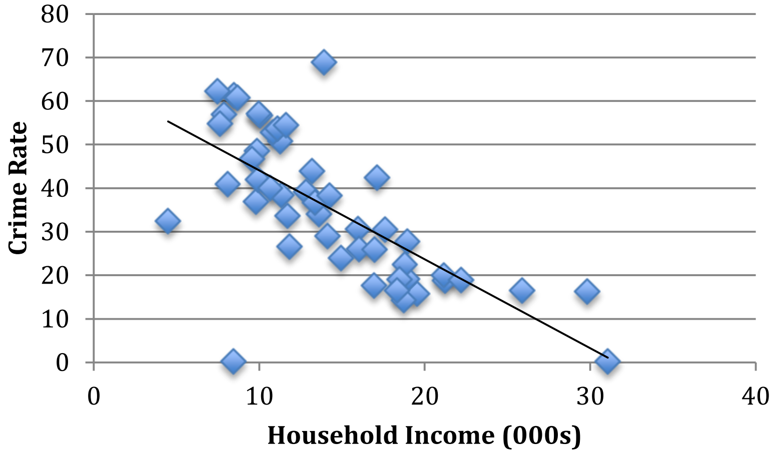
> Thefts in any neighbourhood will be a negative function of income and housing value.

This means that if the hypothesis has explanatory power we would expect the crime rate to be higher in neighbourhoods with lower levels of household income and higher housing values.

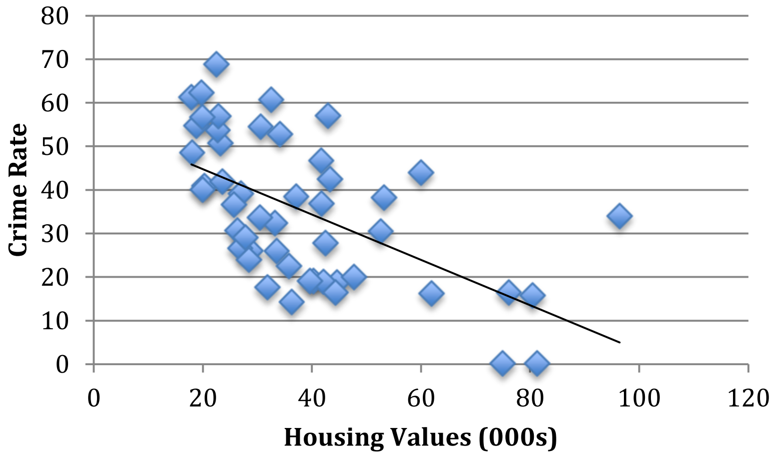
Clearly there would be a deep literature in criminology that would provide motivation for this type of enquiry. These hypotheses are just illustrative and designed to advance the discussion. The important point is that theory always underpins conjecture and empirical analysis. The other cautionary note is that we should not see econometric analysis (for example, regression) as an exercise in truth discovery. Regression results do not prove anything. In some cases, we might be able to say that the estimates we generate provide tentative support for our theoretical conjectures. The regression results can also expose conjectures that are clearly not supported by the data.

However, in spatial data, traditional plots might obscure some of the spatial dependencies and further insights can be gained through visualisation via maps. A popular way of examining spatial patterns is to use ranking or percentile maps, which highlight extreme values, which are defined as observations in the top and bottom one per cent of the distribution. We might consider these extreme values to be outliers although there is no statistical significance attached to the ordering presented.

Figure 3 Cross plots income and crime, housing values and crime, Columbus, Ohio, 1980



Panel A



Panel B

Source: Anselin (1988).

Figure 4 shows the percentile map for crime with the observations ranked into six categories, 0-1 per cent; 1-10 per cent; 10-50 per cent; 50-90 per cent; 90-99 per cent, and 99-100 per cent. The numbers in brackets represent frequency of observations in each band. The high to very high crime rate neighbourhoods are clustered within the inner-city area with the low crime rate regions clustered together around the periphery of the city. Two neighbourhoods are in the top and bottom percentiles of the distribution.

Figure 5 shows the percentile map for household income in Columbus, Ohio in 1980. The lower income households are clustered together, with some exceptions in the inner-city neighbourhoods, while the higher income households are spatially distributed on the periphery of the city.

Figure 4 Percentile Map, Crime, Columbus, Ohio, 1980

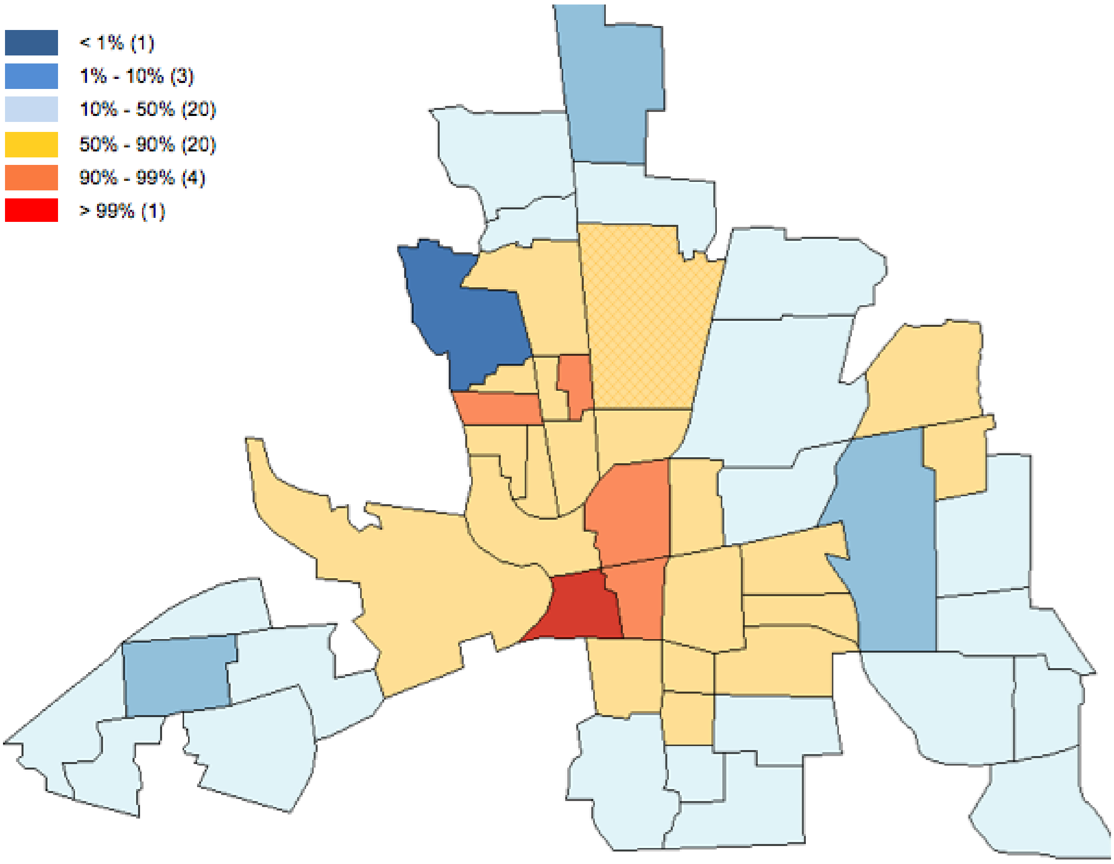
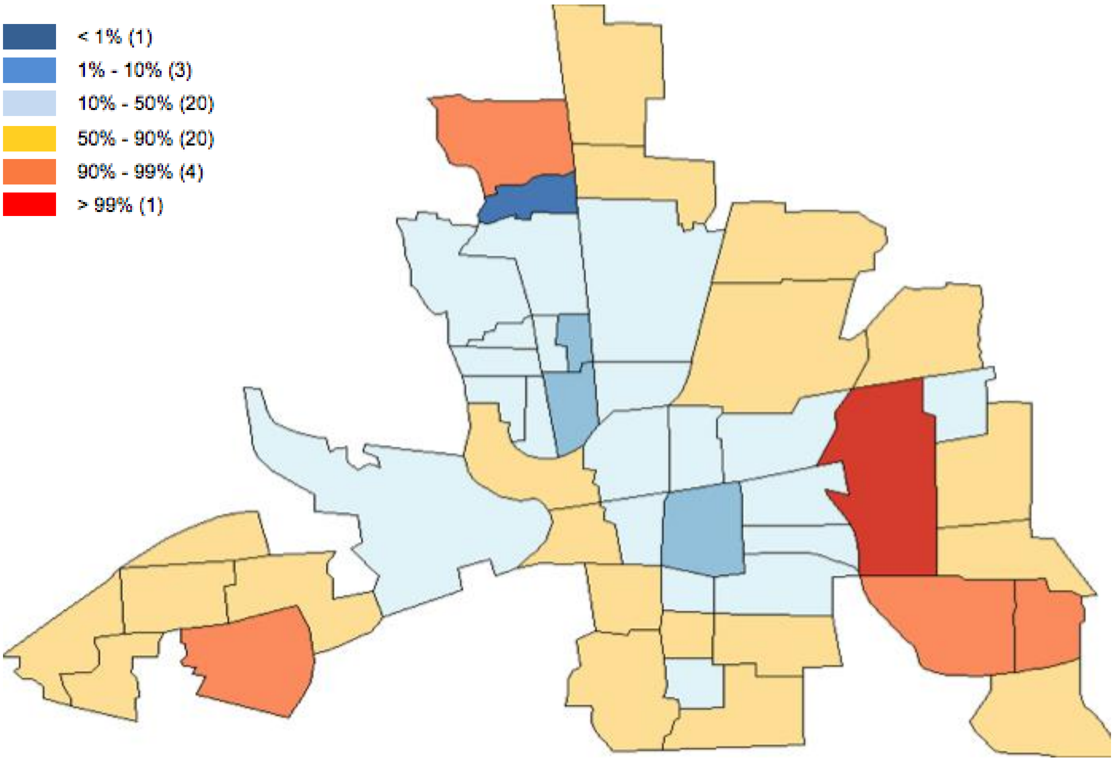
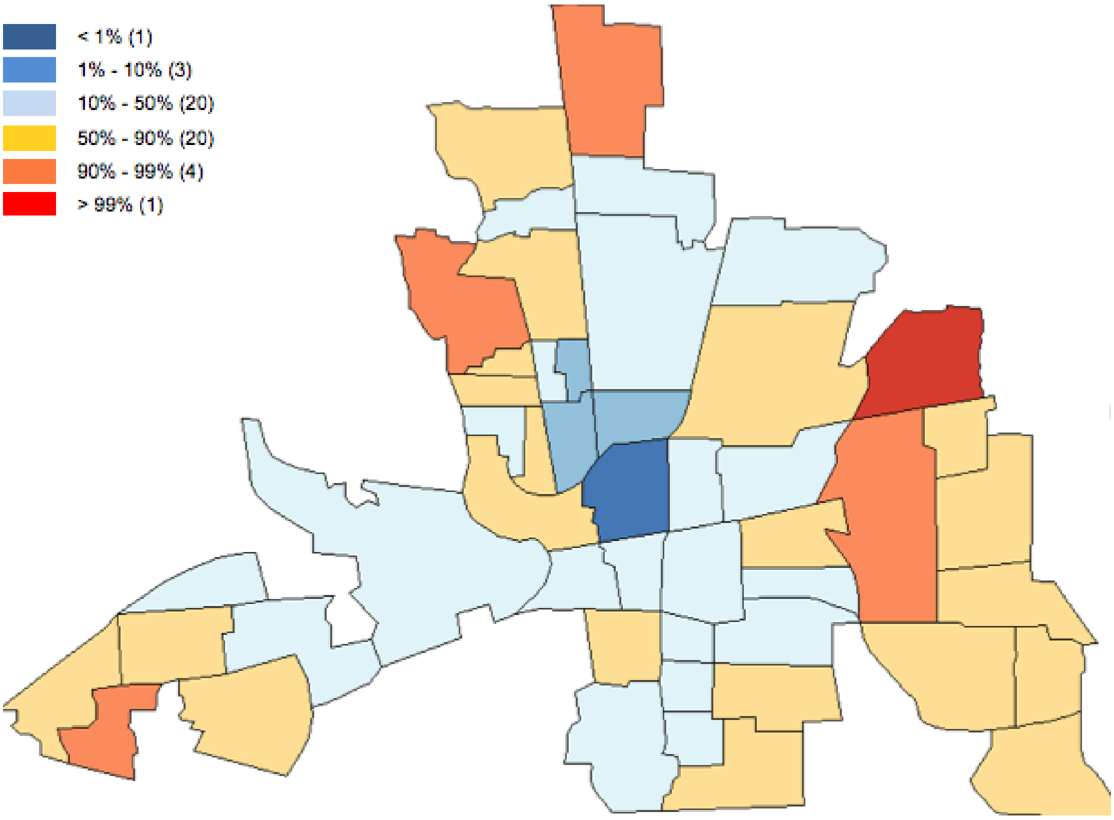


Figure 5 Percentile map for income, Columbus, Ohio, 1980



Finally, Figure 6 shows the percentile map for housing values in Columbus, Ohio in 1980. Consistent with the spatial information shown for household income, the higher valued houses are located in clustered patterns on the periphery of Columbus and the lowest value houses are concentrated increasingly towards the inner city.

Figure 6 Percentile map for housing values, Columbus, Ohio, 1980



Taken together, the percentile maps indicate spatial clustering in the three variables of interest, which may present econometric problems if we ignore the likelihood of spatial autocorrelation. While the percentile maps do not relate any information about the statistical significance of the clustering, the visual patterns alert us to the possibility that these observations should not be treated as being independent, as would be the case in a standard cross-sectional analysis.

4. Ordinary Least Squares regression

Ordinary least squares (OLS) regression is the most widely used modelling method in econometrics and is the first estimation technique that is taught. When people say they have 'run a regression' or fitted a 'linear regression' or 'used least squares' they usually are saying they have used OLS to fit a model to their data. Fitting an equation means that we have used data to quantify an algebraic relationship between variables, with unknown coefficients.

We construct regression models in terms of a dependent variable (CRIME) being "explained" by the explanatory or independent variables (in this case household income and housing values).

We usually express our hypothesis in the form of an algebraic model as follows:

(1) $$CRIME_i = a + b_1 INC_i + b_2 HOVAL_i + e_i$$

In this dataset, they are measured for the i^{th} region hence the subscript. The regression has a constant term, which need not concern us here. Each explanatory variable is multiplied by an unknown parameter or **coefficient** (β_1 and β_2). Together with the constant, the two independent variables comprise the deterministic part of the equation. The task of regression is to quantify these unknown parameters, which define the way in which household income (INC) and housing values (HOVAL) impact on crime.

You will notice an additional parameter ε_i, which is called the **error term** or the stochastic component of the regression. While our hypothesis suggests household and income explain the crime rate in each of the *i* regions, we also accept the dependent variable (CRIME) will have random components. The error term captures the variation in CRIME not explained by INC and HOVAL. There is an error term for each of the 49 observations.

OLS regression makes certain assumptions about the properties of the error term.

- Individual observations are not correlated with each other;
- The errors have a finite variance – so you don't get increasing errors as the value of the explanatory variable rises, for example;
- The errors have a zero mean – so for any 'region' the error is expected to be zero and the expected value of the dependent variable is given by the deterministic part of the equation.

After estimating the equation (finding values for the unknown coefficients and the errors), we would then use the results to progress our analysis in a number of ways. First, we would seek to test the statistical significance of the coefficients, which means we want to be sure that the estimates are not, in fact, indistinguishable from zero. If they were statistically equivalent to zero, then we would conclude there was no relationship between that particular explanatory variable and the dependent variable. Second, we might use the estimated equation to make forecasts (predictions) about the crime rates. As long as the estimates were statistically sound, we might advise the government that if they increased household incomes through job creation programs or social transfers for those households with low income, the crime rate in those neighbourhoods might drop by x per cent. The scale of the policy intervention (the increased income) could then be scaled to match the targeted drop in the crime rate in specific neighbourhoods.

How does OLS provide numerical estimates of the unknown coefficients? There are many ways in which we could produce these numerical estimates. The name OLS refers to a specific way the estimates of the unknown parameters are computed. The practice of regression or 'fitting a line' is the process by which the unknown parameters are picked so that the line 'best' represents the data. What criterion is used to determine 'best'? In OLS, the unknown parameters are estimated by **minimising the sum of the squared residuals** between the data and the model.

The theoretical model (Equation 1) becomes an estimated model, when the unknown coefficients and error terms are quantified. The estimated error terms in the theoretical model (Equation 1) are called **residuals**. The estimated model corresponding to Equation (1) is written as such:

(2) $$\widehat{C}_i = \widehat{a} + \widehat{b}_1 INC_i + \widehat{b}_2 HOVAL_i + \widehat{e}_i$$

where $\widehat{C}_i$ is the estimated value of CRIME in the i^{th} neighbourhood. The hats (^) indicate the parameter is estimated or fitted.

We can thus express the residuals of the estimated model in this way:

$$
\begin{aligned}
\widehat{e}_i &= CRIME_i - \widehat{C}_i \\
&= CRIME_i - \left[\widehat{a} + \widehat{b}_1 INC_i + \widehat{b}_2 HOVAL_i\right]
\end{aligned}
\qquad (3)
$$

If you examine the right-hand side of this expression, you will see that the residual for the i^{th} neighbourhood is the difference between the actual value ($CRIME_i$) in that neighbourhood and the estimated or fitted value, which is the term in the brackets.

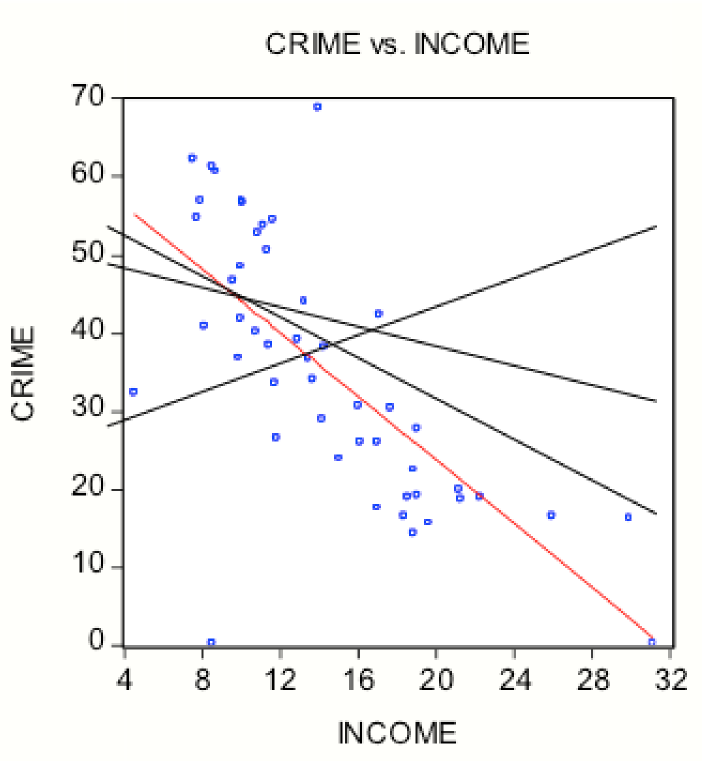
We define the residual sum of squares (RSS) as:

$$
RSS = \sum_{i=1}^{N} \hat{e}_i^2 = \sum_{i=1}^{N} \left[CRIME_i - \left(\hat{a} + \hat{b}_1 INC_i + \hat{b}_2 HOVAL_i\right)\right]^2 \qquad (4)
$$

So the residuals for each observation are squared and then summed overall for N observations. N is the case of the Columbus data set is equal to 49.

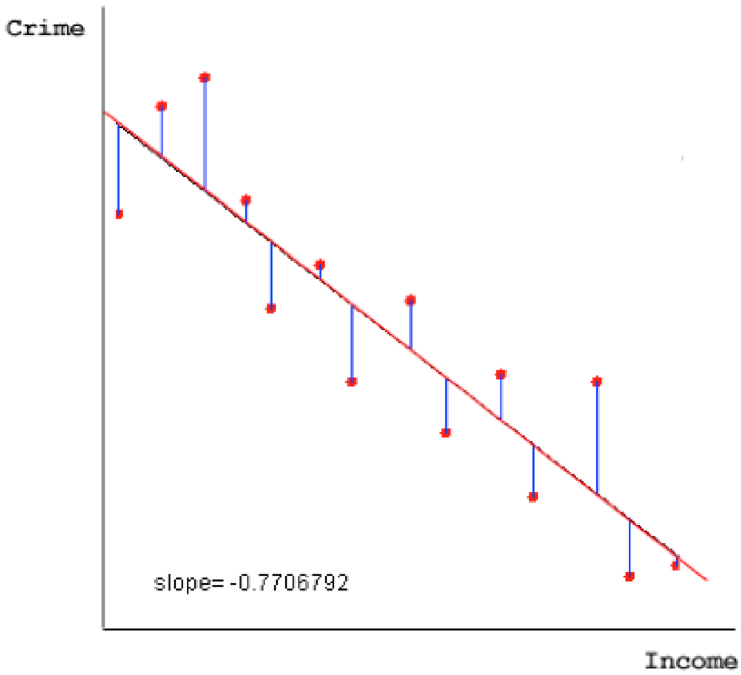
The OLS regression technique selects the values of the unknown parameters such that the RSS is minimised. The estimated parameters define a regression line, which relates the explanatory variable to the dependent variable. In a bi-variate model (only one explanatory variable and a constant), the regression line would be equivalent to the trend line that Excel draws on a scatter plot. In the case of a simple bi-variate model, relating crime to household income as captured in Figure 7, the best line is the red OLS regression line. The other lines would not minimise the RSS for this data.

Figure 7 OLS versus other possible fitted lines.



To cement these ideas more strongly, Figure 8 presents a stylised set of observations for crime and household income (not taken from the Columbus dataset). The observations are shown as red dots. The red line shows the OLS fitted line. The residuals are the blue vertical drop-lines from the actual value (the red observations) to the fitted line. The squared sum of the blue drop-line values is the RSS and OLS minimises that number.

Figure 8 The OLS regression line



What are the properties of the OLS estimator? OLS is known as the Best Linear Unbiased Estimator (BLUE), which means that our inference (hypothesis testing and prediction) will be based on sound statistical principles. OLS estimates of the unknown coefficients are **unbiased** and **best** (efficient).

- Unbiased - an estimate is unbiased if, in repeated trials, it misses in one direction with the same propensity as it misses in the opposite direction. Unbiased estimates have errors, but the errors are distributed around the true value of the dependent variable.
- Best or efficient estimates - the typical error of an estimate is called the standard error. The estimator with the smallest standard error of all possible estimators is called the 'best' estimator or the 'efficient' estimator. The 'best' estimator will produce estimates that are more tightly grouped around the true value of the variable that it is estimating than will any other estimator. Estimators with small standard errors are desirable not the least because they allow for more precise hypothesis tests.

However, these properties only hold if the OLS assumptions relating to the error terms hold. For example, when these assumptions fail, then OLS can produce biased estimates. There are many other issues that arise when conducting regression analysis (choosing functional forms; measurement errors, omitted variables; outliers; etc). For the purposes of this Chapter, we are concerned only with the properties of the residuals relating to their independence.

5. Traditional Econometrics and Spatial Econometrics

Empirical work in regional science uses sample data that is clearly location-specific. Traditionally, regional cross sectional data are viewed as being conceptually identical to cross-sectional data on individuals or businesses at a single location. However, as outlined in the Introduction, spatially adjacent observations are likely to exhibit spatial interdependence, owing to dynamics (such as knowledge flows, trade flows, commuting and social spill-overs), which accompany proximity. Two problems arise in this situation:

- Spatial dependence between the observations;
- Spatial heterogeneity occurs in relationships we are modelling.

These two problems define the ambit of spatial econometrics. In this Chapter we are concerned with the first of these problems, as an introduction to the field of spatial modelling.

OLS regression is unsuitable when there is spatial dependence between the observations (Anselin, 1988). The consequences of ignoring this dependence vary with the type of spatial dependence that is present. We will see that if we ignore what is referred to as the **spatial lag**, we encounter an omitted variable problem (that is, the regression equation excludes an important explanatory variable) and OLS produces biased and inconsistent estimates. If on the other hand we ignore the **spatial error**, we encounter an efficiency problem. In this case, the OLS estimates are unbiased but inefficient, which means that the OLS standard errors and related t-test statistics are biased. The presence of either problem violates the OLS assumption that the sample observations are strictly independent drawings and renders the estimates generated unusable.

Spatial econometrics techniques have been developed to specifically cope with situations where the data observations are spatially related. Spatial econometrics provides ways to modify the standard OLS regression approach to overcome the problem of spatial dependence.

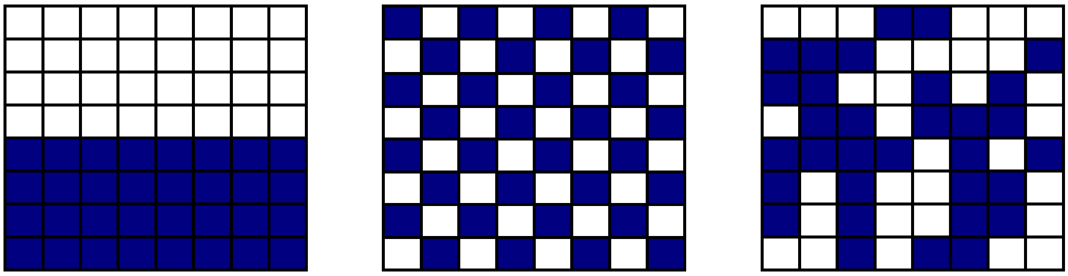
6. Spatial autocorrelation: what is it and how is it measured?

6.1 The concept of spatial autocorrelation

Spatial autocorrelation refers to the formal measure of the extent near and distant things are related. Figure 9, using raster representation, depicts the three types of spatial autocorrelation:

1. Positive spatial autocorrelation occurs when features that are similar in location are also similar in attributes;
2. Negative spatial autocorrelation occurs when features that are close together in space are dissimilar in attributes; and
3. Zero autocorrelation occurs when attributes are independent of location.

Figure 9 Stylised patterns of spatial correlation



(a) Positive spatial correlation (b) Negative spatial correlation (c) Zero spatial correlation

Source: Longley *et al* (2001).

There are two broad reasons proposed as to why spatial dependence may exist between regions. First, data collected on observations associated with spatial units such as used in the ABS Australian Standard Geographic Classification may contain measurement error because the administrative boundaries for data collection do not reflect the underlying processes generating the sample data (Anselin, 1988: 11-12). If social or economic behaviour crosses geographic boundaries we would expect to see very similar results amongst neighbouring regions. For example, mobile workers can cross boundaries to find employment in neighbouring areas, and thus labour force or unemployment measures based on where people live could exhibit spatial dependence.

Second, location and distance are important forces at work in human geography and market activity. For example, labour market outcomes (such as unemployment rates) between neighbouring regions might be clustered because of spatial pattern of employment growth (demand) or the distribution of population characteristics such as job skills (supply), and some mismatch between them. Further, housing has clear spatial dimensions, which may contribute to the clustering of unemployment rates as disadvantaged workers seek cheaper housing (O'Connor and Healy, 2002; Hulse *et al.*, 2003). Mobility patterns are important in determining the extent of spatial dependence between regions. European empirical evidence points to the strong effects of distance as an obstacle to migration. Migration is significantly reduced as distance increases because the costs of moving rise and the benefits from migration become increasingly unknown (Helliwell, 1998; Tassinopolous and Werner, 1999). Spatial impacts can also occur independently of employment patterns, population characteristics and housing patterns due to the functioning of social networks and neighbourhood effects (Borland, 1995; Topa, 2001).

6.2 Representing spatial dependency with spatial weight matrices

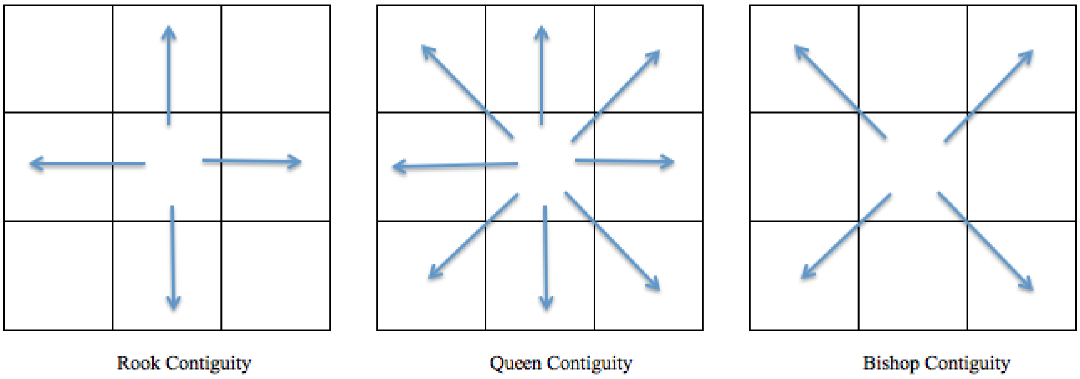
To consider the impacts of spatial dependence we need to be able to quantify "location" in our data. There are a variety of ways in which we can capture the spatial patterning between observations in our dataset. Stetzer (1982: 571) discusses various criteria that might be used, including "connectivity, contiguity, length of common boundary between political units, and various distance decay functions" (see also Hordijk, 1979; Anselin, 1988). For our purposes, we can view the spatial data in terms of **contiguity**, which requires knowledge of the shape and size of the regional unit being observed. Contiguity considers neighbourhood proximity in terms of a shared common border. What do we mean by a shared common border?

There are three types of contiguity patterns, which take their name from the game of Chess (see Figure 10):

- **Rook** contiguity where regions share common sides. In this case, Regions 1 and 2 would be classified as being rook contiguous
- **Queen** contiguity where regions share either a common side or a common vertex. In this case, Regions 4 and 5 meet these criteria and are thus queen contiguous.
- **Bishop** contiguity where regions share a common vertex. Regions 2 and 3 are bishop contiguous.

In our case we will only be concerned with first-order contiguity, which means the regions must have common borders. More complex patterns of *n*-contiguity are sometimes used in spatial analysis.

Figure 10 Contiguity patterns for shared borders



If we take first-order rook contiguity as an example, how might we formalise that spatial relationship in a regression context? In other words, we need a method of quantifying the proximity. First-order contiguity uses **binary** connectivity such that we define an element $w_{ij} = 1$ if regions i and j are contiguous and zero otherwise. In this schema, our concept of connectedness is either "on" or "off". The alternative is to use distance functions that taper as the distance between regio ns increases.

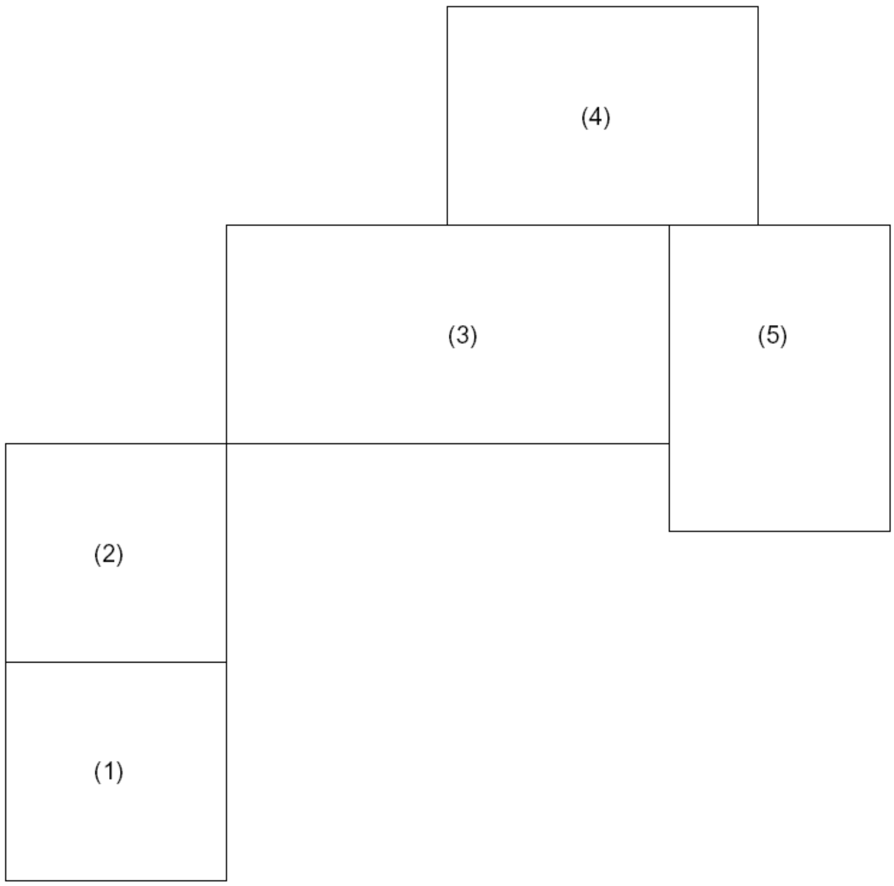
As an example, consider the 3 regions in Figure 11. This square matrix has 3 rows and 3 columns, 9 cells in all. The three main diagonal cells represent each region's relationship with itself while the off-diagonal cells represent the pair-wise relationships. Under rook contiguity, the cells representing regions that are rook-connected are denoted 1 while those that are no connected with such a common border are given the value of 0. We thus expect, for example, regions 1 and 2 to exhibit mutually dependent behaviour but regions 1 and 3 are considered independent of each other. As we will see, this spatial patterning allows us to weight some observations in the regression exercise more than others to reflect the hypothesised spatial dependence. It is for this reason that we refer to such a matrix as a **spatial weights matrix**. This simple quantification of the expected spatial dependence generalises to any number of regions and spatial arrangements. T

Figure 11 Stylised Rook Contiguity and Spatial Weights

Region	1	2	3
1	0	1	0
2	1	0	1
3	0	1	0

Figure 12 presents a more complex map with 5 regions in various states of contiguity. It allows us to generalise the concept of a spatial weights matrix. If we continue with rook contiguity as our representation of connectedness, how might we formalise the spatial relationship shown in Figure 12?

Figure 12 Contiguity Patterns in Spatial Data



In general terms, we define a matrix $\mathbf{W}(n, n)$ where n is the number of regions. There are five regions in Figure 12, so n = 5. Each element in **W** denotes a specific pair-wise relationship between the five regions, that is, whether they are rook contiguous or not. We give an element in **W** the value of 1 if the two regions that the element represents are rook contiguous and zero otherwise. The elements on the main diagonal denote a regions relationship with itself so we give it the value 0. The off-main diagonal elements capture the pattern of spatial dependency assumed.

The element W(2,1) denotes the relationship between Region 2 and Region 1. Are they rook contiguous? If yes, we would declare that element to be 1 if no, the element would be assigned the value of zero. In the case of Figure 10, we can see that Regions 1 and 2 are indeed rook contiguous so we assign W(2,1) = 1. You will note that element W(1,2) is identical to W(2,1), which makes sense because the **W** matrix is symmetrical. The rook contiguous regions are:

- Regions 1 and 2
- Regions 3 and 4
- Regions 3 and 5
- Regions 4 and 5

The resulting spatial weight or pattern matrix, **W** is thus given as:

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

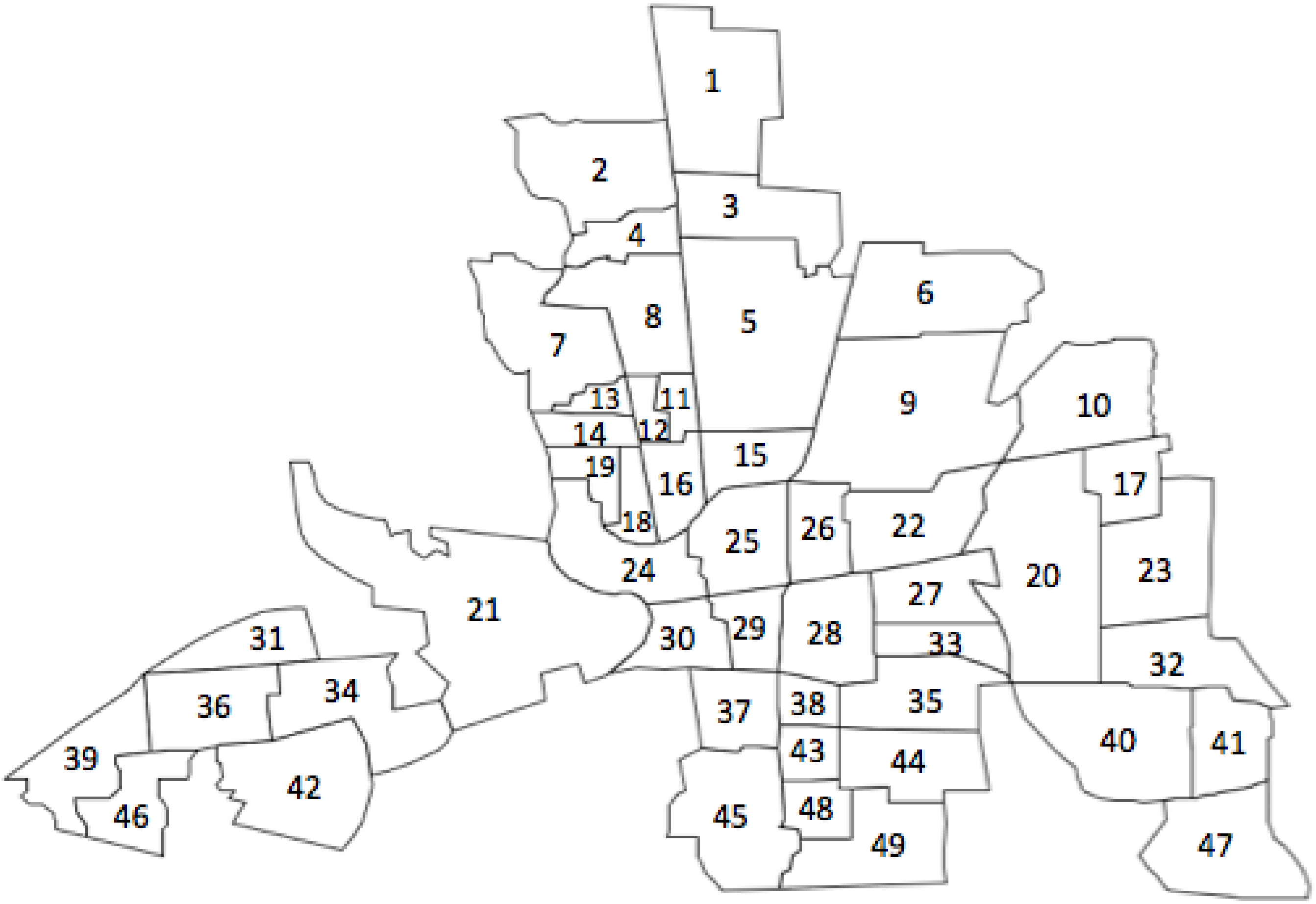
The spatial weight matrix captures in a quantitative manner the spatial patterning based on contiguity, which we posit leads to spatial dependence. Stetzer (1982: 571) notes that spatial weight matrices represent “a priori knowledge of the strength of the relationship between all pairs of places in the spatial system.” The weights are analogous to lag coefficients in autoregressive-distributed lag time series models. Unlike in time-series data where data points are ordered contemporaneously determining the order of observations in space is difficult as it is multidirectional. A ‘spatial order’ is typically imposed in a more or less *ad hoc* fashion. Thus, estimation of spatial autocorrelation is sensitive to the weights employed and the weights embody assumptions about the spatial structure (Molho, 1995: 649). A discussion of this issue is beyond the scope of this Chapter.

In practice, the spatial weight matrix is row standardised. Row standardisation creates proportional weights in situations where a region has an unequal number of neighbours. In other words, the row-standardised weights increase the influence of likely spill-overs where a region has few neighbours relative to regions where spill-overs will occur between many neighbours.

6.3 Spatial contiguity in the Columbus, Ohio dataset

Figure 13 numbers the 49 neighbourhoods in the Columbus, Ohio dataset to help us establish the spatial dependence assumed based on first-order rook contiguity. Table 1 summarises all the first-order rook contiguous cases for the 49 Columbus neighbourhoods.

Figure 13 Columbus, Ohio neighbourhoods



Source: Anselin (1988) Columbus, Ohio dataset (new reordered version).

Table 1 First-order rook contiguity pattern for Columbus Neighbourhoods

Neighbourhood	First-Order Contiguous to							
1	2	3						
2	1	3	4					
3	2	4	5					
4	2	3	5	8				
5	3	8	5	8	9	11	15	
6	5	9						
7	8	13						
8	4	5	7	11	12			
9	5	6	10	15	22	26		
10	9	17	20					
11	5	8	12	16				
12	8	11	13	14	16			
13	7	12	14					
14	12	13	18	19				
15	5	9	16	25				
16	11	12	15	18	24	25		
17	10	20	23					
18	14	16	19	24				
19	14	18	24					
20	10	17	22	23	27	32	33	40
21	24	30	34					
22	9	20	26	27	28			
23	17	20	32					
24	18	19	21	25	30			
25	15	16	24	26	29			
26	9	22	25	28				
27	20	22	28	33				
28	22	26	27	29	33	35	38	
29	25	28	30	37				
30	21	24	29					
31	34	36						
32	20	23	40	41				
33	20	27	28	35				
34	21	31	36	42				
35	28	33	38	44				
36	31	34	39	42	46			
37	29	30	38	43	45			
38	28	35	37	43				
39	36	46						
40	20	32	41	47				
41	32	40	47					
42	34	36						
43	37	38	44	45	48			
44	35	43	48	49				
45	37	43	48					
46	36	39						
47	40	41						
48	43	44	45	49				
49	44	45	48					

Source: Anselin (1988) Table 12.2.

7. Exploratory Spatial Data Analysis (ESDA)

7.1 Overview

The first use for the spatial weights matrix is in **Exploratory Spatial Data Analysis** (ESDA), which is a set of techniques aimed at visualising the spatial distribution of data, identifying 'atypical localisation', detecting patterns of spatial association, that is clusters or hot spots and cold spots, and suggesting the presence of different spatial regimes, where data provide evidence of heterogeneity (Anselin, 1996). It is preliminary to, but informs formal regression modelling.

Measures of spatial dependence or spatial autocorrelation are a way of evaluating the amount of clustering or randomness in the data. Unlike standard measures of concentration these measures impose an explicit geographic structure, which makes them capable of summarising clustering observed via visual inspection of a map and also capable of testing whether these clusters are significantly non-random.

In order to determine if the values of a particular mapped variable (such as the crime rate) deviate from a pattern that would exist if they were randomly assigned, we require an index of comparison. **Global measures** of spatial autocorrelation are such an index, providing evidence of the presence or absence of a stable pattern of dependence across our whole dataset. There is a debate over how spatial autocorrelation can be best characterised, which we will not engage in here (see Cliff and Ord, 1973; Upton and Fingleton, 1985).

Spatial statistics deploy spatial weight matrices, which formalise the level of interdependence between all pairs of regions in the system. Unlike in time-series data where data points are ordered contemporaneously, determining the order of observations in space is difficult as it is multidirectional and because it requires prior knowledge of the nature of dependence in the system. A ‘spatial order’ is typically imposed based on some prior assumption and the estimation of spatial autocorrelation is sensitive to this (Molho, 1995: 649). Following LeSage’s (2005) comments that the main aim of the weighting matrix is to incorporate some notion of proximity into standard statistical tests (and weighting structures that test complex assumptions about the nature of dependence can unnecessarily obscure the simple relationship embodied by Tobler’s Law) it is often reasonable to opt for the most simple conception of spatial interconnectedness, that is ‘first-order contiguity’ – neighbours defined on the basis of regions whose borders touch. In the following discussion we use rook contiguity as explained in Section 6.

Common **global** measures (measures which assess spatial association across a whole dataset) of spatial association include; Moran’s I, Geary’s C and Global G.

The global measures can be decomposed to provide **local measures of spatial association** - LISAs in the case of a Moran statistic Anselin (1995) and Local G in the case of the Getis-Ord statistic (Ord and Getis, 1995). These measures provide more detailed information on the type of spatial association present and indicate the contribution from each region to the overall spatial association. Where regions form a cluster of unusually high values this has been termed a ‘hotspot’. Similarly, a ‘coldspot’ refers to a group of neighbouring regions who exhibit unusually low values. Hotspots and coldspots represent islands of heterogeneity, uncharacteristic deviations from a national spatial pattern and thus are of interest to researchers and policy-makers.

In analysing geographically disaggregated data these measures represent a significant advance on standard measures of concentration (Theil, Gini, Coefficient of variation), because they explicitly incorporate spatial location. Most importantly spatial autocorrelation measures can evaluate the probability that concentrations of high and low values of some variable of interest occur by chance, thus the credibility of a purely visual interpretation of clustering (for instance via a choropleth map) is significantly advanced upon (Frank, 2003: 160).

7.2 Global autocorrelation measures – Moran's *I* statistic

A standard measure of global spatial autocorrelation is provided by Moran's *I* (Moran, 1948). The Moran's *I* statistic provides an indication of the degree of linear association between the observation vector (**x**) and a vector of spatially weighted averages of neighbouring values (**Wx**), where **W** formalises the neighbourhood or contiguity structure of the dataset.

The Moran's *I* ranges from minus one (indicating perfect dispersion) to plus one (indicating perfect correlation) with a zero value corresponding to a random spatial pattern (no spatial autocorrelation). Negative (positive) values thus indicate negative (positive) spatial autocorrelation.

The Moran *I* statistic is computed as:

$$(5) \qquad I_t = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} x_i x_j w_{ij}}{R_b \sum_{i=1}^{R} x_i^{\,2}}$$

where R is the number of regions, R_b is the sum of the weights and when the spatial weighting matrix is row-standardised it simplifies to R. The x is a variable of interest (say the crime rate) in region i (in deviations from the mean). The Moran *I* statistic is easily computed from the residuals of a regression on a constant.

The Moran's *I* statistic can be expressed as a standardised normal Z value for inference purposes. At the 5 per cent level of significance we reject the null of no spatial autocorrelation if the standardised Moran *I* statistic is greater than 1.96 or smaller than -1.96.

Table 2 reports the Moran's *I* statistics for crime, household income and housing values from the Columbus dataset, along with the standardised and *p*-values for each variable. The null hypothesis is that there is no spatial association. It is clear from the test statistics that the three variables exhibit strong spatial interaction and we can safely reject the null hypothesis.

Significant positive dependence is present in all variables using simple first-order rook contiguity as the weighting scheme. This means, for example, that it is much more likely that regions with high (low) crime rates (income or housing values) will have neighbours with high (low) crime rates (income or housing values) than if the distribution of crime, income and housing values in Columbus, Ohio in 1980 were purely random.

Table 2 Moran *I* statistics, Crime, Household Income and Housing Values (under normality)

	Moran *I* statistic	Standardised Deviate	*p*-value
Crime	0.52367021	5.4978	3.845e-08
Income	0.43191278	4.5714	4.846e-06
Housing Values	0.22425202	2.4746	0.01334

The presence of spatial interaction in data samples suggests the need to quantify and model the nature of the spatial dependence in more detail. In the Section 8, we outline the taxonomy of spatial econometric models that can more formally explore the spatial interaction between the variables. However, we can also use other ESDA measures to further examine the nature of the spatial dependencies in the data in terms of spatial clusters with positive or negative spatial autocorrelation and spatial outliers.

7.3 Local Measures of Spatial Association

While our dataset reveals a globally significant trend towards clustering, global measures of spatial autocorrelation offer only an 'average' and can hide interesting micro-concentrations. To overcome this limitation, Local Indicators of Spatial Association (LISAs) have been developed. These indicate if one or more confined areas exhibit substantial deviation from spatial randomness. Local measures are particularly useful in large datasets where spatial association between observations is likely to show instability in the form of "local non-stationarity, spatial regimes and spatial drift" (Anselin, 1996: 112). Anselin (2003: 99 says that "Local spatial autocorrelation analysis is based on the Local Moran LISA statistics … [which yield measures] … of spatial autocorrelation for each individual location".

There are several measures that have been developed to examine local spatial association. In this Chapter we will consider only two: (a) Moran Scatterplots; and (b) LISA cluster maps.

7.4 Moran Scatterplots

Local spatial instability is studied by means of the Moran scatterplot (Anselin, 1996*), which* plots standardised values of the spatial lag ($\mathbf{Wz}$) against the original values ($\mathbf{z}$). A linear regression line is added, which has Moran's I as its slope and can be used to indicate the degree of fit – the extent to which the regression line reflects the overall pattern of association between $\boldsymbol{Wx}$ and $\boldsymbol{x}$. Points that deviate from the Moran regression line may be important outliers or leverage points in the strong global spatial dependence observed for short and long-term unemployment.

The Moran scatterplot also provides further information on the type of spatial dependence, other than whether spatial association is positive or negative.

The four quadrants of the plot represent the four different types of spatial association:

- Upper-right quadrant – high-high, positive autocorrelation, clustering of high values.
- Lower-right quadrant – high-low, negative autocorrelation, outlier, high value among low neighbours.
- Lower-left quadrant – low-low, positive autocorrelation, clustering of low values.
- Upper-left quadrant – low-high, negative autocorrelation, outlier, low value among high neighbours.

The graph scale is in terms of standardised Z values rather than raw data so that the mean of the variable is zero and the units become standard deviations. The Moran regression line is usually displayed and the slope is the value of the Moran I statistic.

Figures 14 to 16 show the Moran Scatter Plots for Crime, Household income and Housing Values, respectively. In each case, there are very few outliers in the dataset. There are many observations representing high-high and low-low clusters.

Figure 14 Moran Scatter Plot, Crime

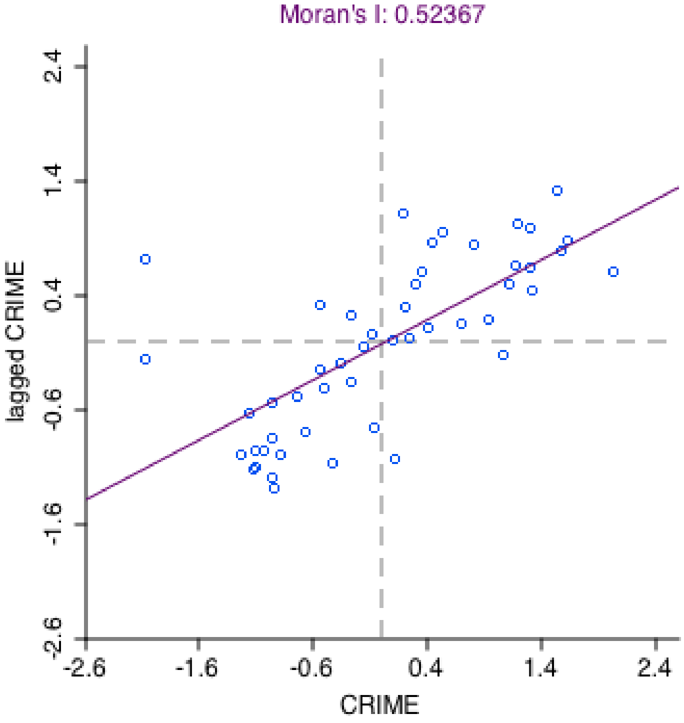


Figure 15 Moran Scatter Plot, Household Income

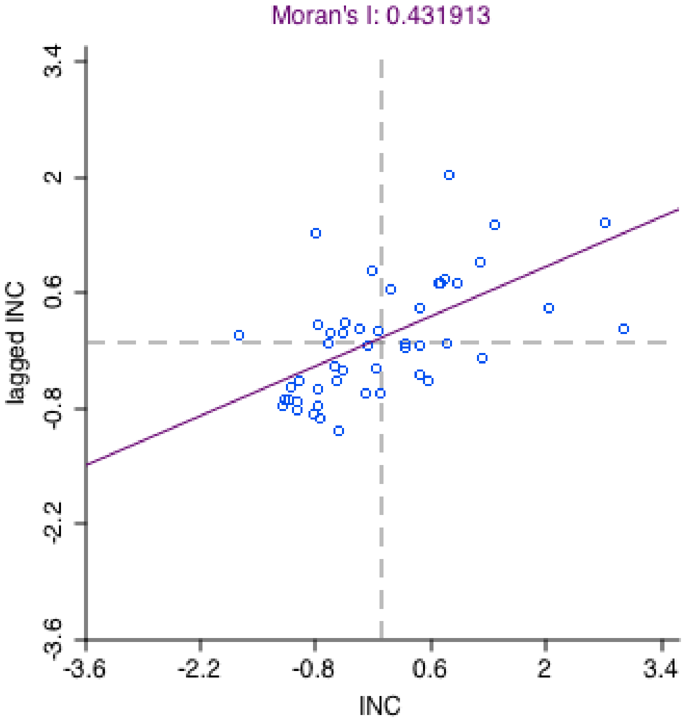
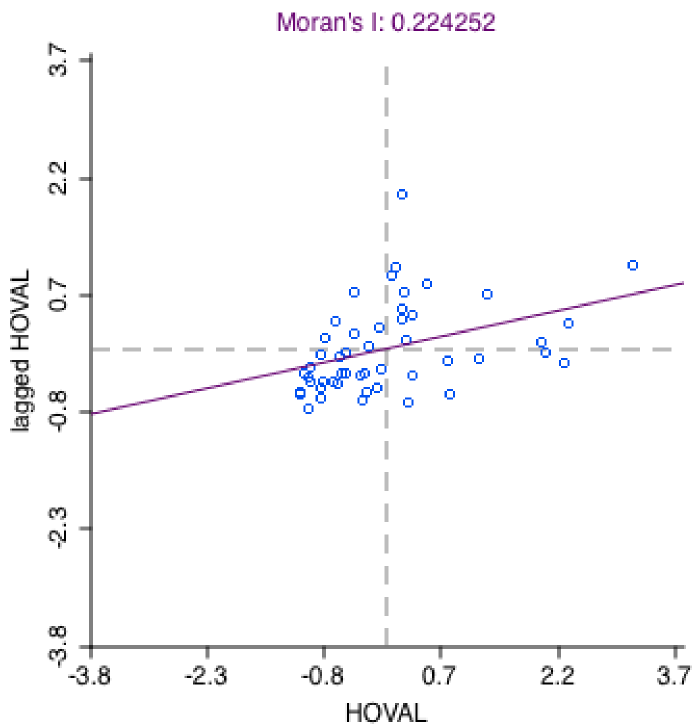


Figure 16 Moran Scatter Plot, Housing Values



7.5 Local Indicators of Spatial Association (LISA)

While the Moran scatterplot provides more detail on the type of spatial clustering, it does not report on the significance of such clustering. Local Indicators of Spatial Association (LISA) have been generated to do this. LISA serve two purposes in ESDA: they indicate local spatial clusters and they perform sensitivity analysis (identify outliers).

Anselin (1995: 94) defines LISA as:

> "a. the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around each observation; and
>
> b. the sum of LISAs for all observations is proportional to a global indicator of spatial association."

As with global measures, LISAs test whether the observed spatial pattern of unemployment amongst SLAs is extreme or is likely or expected, given a random geographic distribution of long-term unemployment.

The LISA statistic can be specified as follows (Le Gallo and Ertur, 2003):

$$I_{i,t} = \frac{(x_{i,t} - \mu_t)}{m_0} \sum_j w_{ij} (x_{j,t} - \mu_t) \;\; with \;\; m_0 = \sum_i \frac{(x_{i,t} - \mu_t)^2}{n} \qquad (9)$$

where $x_{i,t}$ is the observation in the region i for the year t, μ_t is the mean of observations across the regions in the year t and where the summation over j is such that only the neighbouring values are included. Positive values of $I_{i,t}$ mean that there is a spatial cluster of similar values and negative values represent a spatial cluster of dissimilar values.

There are various LISA available, but in this section we consider the cluster maps, which reveal “those locations with a significant Local Moran statistic classified by type of spatial correlation: bright red for the high-high association, bright blue for low-low, light blue for low-high, and light red for high-low … The high-high and low-low locations suggest clustering of similar values, whereas the high-low and low-high locations indicate spatial outliers” (Anselin, 2003: 100). The cluster maps show which regions are driving the global measures of spatial autocorrelation.

Figure 17 shows the LISA cluster map for Crime in Columbus. The legend indicates the frequency of the neighbourhoods in each category. There are 18 neighbourhoods that exhibit statistically significant Moran *I* statistics. The clustering is very pronounced. The high crime rate neighbours are concentrated in the inner city neighbourhoods, while the low crime rate neighbourhoods are clustered around the periphery of the city. Figure 18 shows the LISA cluster map for Household Income, which reveals a similar pattern and one consistent with our initial hypothesis – that high-income areas will have low crime rates and that the behavior that drives high or low crime crosses neighbourhood boundaries such that clusters form.

Figure 17 LISA cluster map for Crime

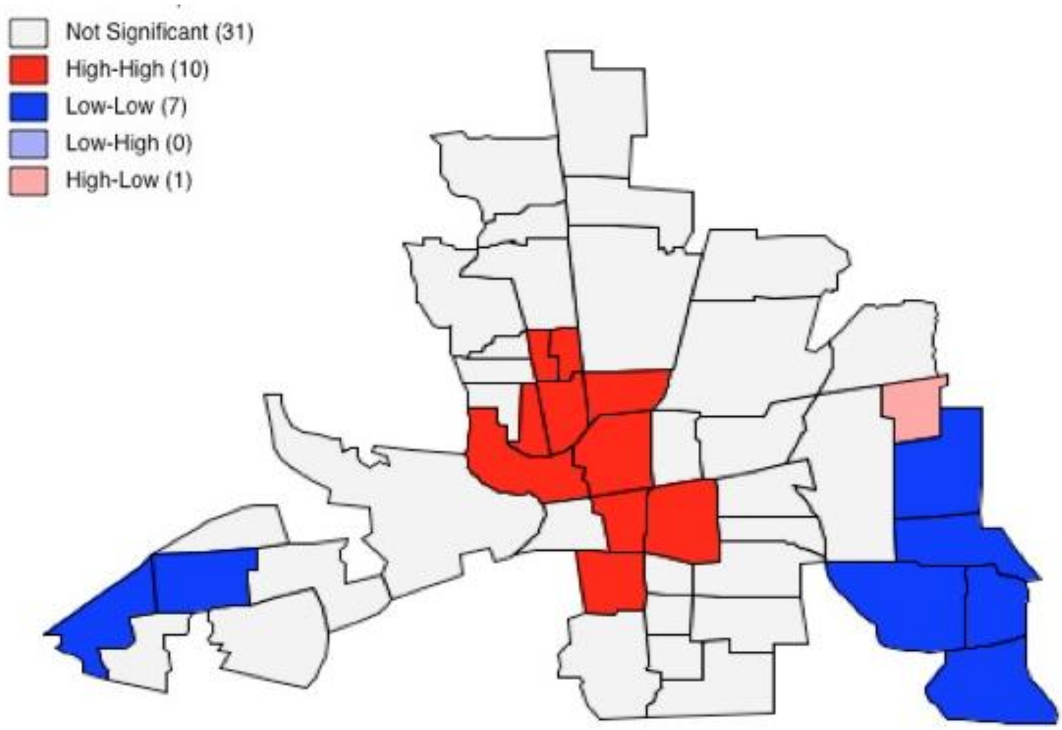
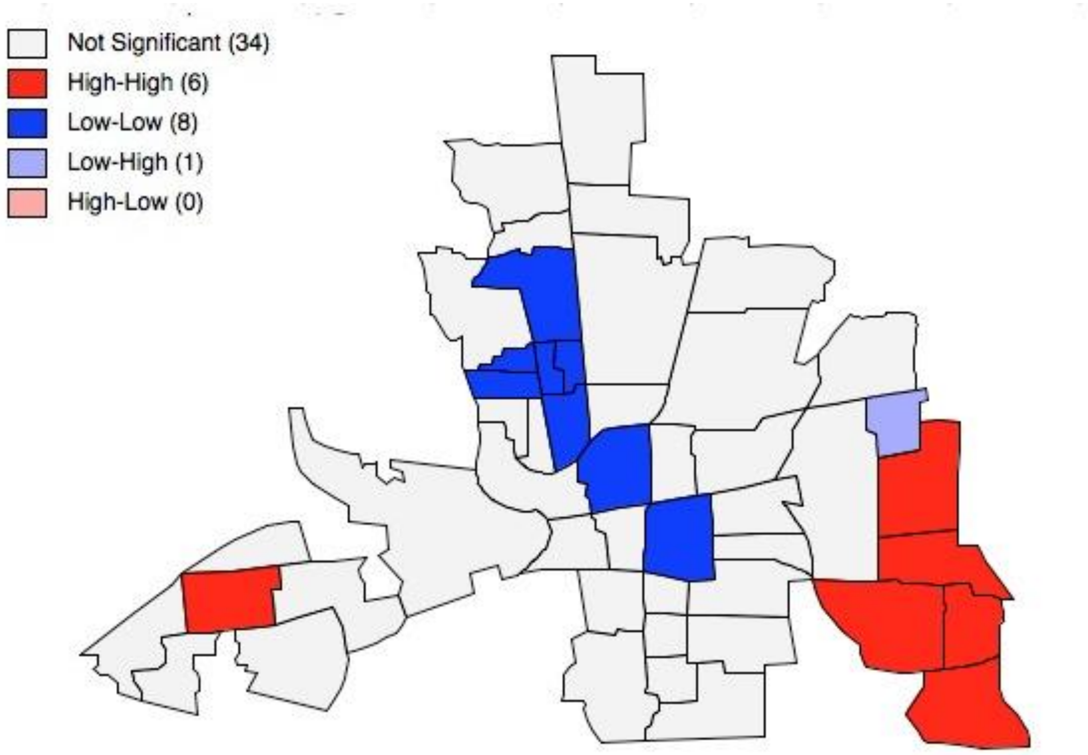


Figure 18 LISA cluster map for Household Income



8. Introduction to spatial econometric models

8.1 A taxonomy of spatial econometric models

To further investigate the way in which the crime rate is influenced by household income and housing values we need to perform more formal econometric modelling. The ESDA has confirmed that there is spatial dependence in the data and the use of OLS is likely to be problematic. In this section, we outlined the various spatial econometric models that are available to researchers seeking to investigate spatially dependent datasets.

The general spatial autoregressive econometric model is the starting point:

(6) $$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$
$$\mathbf{u} = \lambda \mathbf{W}_2 \mathbf{u} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

where $\mathbf{y}$ is a n x 1 dependent variable vector , $\mathbf{X}$ is a n x k explanatory variables matrix (including a constant) with an associated k x 1 vector of parameters $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ is a n x 1 random errors vector. $\mathbf{W_1}$ and $\mathbf{W_2}$ are n x n spatial weight matrices and $\mathbf{W}_{ij}$ is the spatial weight of region i in terms of region j. Table 3 outlines the family of spatial models that can be derived by imposing various restrictions on the general model shown as Equation (6).

The interpretation of the parameters in $\boldsymbol{\beta}$ has similarities with the interpretation of coefficients in a dynamic auto-regressive dynamic lag (ARDL) time series model, where we distinguish between short-run and long-run effects. In the spatial case, the analogy is captured by the concept of the spatial multiplier. We can rewrite the reduced form mean equation as:

(7) $$\mathbf{y} = (1 - \rho \mathbf{W})^{-1} [\boldsymbol{\beta}\mathbf{x} + \boldsymbol{\varepsilon}]$$

where for simplicity we assume well-behaved errors. The marginal effect of an increase in one of the columns of **X** is thus:

(8) $$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \left(\boldsymbol{I} - \rho \mathbf{W}\right)^{-1} \boldsymbol{\beta}$$

The term $\left(\boldsymbol{I} - \rho \mathbf{W}\right)^{-1}$ is the spatial multiplier (see Anselin, 2002).

We can think about this term as spreading the effects of any shocks to the dependent variable across (in this context) space to neighbouring regions. There are thus two effects embedded in the spatial multiplier. If we decompose the spatial multiplier (by geometric expansion, given $|\rho| < 1$) we get:

(9) $$\frac{\partial y}{\partial x} = \boldsymbol{I\beta} + \rho \mathbf{W}\boldsymbol{\beta} + \rho^2 \mathbf{W}^2 \boldsymbol{\beta} + \ldots$$

So the first term ($\boldsymbol{I\beta}$) is termed the direct effect of a marginal change of $\boldsymbol{x}$ on $\boldsymbol{y}$ (operating via the main diagonal). The second term is a matrix with zero values on the main diagonal and the off-diagonal elements capture the local indirect or spillover effects arising from the direct shocks. The third term (and all subsequent higher order terms) capture the induced effects which spill-over into the neighbouring regions (see Abreu *et al.*, 2004).

In other words, the spatial lag model is a way of capturing interdependency between the data points across space (or across any cross sectional data set where the observations are not independent).

Table 3 Taxonomy of spatial econometric models

Model	Specification	Restrictions	Comments
Ordinary Least Squares	$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$	$\mathbf{W}_1 = \mathbf{0}$ $\mathbf{W}_2 = \mathbf{0}$	No spatial effects
First-order spatial autoregressive (FAR) model	$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$	$\mathbf{X} = \mathbf{0}$ $\mathbf{W}_2 = \mathbf{0}$	
Mixed autoregressive-spatial autoregressive (SAR) model	$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$	$\mathbf{W}_2 = \mathbf{0}$	ρ measures the degree of spatial dependence. In this study it is the average influence of unemployment rates in neighbouring regions on the unemployment rate in region *i*.
Spatial autocorrelation (SEM) model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ $\mathbf{u} = \lambda \mathbf{W}_2 \mathbf{u} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$	$\mathbf{W}_1 = \mathbf{0}$	Spatial autocorrelation may be due to measurement problems (rather than endogenous effects occurring between regions).
Spatial Durbin (SDM) model	$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{W}_1 \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$	$\mathbf{W}_2 = \mathbf{0}$	Spatially weighted term added to the FAR model. The parameters ρ and γ measure the strength of the spill-over effects. One or more **X** variables can be spatially lagged.
General spatial model (SAC)	$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ $\mathbf{u} = \lambda \mathbf{W}_2 \mathbf{u} + \boldsymbol{\varepsilon}$ $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$		Combines the SAR and SEM models. λ measures the degree of spatial residual correlation.

Source: Anselin, 1988.

8.2 Model selection methods

The issue of model selection techniques (specification strategies) remains contentious in the spatial econometric literature although some consensus is emerging. One viewpoint is that the researcher should not engage in a specification search but rather pre-filter the data, netting out any inherent spatial dependence (for example, Getis, 1995). The spatially-filtered data can then be approached using conventional OLS estimation.

The alternative approach to ‘filtering’ can be cast, once again, in the broader debate common among time series econometricians. Two options appear possible. First, should we proceed with a specific-to-general approach (the so-called ‘classical approach’), which begins with the simplest OLS regression and then uses appropriate tests of restrictions (Lagrange Multiplier tests) to assess the statistical validity of a range of ‘added variables’ including the presence of spatial dependence? In this case, the specification search is less transparent and the researcher would ultimately choose the model with some highest test value. For example, Anselin (1992) suggests that LM tests could provide the basis for the choice between the SEM and the SAR model. We can test whether $\lambda = 0$ in the SEM model and whether $\rho = 0$ in the SAR model. The model with the largest test statistic would be rejected.

Second, as an alternative, we might follow the Hendry general-to-specific approach, where the researcher deliberately sets out with an over-parameterised model, which in this context would be include all the spatial effects, and then ‘test down’ using valid simplifying restrictions to the parsimonious form. Florax *et al.*, (2003) used Monte Carlo simulation to demonstrate that the classical approach provides for better inference than the Hendry approach.

Other issues regarding parsimony of the spatial weights matrix also arise – in the sense, that the asymptotic properties of the estimators are unclear after a certain number of non-zero elements is included.

8.3 Spatial autocorrelation diagnostic tests

We employ the standard spatial diagnostic tests to test for spatial autocorrelation in the residuals of the OLS regression and the SAR models. These tests are outlined in LeSage (1999) and are summarised as follows:

<u>Moran *I*-statistic</u> (Cliff and Ord, 1981) is written as:

(10) $$\mathbf{I} = \mathbf{e}'\mathbf{W}\mathbf{e} / \mathbf{e}'\mathbf{e}$$

where **e** is the regression residuals. The ***I*** statistic has an asymptotic distribution that corresponds to the standard normal distribution after subtracting the mean and dividing by the standard deviation of the statistic (Anselin, 1988: 102). We thus interpret the standardised version as rejecting the null of no spatial autocorrelation if its value exceeds 1.96 (at the 5 per cent level).

<u>Likelihood ratio test</u> compares the LR from the OLS model to the LR from the SEM model and this statistic is asymptotically distributed as $\chi^2(1)$. We reject the null of no spatial autocorrelation if the test statistic exceeds 3.84 (at the 5 per cent level) and 6.635 (at the 1 per cent level).

<u>Wald test</u> (Anselin, 1988: 104) is asymptotically distributed as $\chi^2(1)$. We reject the null of no spatial autocorrelation if the test statistic exceeds 3.84 (at the 5 per cent level) and 6.635 (at the 1 per cent level).

<u>Lagrange Multiplier</u> (LM) test (Anselin, 1988: 104) uses the OLS residuals **e** and the spatial weight matrix **W**, and is computed as:

(11) $$LM = (1/T)\left[\left(\mathbf{e}'\mathbf{W}\mathbf{e}\right)/\sigma^2\right]^2 \sim \chi^2(1)$$
$$T = tr\left(\mathbf{W} + \mathbf{W}'\right)\mathbf{W}$$

<u>Spatial error residuals LM test</u> (Anselin, 1988: 106) is based on the residuals of the SAR model to determine “whether inclusion of the spatial lag term eliminates spatial dependence in

the residuals of the model" (LeSage, 1999: 75). The test requires the spatial lag parameter r is non-zero in the model. The test produces a LM statistic, which is asymptotically distributed as $\chi^2(1)$. As before, we reject the null of no spatial autocorrelation if the test statistic exceeds 3.84 (at the 5 per cent level) and 6.635 (at the 1 per cent level).

9. Results and Analysis

Table 4 reports the various regression results using the R software package **spdep**. We compare the OLS, SAR and SEM models in this Chapter. The numbers in parentheses below the coefficient estimates are *t*-statistics (for OLS) and *z*-values for the Maximum Likelihood (ML) spatial equations. The first column of results reports the estimates from the OLS regression, which gives no special attention to the spatial properties of the data. The crime rate is regressed on a constant term, income (INC) and housing values (HOVAL).

The estimated coefficients on the income and housing values variables are highly significant (*t*-statistics above 2) and of the hypothesised signs. That is, the higher is the income and the housing values of a neighbourhood; the lower is the expected crime rate. The model would also predict, for example, that if household incomes rose in a neighbourhood, the crime rate would be expected to fall.

The R^2 statistic indicates the goodness of fit of the equation (how strong the relationship is between the variables). The adjusted R^2 statistic controls for loss of degrees of freedom as more regressors are added and is also above 0.5. Taken together the OLS equation is a reasonable fit to the data. In professional practice, we would also report the test statistics for a range of advanced diagnostic tests, which aim to examine the assumptions we made about the residuals. These would include tests for normality, heteroscedasticity and serial correlation. We ignore these statistics in this Chapter to focus only on the issue of spatial dependence.

How can we tell if there is spatial autocorrelation in the error term of this estimated equation? After running the OLS regression we would subject the estimated residuals to a battery of tests for spatial autocorrelation (as outlined in Section 8.4). Table 1 shows the results for the standardised normal Moran *z*-statistic and the LM $\chi 2(1)$ statistic. Both are high significant indicating the strong presence of spatial autocorrelation in the OLS residuals.

The second column of results shows the maximum likelihood estimation results for Mixed autoregressive-spatial autoregressive (SAR) model, which adds the spatial lag term to the explanatory variables. The spatial lag coefficient is ρ (reported in the Table as rho) and the number in parenthesis immediately below it is the standardised z-value.

First, note that the estimated coefficients retain their hypothesised signs but the size of the coefficient on income is now much lower. There is very little difference in the coefficient on Housing Values. Second, the coefficient on the spatial lag (Rho) is highly significant with an asymptotic z-value of 3.69 and positive. This indicates that there is significant spatial dependence between rook-contiguous crime rates such that a higher rate in region *i* will tend to push up the rate in region *j*, independent of the impact of household income and housing values.

A Moran's *I* test under normality assumptions on the SAR equation residuals produced a *p*-value of 0.5938 indicating that once the spatially lagged dependent variable was included in the model there is no further evidence of spatially autocorrelated errors.

The OLS equation residuals exhibited strong spatial autocorrelation and despite the finding that there was no further evidence of spatially autocorrelated errors once the spatially lagged dependent variable was included (the SAR) model, it is sometimes useful to estimate the Spatial autocorrelation (SEM) model. The results are in the third column.

Table 4 Spatial econometric results for dependent variable, CRIME

	OLS	SAR ML	SEM ML
Constant	68.62	46.85	61.05
	(14.49)	(6.41)	(11.49)
Income	-1.60	-1.07	-1.00
	(4.78)	(3.45)	(2.95)
Housing Values	-0.27	-0.27	-0.31
	(2.65)	(3.00)	(3.33)
Rho		0.40	
Asymptotic Z		(3.35)	
Lambda			0.52
Asymptotic Z			(3.69)
R^2	0.552		
Adjusted R^2	0.533		
Log likelihood		-183.1683	-184.1552
Spatial Autocorrelation Tests			
Standardised Moran z-statistic	2.952		
LM test $x^2(1)$	5.723		

ML is Maximum Likelihood estimation.

10. Conclusion

Typically social and economic data have some spatial dimension. Statistically, it is traditional for regionally based cross-sectional data to be viewed as conceptually identical to cross-sectional data on individuals or businesses at a single location. However, spatially adjacent observations are likely to exhibit spatial interdependence, owing to dynamics (knowledge flows, trade flows, commuting and social spill-overs), which accompany proximity. We consider that observations that are near to each other in space are more likely to be co-related in some way than more distant observations.

Spatial autocorrelation refers to the formal measure of the extent near and distant things are related. Ignoring dependence between neighbouring regions will lead to problematic regression results. In the last twenty years, a range of spatial regression techniques have been developed to measure latent forces of interaction and handle data that violate standard statistical assumptions of independence.

Using the Columbus, Ohio dataset to demonstrate the concepts involved, we found that the Ordinary Least Squares (OLS) regression of the crime rate on household income and housing values revealed significant spatial autocorrelation in the residuals. The spatial dependence likely indicates the presence of interactions between spatially proximate regions (keeping in mind sensitivities to model misspecification and the weighting structure employed). In that context, the spill-overs between regions will magnify local responses.

In this Chapter, we have provided an elementary introduction to the problem of spatial interdependence and outlined ways of detecting it and resolving it, in order to produce robust statistical analysis.

References

Abreu, M., Groot, H. and Florax, R. (2004) 'Spatial Patterns of Technology Diffusion: An Empirical Analysis Using TFP', ERSA Conference, Porto.

Abbott (1997) 'Of time and space: The contemporary relevance of the Chicago school', *Social Forces*, 75. 1149-1182.

Akerlof, G. (1997) 'Social Distance and Social Decisions', *Econometrica*, 65(5), 1005-1007.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (1992) *SpaceStat Tutorial. A Workbook for Using SpaceStat in the Analysis of Spatial Data*, Regional Research Institute, West Virginia University.

Anselin, L. (1995), 'Local Indicators of Spatial Association – LISA', *Geographical Analysis*, 27, 92-115.

Anselin, L. (1996) 'The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association', in Fischer, M. M., Scholten, H.J. and Unwin, D. (eds) *Spatial Analytical Perspectives on GIS*, Taylor & Francis, London, 111-125.

Anselin, L. (2002) 'Spatial Externalities, Spatial Multipliers and Spatial Econometrics', *mimeo*, Department of Agricultural and Consumer Economics,

Anselin, L. (2003) *Exploring Spatial Data with GeoDa: A Workbook*, Center for Spatially Integrated Social Science, (Revised 2005).

Banerjee, A.V. (1992) 'A Simple Model of Herd Behavior'. *Quarterly Journal of Economics* 107(3), 797-818.

Bivand, R.S. (2006) 'Implementing Spatial Data Analysis Software Tool in R', *Geographical Analysis*, 38, 23-40.

Borland, J. (1995) 'Employment and income in Australia – Does the neighbourhood dimension matter?', *Australian Bulletin of Labour*, 21, 281-293.

Cliff, A.D. and Ord, J.K. (1973) *Spatial Autocorrelation*, Pion, London.

Cliff, A.D. and Ord, J. (1981) *Spatial Processes: Models and Applications*, Pion, London.

Durlauf, S. (2003) 'Neighborhood Effects', *mimeo*, Department of Economics, University of Wisconsin.

Florax, R. and Nijkamp, P. (2003) 'Misspecification in Linear Spatial Regression Models', Tinbergen Institute Discussion Papers 03-081/3, Tinbergen Institute.

Frank, A. (2003) 'Using Measures of Spatial Autocorrelation to Describe Socio-economic and Racial Patterns in US Urban Areas', in Kidner, D., Higgs, G. and White, S (eds), *Socio-Economic Applications of Geographic Information Science*, Taylor & Francis, London.

Getis, A. (1995) ‘Spatial Filtering in a Regression Framework: Experiments on Regional Inequality, Government Expenditures, and Urban Crime’, in Anselin, L. and R.Florax, R. (eds) New Directions in Spatial Econometrics, North Holland, Amsterdam.

Goodchild, M, Anselin, L., Applebaum, R. and Herr Harthorn, B., (2000) ‘Toward Spatially Integrated Social Science’, *International Regional Science Review*, 23(2), 139-159.

Hawley, A. (1950) *Human Ecology: A Theory of Community Structure*, The Ronald Press Company, New York.

Helliwell, J.F. (1998) *How Much Do National Borders Matter?,* Brookings, Washington DC.

Hordijk, L. (1979) ‘Problems in Estimating Econometric Relations in Space’, *Papers of the Regional Science Association*, 42, 99-115.

Hulse, K., Randolph, B., Toohey, M., Beer, G. and Lee, R. (2003) *Understanding the Roles of Housing Costs and Housing Assistance in Creating Employment Disincentives*, Positioning Paper, Australian Housing and Urban Research.

Krugman, P. (1991) *Geography and Trade*, MIT Press, Cambridge.

Le Gallo, J. and Ertur, C. (2003) ‘Exploratory Spatial Data Analysis of the Distribution of Regional Per Capita GDP in Europe, 1980-1995’, *Papers in Regional Science*, 82, 175-201.

LeSage, J.P. (1999) ‘Spatial Econometrics’, University of Toledo, http://www.rri.wvu.edu/WebBook/LeSage/spatial/spatial.html.

Le Sage, J. (2005) *Using the Spatial Econometrics Toolbox for Estimation of Spatial Regression : Models and Mapping in MATLAB*, Keynote Address to ARCRNSISS Workshop – Methodology, Tools and Techniques Paradigm Forum, 15-17 June, University of Newcastle.

Longley P., Goodchild, M., Maguire, D. and Rhind, D., (2001) *Geographic Information Systems and Science*, John Wiley & Sons, New York.

Molho, I. (1995) ‘Spatial Autocorrelation in British Unemployment’, *Journal of Regional Science*, 35(4), 641-658.

Moran, P. (1948) ‘The Interpretation of Statistical Maps’, *Journal of the Royal Statistical Society*, Series B 10: 243-251.

Niebuhr, A. (2002) ‘Spatial dependence of regional unemployment in the European Union’, HWWA Discussion Papers 186, Hamburg Institute of International Economics (HWWA).

O'Connor, K. and Healy, E. (2002) ‘The Links Between Labour Markets and Housing Markets in Melbourne’, Australian Housing and Urban Research Institute, Swinburne-Monash Research Centre.

Ord, J.K. and Getis, A. (1995) ‘Local Spatial Autocorrelation Statistics: Distributional Issues and an Application’, *Geographical Analysis*, 27, 286-306.

Park, R.E., Burgess, E.W. and McKenzie, R.D. (eds) (1925) *The city*, University of Chicago Press, Chicago, Illinois.

Park, R.E. (1936) ‘Human ecology’, *American Journal of Sociology*, 42, 1-15.

Stetzer, F. (1982) ‘Specifying Weights in Spatial Forecasting Models: The Results of Some Experiments’, *Environment and Planning*, 14, 571-584.

Tassinopoulos, A. and Werner, H. (1999) ‘To Move or Not to Move. Migration in the European Union’, IAB Labour Market Research Topics 35, IAB, Nurenberg.

Tobler, W. R. (1970) ‘A computer movie simulating urban growth in the Detroit region’, *Economic Geography*, 46, 234–40.

Topa, G. (2001) ‘Social Interactions, Local Spillovers and Unemployment’, *Review of Economic Studies*, 68, 261-295.

Upton, G.J. and Fingleton, B. (1985) *Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data*, Wiley, Toronto.

Wilson, W.J. (1987) *The Truly Disadvantaged*, University of Chicago Press, Chicago.

---

1 The author is the Director of Centre of Full Employment and Equity and Professor of Economics at the Charles Darwin University, Australia. He acknowledges excellent research assistance from Michael Flanagan.