AURIN Project

Project Title

Spatial Statistics and Econometrics e-Tools

Lead organisation, group or individual name

Centre of Full Employment and Equity

Date

March 2015

About

The Spatial Statistics and Econometric e-Tools are a suite of tools for carrying out spatial data analyses implemented through the Australian Urban Research Infrastructure Network (AURIN) portal. These tools provide users with a freely available graphical user interface to perform spatial data analyses, integrated into the AURIN portal where users can either download data in the portal or upload their own datasets. The tools that are available include methods for designing spatial weight matrices, detecting spatial autocorrelation in variables, running spatial regression models as well as spatial indices and concentration/inequality measures. The e-tools make use of the R statistical package, utilising the various add-on packages that deal with spatial data analysis, and are fully integrated with the other capabilities of the AURIN portal.

Help File

1. Introduction

The portal of the Australian Urban Research Infrastructure Network (AURIN) website allows users to perform various statistical functions using a graphical user interface. One group of tools that can be implemented come under the heading of Spatial Statistics. As the name suggests these tools allow the user to implement a range of spatial statistical analyses as well as performing spatial econometric tests.

The Spatial Statistics and Econometrics e-Tools (hereafter e-tools) utilises the statistical package R to carry out the calculations. R is an open source statistical programming environment that is widely used throughout the academic and business world. It was designed as a "suite of software facilities for data manipulation, calculation and graphical display," but has been used mostly by people as a statistics system (Venables *et al.*, 2013: 2).

Very few statistical techniques are actually provided in the R environment, most are implemented in packages that are added to the R software. Some packages are supplied in the basic downloadable version of R, but most are available to be downloaded through the CRAN internet site (http://CRAN.R-project.org). The Tools make use of routines that are in the spdep package, which deal specifically with spatial statistics and econometrics (see Bivand *et al.*, 2008 and Bivand, 2013).

The area of spatial statistics has been around since the development of Moran's much used *I* statistic (Moran, 1948). Since then a large body of spatial statistical techniques have been developed (see Cressie, 1993 for a comprehensive summary). Among the contributions was Tobler's (1970: 236) first law of geography that "everything is related to everything else, but near things are more related than distant things" which implied that spatial autocorrelation was the norm, rather than the exception. The Tools implement a number of the more widely used spatial statistics.

Econometricians, meanwhile, were preoccupied with the time dimension and overcoming the problem of serial dependence, that they neglected the spatial autocorrelation problem inherent in cross-sectional data. The physical sciences had widely accepted the relevance of spatial autocorrelation long before it began to be recognised in mainstream econometrics. Anselin's (1988a) seminal book built on the spatial statistics literature to allow for the presence of spatial autocorrelation in econometric modelling. The Tools allow the user to perform spatial regressions using a variety of spatial econometric models that have been prevalent throughout the literature.

The increase in acceptance and popularity of spatial statistics and spatial econometric modelling means a suite of tools such as the ones the Tools provide is of great benefit to the econometric community. The main advantage of such a website is that it makes somewhat complex statistical procedures accessible to a wide range of prospective users who would find their capabilities invaluable. Potential users of such techniques who may baulk at having to learn a programming language or commands can undertake their work without the added concern of training in a particular language or statistical package.

Despite the ease with which these complex procedures can be performed, it is still vitally important that the user has an understanding of what they are doing at each step. Indeed, the success of a user's implementation of the various components contained in the Tools is contingent on the user's understanding of when to use the components and what each does. This help file provides a brief outline of how and when a user should undertake each of the

components; however this should not replace a thorough understanding of the theory behind these routines. To that end, the references provided herein will be most useful for inexperienced users of spatial statistics and econometrics.

Throughout this help file we introduce each of the components from a theoretical perspective and then outline the required inputs and expected outputs for the Tools. Following this introduction is a discussion of the concept of spatial autocorrelation. We then turn to spatial weight matrices and describe the three types of spatial weight matrices that are available in the Tools. Section 4 looks at global measures of spatial autocorrelation, while Section 5 discusses local measures of spatial autocorrelation and the corresponding components that are available for each. In Section 6 we discuss spatial regression and look at the different diagnostic tests and models that are available. Section 7 provides details on the spatial indices that are in the Tools. Section 8 discusses three measures of concentration that are available on the portal. Section 9 provides an overview of the Concordance e-Tool that is hosted on the CofFEE website as part of the AURIN project. Concluding remarks follow.

2.   The concept of spatial autocorrelation

Spatial autocorrelation refers to the formal measure of the extent near and distant things are related. Figure 2.1 below, 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.

There are two reasons proposed as to why spatial dependence may exist between regions. First, data collected on observations associated with spatial units may contain measurement error because the delineated boundaries for data collection do not reflect the underlying processes generating the sample data (Anselin, 1988a: 11-12). The most common example of this is where administrative regions are used as the basis for the dissemination of economic data. If social or economic phenomena cross geographic boundaries we would expect to see

Figure 2.1 Stylised patterns of spatial correlation

| a) positive spatial correlation | b) negative spatial correlation | c) zero spatial correlation |
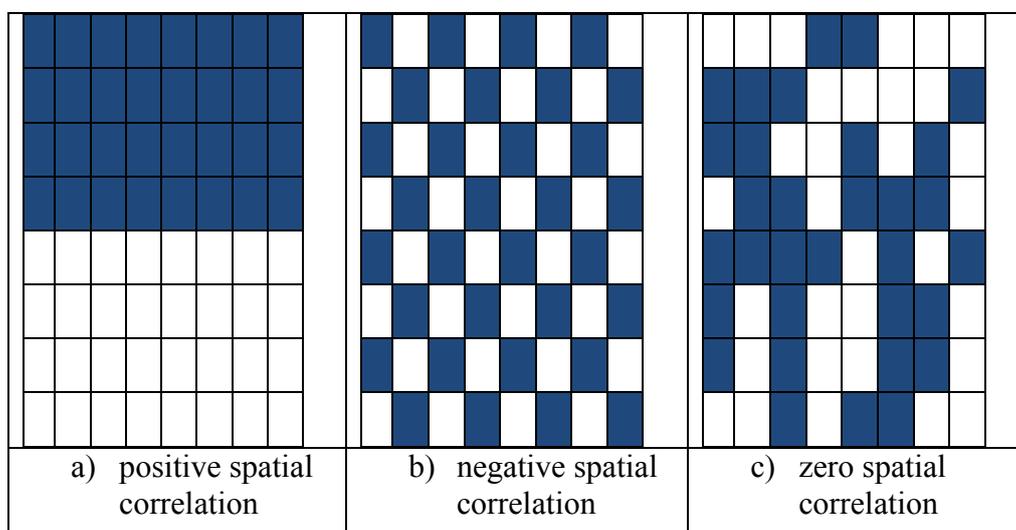
Source: Longley *et al*. (1996)

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. This is the result of what Openshaw (1984: 3) termed the Modifiable Areal Unit Problem (MAUP), where the "the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating."

Second, location and distance are important forces at work in human geography and market activity. For example, clustering of unemployment rates might occur because of the 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 then becomes an important factor in determining the extent of spatial dependence. Neoclassical explanations for regional unemployment differentials revolve around the rigidity of wages and the imperfect mobility of labour resources (Debelle and Vickery, 1999). 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).

3.   Spatial Weight Matrices

Spatial statistics and spatial econometrics analyses require the specification of spatial weight matrices which Stetzer (1982: 571) notes 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).

Spatial weight matrices represent the strength of the connection between all pairs of places in the system. Two broad types of matrices dominate the literature: one based on contiguity of areas and one based on relative distance between areas. While the early measures of spatial weight matrices primarily used contiguity as their basis, more recent measures have involved notions of distance.

The spatial weights matrix, W, is characterised by each element, w_{ij}, reflecting the connection between place *i* and place *j*. In a first order queen contiguity matrix, $w_{ij} = 1$ if the boundaries of two areas touch, and $w_{ij} = 0$ otherwise. Higher order contiguity can also be implemented. Distance threshold matrices give $w_{ij} = 1$ if the distance between *i* and *j* is within a threshold distance, and zero otherwise. Distance decay matrices have matrix elements that get smaller the greater the distance between two places; these may or may not have a threshold distance above which $w_{ij} = 0$.

There is much conjecture as to the most appropriate way to design a spatial weight matrix. Anselin and Bera (1998: 243) acknowledge that "the specification of which elements are non-

zero in the spatial weights matrix is a matter of considerable arbitrariness." Further, Bill (2005: 6) cites a presentation by LeSage that "the main aim of the weighting matrix is to incorporate some notion of proximity into standard statistical tests."

3.1   Contiguous Spatial Weight Matrix

This component computes a contiguous spatial weight matrix for a chosen shapefile. A contiguous spatial weight matrix is one where neighbours are determined by whether they are contiguous or not. Queen contiguity exists if two areas share a single boundary point. Rook contiguity exists if two areas share more than one boundary point (this is not necessarily a boundary line). First order contiguity requires two regions to share a common boundary or point, while higher order neighbours can also be calculated.

- Input parameters
  - o Dataset source – the shapefile to be used.
  - o Contiguity Spatial Weight Matrix Type – tick indicates queen contiguity, left blank indicates rook contiguity. Queen contiguity requires neighbouring regions to share a single point, while rook contiguity requires neighbours to share more than one boundary point.
  - o Contiguity Spatial Weight Matrix Order – integer for the order of contiguity (normally is 1, but can be up to 5)
  - o Contiguity Spatial Weight Matrix Style – style of the matrix. Options are: binary, meaning the matrix will contain a 1 for regions that are neighbours and a 0 otherwise; row-standardised, which row-standardises the binary matrix; globally-standardised, which globally-standardises the binary matrix.
  - o Contiguity Spatial Weight Matrix Island Parameter – tick indicates islands are expected, blank indicates islands are not expected. Islands are areas with no neighbours as defined by the type of matrix being formed. If the user leaves this blank and islands are found a message will be returned. Islands are always allowed in the spatial weight matrices.
- Expected outputs
  - o The spatial weight matrix, which is a full matrix of size n x n where n is the number of regions in the dataset source.

3.2   Distance-based Spatial Weight Matrices

This component computes a spatial weight matrix that is calculated using a distance-based constraint. There are two options of matrices formed from this: k-nearest-neighbours, where each region is a neighbour with its k closest regions; and distance-threshold, where regions are neighbours with all other regions within a threshold distance of themselves.

- Input parameters
  - o Dataset source – the shapefile to be used.
  - o Distance Spatial Weight Matrix Type – options are distance-threshold or k-nearest-neighbours. In a k-nearest-neighbours matrix, regions are neighbours with their k closest regions; while in a distance-threshold matrix, regions are neighbours with other regions within the threshold distance, where distances are taken from centroids of the regions.
  - o Distance Spatial Weight Matrix Threshold Distance or k Parameter – if the user has chosen to calculate a distance-threshold matrix this is the threshold distance, above which regions are non neighbours. If the user has chosen a k-nearest-neighbours matrix this is the value of k, the number of closest regions to be considered neighbours.

- Distance Spatial Weight Matrix Style – style of the matrix. Options are: binary, meaning the matrix will contain a 1 for regions that are neighbours under the definition of the matrix type and a 0 otherwise; row-standardised, which row-standardises the binary matrix; globally-standardised, which globally-standardises the binary matrix.
- Distance Spatial Weight Matrix Island Parameter – tick indicates islands are expected, blank indicates islands are not expected. Islands are areas with no neighbours as defined by the type of matrix being formed. If the user leaves this blank and islands are found a message will be returned. Islands are always allowed in the spatial weight matrices.
- Expected outputs
  - The spatial weight matrix, which is a full matrix of size n x n where n is the number of regions in the dataset source.

3.3   Distance Decay Spatial Weight Matrices

This component computes a spatial weight matrix that is termed distance decay, that is, it provides greater weighting to regions that are closer to each other than regions that are further apart. There are two options of matrices that are formed from this:

1. Inverse distance matrix. The weighting between regions i and j, $w_{ij} = 1/d_{ij}$, for $d_{ij} < D_{max}$; $w_{ij} = 0$ otherwise, where d_{ij} is the distance between the centroids of regions i and j and D_{max} is a threshold distance.
2. Exponential distance decay matrix. The weighting between regions i and j, $w_{ij} = \exp(-cd_{ij})$ for $d_{ij} < D_{max}$; $w_{ij} = 0$ otherwise, where c is the distance decay parameter.

- Input parameters
  - Dataset source – the shapefile to be used.
  - Distance Decay Spatial Weight Matrix Type – options are inverse-distance or exponential-distance-decay. See above for definition and calculation of each of these.
  - Distance Decay Spatial Weight Matrix Threshold Distance – this is the threshold distance, in km, above which the weights between regions will be 0. If the user wants all regions to be considered neighbours to all other regions ensure the threshold distance is at least the distance between the two regions which are the greatest distance apart.
  - Distance Decay Spatial Weight Matrix Exponential Decay Parameter – this is the value of c in the exponential distance decay formula given in point 2 above (traditionally between 0.1 and 0.9).
  - Distance Spatial Weight Matrix Style – style of the matrix. Options are: absolute, meaning the matrix will contain the value calculated from the corresponding formula above, or a 0 if the distance between the regions is above the distance threshold; row-standardised, which row-standardises the absolute matrix; globally-standardised, which globally-standardises the absolute matrix.
  - Distance Spatial Weight Matrix Island Parameter – tick indicates islands are expected, blank indicates islands are not expected. Islands are areas with no neighbours as defined by the type of matrix being formed. If the user leaves this blank and islands are found a message will be returned. Islands are always allowed in the spatial weight matrices
- Expected outputs

o The spatial weight matrix, which is a full matrix of size n x n where n is the number of regions in the dataset source.

Note: All components that follow allow for no-neighbour regions in the spatial weight matrix by default.

4. Global Measures of Spatial Autocorrelation

In order to understand the patterning of economic phenomena in large and detailed maps, we need a way of capturing the properties of the data in a simple summary. Global measures of spatial autocorrelation provide 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).

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 represented in a map deviate from ones which would exist if they were randomly assigned, requires 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. Some debate exists over how spatial autocorrelation can be best characterised. Upton and Fingleton (1985) defined spatial autocorrelation as a property that mapped data displays whenever it exhibits an organised pattern. A lack of spatial autocorrelation is found when the patterning of data on a map doesn't deviate significantly from one in which each Xi value is assigned randomly with equal probability to each (i,j) location. Earlier Cliff and Ord (1973) had argued that spatial autocorrelation exists when the distribution of some quality or quantity in a region makes its presence more or less likely in neighbouring regions. The former definition implies that spatial autocorrelation is simply the non-random patterning of data, the latter implies a dependence springing from the proximity of neighbouring regions (after controlling for other factors).

Most published measures of spatial autocorrelation can be reworked as a normalised cross-product statistic that indexes the degree of relation between corresponding observations in two matrices (Sawanda, 2005). The first is a weighting matrix which specifies the degree of interrelatedness (bi,j), between a set of n locations or observations. The other reflects a definition of 'similarity' – usually correlatedness or concentration (ai,j) – between some variable x over n locations (Hubert and Arabie, 1991; Sawanda, 2005):

(1) $\quad \tau_{obs} = \sum_{i \neq j} a_{ij} b_{ij}$

Thus spatial autocorrelation compares two sets of similarities, similarities in attributes and similarities in location. The relative size of τ_{obs} can be assessed by constructing a *p*-value, for the index using as a reference the distribution of values under an assumption of randomness (usually taken to be normal).

4.1 Moran's *I*

This component computes the Moran's *I* global measure of spatial autocorrelation statistic and related information. The Moran's *I* (Moran, 1948) 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 is the spatial weight matrix which formalises the neighbourhood or contiguity structure of the dataset.

The Moran's *I* statistic is computed as:

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

where *R* is the number of regions in the dataset; R_b is the sum of the weights which simplifies to *R* if the spatial weight matrix is row-standardised; x_i is the variable that we are testing and is measured as deviation from the mean, i.e. $x_i = X_i - \bar{X}$. The location variable for the area's proximity is given by w_{ij} which is the element from the corresponding spatial weight matrix, as described in the previous section.

The Moran's *I* statistic can be expressed as a standardised normal *Z*-score for inference purposes, computed by:

(3) $Z_I = \dfrac{I - E(I)}{sd(I)}$

where *I* is Moran's *I* statistic, *E(I)* is the theoretical mean and *sd(I)* the theoretical standard deviation of Moran's *I* statistic. Anselin (1992, p. 134) points out the different assumptions that can be made about the data that effect the calculation of the standard deviation. The component allows for the standard deviation to be calculated under the assumption of randomisation or normality, both of which allow the computed *Z*-score to follow a normal distribution (asymptotically), so its significance can be evaluated by means of a standard normal table.

Alternatively, a user may prefer either a Saddlepoint approximation (Tiefelsdorf, 2002) or an exact test (Tiefelsdorf, 2000) which are also provided. In general these make little difference to the significance of global tests unless *R* is quite small. If the user would prefer a permutation test, this facility is provided in the Moran's *I* permutation test component.

The range of possible Moran's *I* values are between minus one and one. An estimate of zero implies no spatial autocorrelation. For a significant estimate the closer it gets to one, the greater the degree of positive spatial autocorrelation; while the closer it is to minus one indicates stronger negative spatial autocorrelation.

Anselin (1996) shows the relationship between the Moran Scatterplot and the Moran's *I* statistic. The Moran Scatterplot visualises and identifies the degree of local spatial instability in spatial association that is present in the Moran's *I* statistic. The Moran's *I* statistic can be interpreted as the regression coefficient in a bivariate regression of the spatially lagged variable, *Wx*, on the original variable, *x* (in deviations from the mean). The spatially lagged variable, *Wx*, is the average of observations at neighbouring locations, that is, locations for which $w_{ij} \neq 0$.

This interpretation of Moran's *I* easily translates to a bivariate scatterplot with *x* on the x-axis, *Wx* on the y axis and Moran's *I* being the slope of the linear best fit. The scatterplot is useful in identifying those observations that do not conform, and that differ significantly from the global Moran's *I* in magnitude and/or direction. The scatterplot centres on the point where the mean of the variable meets the mean of the lagged variable, and the four quadrants of the plot relative to this point give information about the type of association that is present. The upper right and lower left quadrants represent positive spatial association, while the upper left and lower right quadrants show those observations that have negative spatial association. The densities of each of the quadrants indicate which spatial pattern dominates and also provides

information on the distribution of the individual spatial associations and the contribution of each to the global statistic. Further, the plot shows outliers and leverage points and can provide an overall picture of the consistency of the global indicator. The Moran Scatterplot is related to the Local Moran's *I* below. There is a separate tool available on the portal under the Charts group specifically designed to output a Moran's *I* Scatter Plot, which has input parameters the same as those below.

- Input parameters

    o Dataset Input – the dataset that contains the variable to be tested.
    o Key Column – the key column of the dataset
    o Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above
    o Variable – the variable to be tested.
    o Alternative Hypothesis – indicates the alternative hypothesis; can be two.sided, greater, meaning one sided greater than or less, meaning one sided less than.
    o Inference – a tick box indicating the assumption under which the variance should be calculated. A tick indicates randomisation, blank indicates normality.

- Expected outputs

    o Text output with the following:
        ▪ Moran's *I* with the corresponding *Z*-score and *p*-value
        ▪ Moran's *I* with its expectation and variance
        ▪ Moran's *I* with its Saddlepoint approximation for the standard deviate and corresponding *p*-value
        ▪ Moran's *I* with its exact standard deviate and corresponding *p*-value
        ▪ The alternative hypothesis
        ▪ The assumption under which the variance was calculated
    o Dataset output with the following columns of the same length as the input variable:
        ▪ Key column
        ▪ Variable
        ▪ Lagged variable
        ▪ Variable scaled
        ▪ Lagged variable scaled

4.2 Moran's *I* Permutation Test

This component allows the user to compute the global Moran's *I* statistic by using random permutations of *x* for the given spatial weights matrix, "to establish the rank of the observed statistic in relation to the (number of) simulated values" (Bivand, 2013: 119). In this test the observed values are randomly assigned to areas and the *I* statistic is simulated for each permutation.

- Input parameters
    o Dataset Input – the dataset that contains the variable to be tested.
    o Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above
    o Variable – the variable to be tested.
    o Alternative Hypothesis – indicates the alternative hypothesis; can be greater, meaning one sided greater than or less, meaning one sided less than.

- o Number of Simulations – the number of simulations performed for the test.

- • Expected outputs
  - o Text output with the following:
    - ▪ Moran's *I* with rank and pseudo *p*-value
    - ▪ Moran's *I* with its bias and standard error
    - ▪ The alternative hypothesis
    - ▪ The number of simulations

4.3 Geary's *C*

Geary's *C* is computed by (Cliff and Ord, 1981):

$$(4) \qquad C = (N-1)\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(X_i - X_j)^2}{2W\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Where X_i and X_j are the values for regions *i* and *j* respectively, $\bar{X}$ is the mean of the variable, W is the global sum of the weights, i.e. $W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$. The spatial weight matrix that is employed by this component must be symmetric. Hence if the selected spatial weight matrix is not already symmetric, part of the process will include converting the spatial weight matrix so it becomes symmetric.

Given the null hypothesis is one of no global spatial autocorrelation, the expected value of Geary's *C* equals 1. If *C* is larger than 1, the distribution of the variable being tested is characterised by negative spatial autocorrelation. If *C* is smaller than 1, then the distribution of the variable displays positive spatial autocorrelation. As with Moran's *I*, inference is based on Z-scores:

$$(5) \qquad Z_C = \frac{C-1}{sd(C)}$$

While Moran's *I* is based on the cross-products of the deviations from the mean, Geary's *C* is based on the deviations in responses of each observation with one another. This in effect means that Moran's *I* is more sensitive to extreme values and is more a global measure, whereas Geary's *C* is more sensitive to differences between values in neighbouring areas.

- • Input parameters
  - o Dataset Input – the dataset that contains the variable to be tested.
  - o Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
  - o Variable – the variable to be tested.
  - o Alternative Hypothesis – indicates the alternative hypothesis; can be two.sided, greater, meaning one sided greater than or less, meaning one sided less than.
  - o Inference – a tick box indicating the assumption under which the variance should be calculated. A tick indicates randomisation, blank indicates normality.

- • Expected outputs
  - o Text output with the following:
    - ▪ Geary's *C* with the corresponding *Z*-score and *p*-value
    - ▪ Geary's *C* with its expectation and variance
    - ▪ The alternative hypothesis
    - ▪ The assumption under which the variance was calculated

4.4 Getis/Ord Global *G*

The focus of the family of *G* statistics derived by Getis and Ord (1992; Ord and Getis, 1995) is on spatial autocorrelation at the local level (local G_i statistics). However, they did include a global measure of spatial autocorrelation. The Global *G* is defined as (Getis and Ord, 1992: 194):

$$(6) \qquad G(d) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d) X_i X_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j}$$

Following the recommendation of Bivand (2013: 57) all spatial weight matrices will be converted to binary weights (style "B") if they are not already.

The *G(d)* statistic is concerned with the overall concentration or lack of concentration in all pairs that are neighbours given the definition of neighbouring areas. The variable must contain only positive values to be used.

- Input parameters
  - Dataset Input – the dataset that contains the variable to be tested.
  - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
  - Variable – the variable to be tested.
  - Alternative Hypothesis – indicates the alternative hypothesis; can be two.sided, greater, meaning one sided greater than or less, meaning one sided less than.

- Expected outputs
  - Text output with the following:
    - Global *G* with the corresponding *Z*-score and *p*-value
    - Global *G* with its expectation and variance
    - The alternative hypothesis

4.5 Bivariate Moran's *I*

This component uses the Moran's *I* global measure of spatial autocorrelation to measure the influence one variable has on the occurrence of another variable in close proximity. The original Moran's *I* statistic measured the degree of linear association of the values of a variable in neighbouring regions. The bivariate Moran's *I* measures the degree of linear association between one variable, *x*, and a different variable, *y*, in neighbouring regions.

The bivariate Moran's *I* statistic is computed as:

$$(7) \qquad I_t = \frac{R \sum_{i=1}^{R} \sum_{j=1}^{R} x_i y_j w_{ij}}{R_b \sum_{i=1}^{R} x_i^2}$$

where *R* is the number of regions in the dataset; R_b is the sum of the weights which simplifies to *R* if the spatial weight matrix is row-standardised; x_i is the first variable and is measured as deviation from the mean, i.e. $x_i = X_i - \bar{X}$; y_i is the second variable and is also measured as deviation from the mean, i.e. $y_i = Y_i - \bar{Y}$. Once again w_{ij} is the spatial weight element, as described previously.

In a similar way to the Moran's *I*, the bivariate Moran's *I* statistic can be expressed as a standardised normal *Z*-score for inference purposes, computed by:

$$(8) \qquad Z_I = \frac{I - E(I)}{sd(I)}$$

where *I* is Moran's *I* statistic, *E(I)* is the theoretical mean and *sd(I)* the theoretical standard deviation of Moran's *I* statistic.

Just as the Moran's *I* easily translated to a bivariate scatterplot (Moran scatterplot), the same can be done for bivariate Moran's *I*. The bivariate Moran's *I* statistic can be interpreted as the regression coefficient in a bivariate regression of the spatially lagged second variable, *Wy*, on the original variable, *x* (in deviations from the mean). The spatially lagged second variable, *Wy*, is the average of observations of the second variable at neighbouring locations, that is, locations for which $w_{ij} \neq 0$. For the scatterplot the first variable, *x*, is placed on the x-axis and the lagged version of the second variable, *Wy*, lies on the y-axis. Once again the slope of the linear best fit is the bivariate Moran's *I* value.

- Input parameters
    - Dataset Input – the dataset that contains the variables to be used for the test.
    - Key Column – the key column of the dataset.
    - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
    - X Variable – the first (or *x*) variable.
    - Y Variable – the alternative (or *y*) variable.
    - Alternative Hypothesis – indicates the alternative hypothesis; can be two.sided, greater, meaning one sided greater than or less, meaning one sided less than.
    - Inference – a tick box indicating the assumption under which the variance should be calculated. A tick indicates randomisation, blank indicates normality.

- Expected outputs
    - Text output with the following:
        - Bivariate Moran's *I* with the corresponding *Z*-score and *p*-value
        - Bivariate Moran's *I* with its expectation and variance
        - The alternative hypothesis
        - The assumption under which the variance was calculated
    - Dataset output with the following columns of the same length as the input variable:
        - Key column
        - X variable
        - Y variable
        - Lagged Y variable
        - X variable scaled
        - Lagged Y variable scaled

5.   Local Measures of Spatial Autocorrelation

While a dataset may reveal 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 measures of spatial association 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).

5.1　Local Moran's *I*

While the Moran scatterplot provides more detail on the type of spatial clustering, it does not report on the significance of such clustering. The local Moran's *I* statistic has been generated to do this. Anselin (1995) defines a class of local indicators he calls Local Indicators of Spatial Association (LISAs), of which local Moran's *I* is one. The local Moran's statistic offers insight into the behaviour of data at local levels, by providing a decomposition of the Moran's *I* global statistic into the degree of spatial association associated with each observation.

LISAs 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 a variable of interest amongst areas is extreme or is likely or expected, given a random geographic distribution of the variable.

Local Moran's *I* is calculated as:

(9) $$I_i = x_i \sum_j w_{ij} x_j$$

where the x_i and x_j are the observations in deviations from the mean and w_{ij} is the spatial weight matrix element. Positive values of I_i mean that there is a spatial cluster of similar values and negative values represent a spatial cluster of dissimilar values.

Significance testing of the local Moran statistics can be somewhat problematic. Unlike the G_i and G_i^* statistics (see below), the local Moran does not conform to a common distribution and so the test under a normality assumption should be treated with caution. Instead Anselin (1995) suggests a conditional randomisation or permutation approach to give so-called pseudo significance levels. Tiefelsdorf (2000) published the exact reference distribution of Moran's *I*, but in a later paper recommends an application of the Saddlepoint approximation, as it "outperforms other approximation methods with respect to its accuracy and computational costs" (Tiefelsdorf, 2002: 187). The Local Moran's *I* component in the Tools provides results for a normal distribution, a Saddlepoint approximation of the standard deviate and the exact standard deviate.

- Input parameters
  - Dataset Input – the dataset that contains the variable to be tested.
  - Key Column – the key column of the dataset.
  - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above
  - Variable – the variable to be tested.
  - Test – the type of approximation to the reference distribution to be conducted; either normal, saddlepoint or exact.
  - Significance Level – the significance level to be applied for the Local Moran's *I* map; can be 0.1, 0.05, 0.01 or 0.001.

- o Alternative Hypothesis – indicates the alternative hypothesis; can be two.sided, greater, meaning one sided greater than or less, meaning one sided less than.
    - o Spatial Weight Matrix Style – the style of the spatial weight matrix being used. Options are: binary, which can either be binary or absolute; row-standardised, which row-standardises the binary matrix; globally-standardised, which globally-standardises the binary matrix. If the user indicates an option different to how the spatial weight matrix was originally registered in the Data Cart, it will be transformed here to the style the user indicates, for this test only.

- • Expected outputs
    - o Text output referring the user to the dataset output
    - o Dataset output with the following columns of the same length as the input variable
        - ▪ Key column
        - ▪ Local Moran's *I* values
        - ▪ Local Moran's *I* standard deviates (or saddlepoint or exact approximation)
        - ▪ Local Moran's *I* *p*-values
        - ▪ Variable
        - ▪ Lagged variable
        - ▪ Variable scaled
        - ▪ Lagged variable scaled
        - ▪ Map group based on the significance level selected by the user, with regions having one of 5 classifications: 0 – not significant/null, 1 – high-high, 2 – high-low, 3 – low-high, 4 – low-low.
        - ▪ Map group based on the normal *p*-values and the significance level selected by the user, with regions having one of 5 classifications: 0 for regions that not significant/null, 1 for regions with high values surrounded by regions with high values (high-high), 2 for regions with high values with neighbours having low values (high-low), 3 for low-high regions and 4 for low-low regions.
        - ▪ Map group name based on the normal *p*-values and the significance level selected by the user. The following are the corresponding names of the group: 0 – "Not Significant", 1 – "High-High", 2 – "High-Low", 3 – "Low-High", 4 – "Low-Low".

To visualise the Local Moran's *I* map, the user must use the Spatialise Aggregated Dataset tool under the Spatial Data Manipulation group. Once the geometries have been added to the dataset, the user should choose a Choropleth map under the Map Visualisations group of Maps, Charts and Graphs. After selecting the correct column the user should select 5 for the Number of Classes, a palette type of Diverging and the Palette: GreyRedBlue.

5.2   Getis/Ord Local G_i

The G_i-statistics are known to be useful for identifying "hot and/or cold spots" and to check for heterogeneity in the dataset. G_i (Getis and Ord, 1992) statistics are the ratio of the sum of values in neighbouring locations, defined by a given distance, to the sum over all observations.

Like Local Moran's *I*, G_i statistics can detect local 'pockets' of dependence that may not show up when using global statistics, for example they isolate micro-concentrations in the data which are otherwise swamped by the data's overall randomness. The form of the X matrix is additive, either (X_j) or $(X_i + X_j)$ for the self-included measure. This contrasts with the Moran statistic where the matrix is in the form $(X_i - \bar{X})(X_j - \bar{X})$ or Geary's *C* which specifies the matrix in the form $(X_i - X_j)^2$. As G_i measures are measures of concentration, they are in contrast to the Moran statistic which examines the correlation or covariance of values in neighbouring regions compared to the data's overall variance (while Geary's C calculates differences). G_i statistics evaluate association by examining 'additive qualities' (Getis and Ord, 1996, p. 262), they compare local weighted averages to global averages to isolate 'coldspots' and 'hotspots'.

The interpretation of G_i is also somewhat different from other measures of spatial association. When G_i is larger than its expected value, the overall distribution of the variable being measured can be seen as characterised by positive spatial autocorrelation, with high value clusters prevalent. If G_i is smaller than its expected value, then the overall distribution of the variable being measured is still characterised by positive spatial autocorrelation, but with low values clustered together.

A special feature of this statistic is that it equals zero when all *X* values are the same. Also while the weighted value of an *X* value might be expected to rise with the number of neighbours (or weighted regions), all else equal, a region that has a greater number of neighbours does not receive a greater G_i. Only when the observed estimate in the vicinity of the region *i* varies significantly from the mean of the variable does G_i rise (Getis and Ord, 1992).

A slightly different form of G_i was suggested by Ord and Getis (1995). $G_i(d)$, originally proposed for elements of a symmetric binary weights matrix, was extended to variables that do not have a natural origin and to non-binary standardised weight matrices (Ord and Getis, 1995: 289): This statistic for each region *i* is:

(10) $$G_i(d) = \frac{\sum_j w_{ij}(d)X_j - W_i\mu}{\sigma\{[(n-1)S_{1i} - W_i^2]/(n-1)\}^2} , \qquad j \neq i$$

where w_{ij} is the spatial weight matrix element, X_j is the variable, $W_i = \sum_j w_{ij}$, $S_{1i} = \sum_j w_{ij}^2$ and μ and σ are the usual sample mean and standard deviation for the sample size n-1. *d* is the threshold distance from i. $G_i^*(d)$ includes the case where $j = i$.

The Getis-Ord local G_i are statistics for local spatial association but are not LISAs given the criteria established by Anselin (1995). Their individual components are not related to the global statistic of spatial association (*G*). Anselin notes that "this requirement is not needed for the identification of local spatial clusters, but it is important for the second interpretation of a LISA, as a diagnostic of local instability in measures of global spatial association (for example in the presence of significant global association)" (Anselin, 1995: 102).

The results firstly produce the G_i for each area *i* as a standardised *Z*-score. Getis and Ord (1992) argued that inference, as with global measures are based on calculating a standardised value and comparing this against a null which is assumed to follow a normal distribution. However a normally distributed null may not be an appropriate assumption, as Local G_i are not independent of each other by design (Ord and Getis, 1995). By definition one region may appear in a number of different region's weighting vectors. This raises the general issue that for local measures of spatial autocorrelation inference is complicated as statistics will be correlated when weights contain the same elements which they do. This is a problem of multiple statistical comparison and reflects "the built-in correlatedness of measures for

adjoining locations" (Anselin, 1995: 112). This requires a more stringent test to be able to assert spatial non-randomness, that is, to assert the presence of spatial autocorrelation at the local level. Anselin (1995: 96) notes "This means 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 α/m (Bonferroni) or $1 - (1 - \alpha)^{1/m}$ (Sidák)." Ord and Getis (1995) argue that the choice of a standard measure is not clear cut, although the conservative approach is to apply the Bonferroni adjustment.

The component allows users to compute both the G_i and G_i^* (the self-included measure) statistics and provides a Z-score with both a normal and Bonferroni p-value.

- Input parameters
  - Dataset Input – the dataset that contains the variable to be tested.
  - Key Column – the key column of the dataset.
  - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
  - Variable – the variable to be tested.
  - Matrix Type – tick indicates the inclusion of the self-weight $w_{ii} > 0$ in the matrix so the test becomes G_i^* (see above).
  - Significance Level – the significance level to be applied for the Local G map; can be 0.1, 0.05, 0.01 or 0.001.

- Expected outputs
  - Text output referring the user to the dataset output
  - Dataset output with the following columns of the same length as the input variable
    - Key column
    - Variable
    - Local G_i Z-scores
    - Local G_i p-values (normal)
    - Local G_i p-values (bonferroni adjusted)
    - Map group based on the normal p-values and the significance level selected by the user, with regions having one of 3 classifications: 0 for regions that not significant/null, 1 for regions that show a concentration of high values, 2 for regions that show a concentration of low values.
    - Map group name based on the normal p-values and the significance level selected by the user. The following are the corresponding names of the group: 0 – "Not Significant", 1 – "High", 2 – "Low".
    - Map group based on the bonferroni adjusted p-values and the significance level selected by the user, with the same three classifications as for the normal distribution.
    - Map group name based on the bonferroni p-values and the significance level selected by the user. The same corresponding names apply: 0 – "Not Significant", 1 – "High", 2 – "Low".

As with the Local Moran's I, to visualise the Local G_i map, a user must first add geometries to the output dataset by using the Spatialise Aggregated Dataset under the Spatial Data Manipulation. Then in the Choropleth option of Map Visualisations, the user must choose 3 for the Number of Classes, and then once again choose a palette type of Diverging and a palette GreyRedBlue.

6. Spatial Regression

When employing data collected across various geographic points, it may not be appropriate to view it as conceptually identical to cross-sectional data on individuals or businesses at a single location. Spatially adjacent observations are likely to exhibit spatial interdependence, owing to dynamics which accompany proximity. This reflects Tobler's (1970) maxim that 'everything is related to everything else but near things are more related than distant things'. The presence of spatial autocorrelation in data used in ordinary least squares (OLS) models violates standard statistical techniques that assume independence among observations. Ignoring this dependence between neighbouring regions will lead to inefficient and/or biased regression results (Anselin, 1988a).

6.1 Spatial Autocorrelation Diagnostic Tests

There are a number of diagnostic tests that can be performed to test whether an ordinary least squares (OLS) regression is the most appropriate type of model to use. These tests are carried out on the results of an OLS regression and can determine whether the data being modelled violates the OLS assumptions of independence among observations. Further they also provide insight into which of the spatial econometric models (explained in the next section) is the most appropriate to use.

6.1.1 Moran's *I* on residuals

The first of our spatial diagnostics performs a Moran's *I* test on the residuals of a linear regression to determine whether the error terms are spatially correlated. This Moran's I test treats the residuals slightly differently to how the original Moran's *I* test treats a variable (Cliff and Ord, 1973):

(11) $\quad I = e'We/e'e$

where e is the regression residuals and W the spatial weight matrix. As with the original Moran's I, 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, 1988a: 102). A significant result would indicate that there are indeed spatially correlated error terms.

To implement this component the user must outline the regression equation they want to test.

- Input parameters
    - Dataset Input – the dataset that contains the variables to be used in the regression.
    - Key Column – the key column of the dataset.
    - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
    - Regression Dependent Variable – the dependent variable of the regression equation.
    - Regression Independent Variables – the independent variable(s) of the regression equation.
    - Alternative Hypothesis – indicates the alternative hypothesis; can be two.sided, greater, meaning one sided greater than or less, meaning one sided less than.

- Expected outputs

- o Text output with the following:
    - ▪ Moran's *I* with the corresponding *Z*-score and *p*-value
    - ▪ Moran's *I* with its expectation and variance
    - ▪ The alternative hypothesis

6.1.2 Lagrange Multiplier tests

Anselin (1988b) proposed Lagrange Multiplier diagnostic tests for OLS models where the focus is on detecting model misspecification due to spatial dependence in one of two forms. The first of these tests whether there is an omitted spatially lagged dependent variable (LMlag), while the second tests for spatial error autocorrelation (LMerror). Anselin *et al* (1996), building on the work from Bera and Yoon (1993), extended this to include more robust tests that include the possible presence of the other form of spatial dependence. Specifically these are a test for an omitted spatially lagged dependent variable in the possible presence of spatial error autocorrelation (RLMlag), as well as a test for spatial error autocorrelation in the possible presence of a spatially lagged dependent variable (RLMerror). In addition they derive a portmanteau test that incorporates both aspects by combining the test for spatial error autocorrelation with the test for an omitted spatially lagged dependent variable in the presence of spatial error autocorrelation (SARMA). The tests are all based on the results of ordinary least-squares (OLS) estimation.

If we use the notation which is explained fully in the next section, the LM test for an omitted spatially lagged dependent variable tests that $\rho = 0$, while the LM test for error autocorrelation tests whether $\lambda = 0$. LM tests can be used then to determine which of the models below is the more appropriate (see for example Elhorst, 2010) If, for example we find that indeed $\lambda = 0$, but $\rho \neq 0$ then the appropriate model may be the spatial lag model (see next section):

(12) $y = \rho W y + X\beta + \varepsilon$

The LM statistics follow a chi-squared distribution with the number of degrees of freedom being the number of coefficients being tested. Hence the LM statistics LMlag, LMerror, RLMlag and RLMerror are asymptotically chi-squared distributed with one degree of freedom. This would mean that, for the LMlag test for example, we reject the null hypothesis of not requiring a spatially lagged dependent variable if the test statistic exceeds 3.84 at the 5 per cent level. The SARMA is asymptotically chi-squared distributed with 2 degrees of freedom.

As with the previous component to implement this component the user must outline the regression equation they want to test.

- • Input parameters
    - o Dataset Input – the dataset that contains the variables to be used in the regression.
    - o Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above
    - o Regression Dependent Variable – the dependent variable of the regression equation.
    - o Regression Independent Variables – the independent variable(s) of the regression equation.

- • Expected outputs
    - o Text output with the following:
        - ▪ LMerror statistic, degrees of freedom and *p*-value

- LMlag statistic, degrees of freedom and *p*-value
- RLMerror statistic, degrees of freedom and *p*-value
- RLMlag statistic, degrees of freedom and *p*-value
- SARMA statistic, degrees of freedom and *p*-value

6.2   Spatial Econometric Models

There are a suite of spatial econometric models that have been developed that attempt to overcome the spatial autocorrelation problem and allow econometric modelling to produce unbiased and efficient estimates. Anselin (1988a: 34) introduced the model specification which applies to "situations where observations are available for a cross-section of spatial units, at one point in time." These models are calculated using maximum likelihood estimation.

The issue of model selection techniques (specification strategies) remains contentious in the spatial econometric literature. 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 and the suite of spatial econometric models would not be required.

The alternative approach to 'filtering' can be cast in the broader debate common among time series econometricians. Two options appear possible. First, we would proceed with a specific-to-general approach (the so-called 'classical approach'), which begins with the most simple OLS regression and then uses appropriate Lagrange Multiplier (LM) tests to test a range of 'added variables' including the presence of spatial dependence. In this case, the researcher would ultimately choose the model with some highest test value. 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.

Obviously, the user is free to use the components in whatever order they deem fit for their purpose. We would assume most spatial econometricians use the specific-to-general approach whereby they begin with the linear regression model.

6.2.1 OLS model

The OLS model is not strictly part of the Spatial Statistics Tools, as it can be found under Statistics in the portal. However, it is the traditional starting point for model selection. This model assumes no spatial effects – the presence of spatial effects would produce biased estimates:

(13) $y = X\beta + \varepsilon$

$\varepsilon \sim N(0, \sigma^2 I_n)$

where *y* is the *n* x *1* dependent variable vector (*n* being sample size), *X* the *n* x *k* matrix of independent variables (*k* being the number of independent variables), *β* the *k* x *1* vector of parameter estimates associated with the independent variables and *ε* the error term which is normally distributed with error variance matrix $\sigma^2 I_n$.

6.2.2 First order spatial autoregressive model

In this model we introduce spatial effects, by including just a spatially lagged dependent variable:

(14) $y = \rho Wy + \varepsilon$

     $\varepsilon \sim N(0, \sigma^2 I_n)$

where W is the n x n spatial weight matrix and ρ is the coefficient of the spatially lagged dependent variable, Wy.

- Input parameters
    - Dataset Input – the dataset that contains the variable to be used in the spatial regression.
    - Key Column – the key column of the dataset.
    - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above
    - Dependent Variable – the dependent variable of the regression equation – also the only variable used in first order spatial lag.

- Expected outputs
    - Text output with the following:
        - The estimate for the coefficient of the spatially lagged dependent variable, ρ (rho), its standard error, Z-score and p-value
        - The estimates for the intercept, its standard error, Z-score and p-value
        - Log Likelihood (LL)
        - Sigma squared (σ^2)
        - Asymptotic coefficient covariance matrix for (σ and ρ)
        - Sigma (σ)
        - Akaike Information Criterion (AIC)
        - Likelihood Ratio (LR) test result
        - Wald test result and p-value
        - LM test on the residuals of the spatial lag regression
    - Dataset output with the following columns of the same length as the input variable
        - Key column
        - Dependent variable
        - Residuals

6.2.3 Spatial Lag Model

Also called the mixed regressive-spatial autoregressive model, we alter the first-order spatial autoregressive model by introducing a matrix of independent variables. This model combines the matrix of explanatory variables with a vector of coefficient parameters, with the spatially lagged dependent variable from the previous model:

(15) $y = \rho Wy + X\beta + \varepsilon$

     $\varepsilon \sim N(0, \sigma^2 I_n)$

In this model, the spatially lagged dependent variable is correlated with the error terms (see Anselin, 1988a: 58).

- Input parameters
    - Dataset Input – the dataset that contains the variables to be used in the spatial regression.
    - Key Column – the key column of the dataset.
    - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.

- Dependent Variable – the dependent variable of the regression equation.
- Independent Variables – the independent variable(s) of the regression equation.

- Expected outputs
  - Text output with the following:
    - The estimate for the coefficient of the spatially lagged dependent variable, ρ (rho), its standard error, Z-score and p-value
    - The estimates for the regression coefficients, their standard error, Z-score and p-value
    - Log Likelihood (LL)
    - Sigma squared (σ^2)
    - Asymptotic coefficient covariance matrix for (σ, ρ and β)
    - Sigma (σ)
    - Akaike Information Criterion (AIC)
    - Likelihood Ratio (LR) test result
    - Wald test result and p-value
    - LM test on the residuals of the spatial lag regression
  - Dataset output with the following columns of the same length as the input variable
    - Key column
    - Dependent variable
    - Independent variable(s)
    - Residuals

6.2.4 The spatial error model

The spatial error model, as Anselin (1988a: 35) points out combines "the linear regression model with a spatial autoregressive disturbance" as follows:

(16) $y = X\beta + \mu$

$\mu = \lambda W\mu + \varepsilon$

$\varepsilon \sim N(0, \sigma^2 I_n)$

where μ is the disturbance term, which has its own spatial autoregressive structure and λ is the coefficient of the spatially lagged error term. The presence of the spatially autocorrelated error term "is analogous to the serial correlation problem in time series models" (LeSage, 1998: 50). This model is traditionally used if it is determined that there is spatial autocorrelation in the residuals of a standard least squares regression.

- Input parameters
  - Dataset Input – the dataset that contains the variables to be used in the spatial regression.
  - Key Column – the key column of the dataset.
  - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
  - Dependent Variable – the dependent variable of the regression equation.
  - Independent Variables – the independent variable(s) of the regression equation.

- Expected outputs
  - Text output with the following:

- The estimate for the coefficient of the spatially lagged error term, λ (lambda), its standard error, Z-score and *p*-value
- The estimates for the regression coefficients, their standard error, Z-score and *p*-value
- Log Likelihood (LL)
- Sigma squared (σ^2)
- Asymptotic coefficient covariance matrix for (σ, λ and β)
- Sigma (σ)
- Akaike Information Criterion (AIC)
- Likelihood Ratio (LR) test result
- Wald test result and *p*-value
  - Dataset output with the following columns of the same length as the input variable
    - Key column
    - Dependent variable
    - Independent variable(s)
    - Residuals

6.2.5 The general spatial autoregressive model

This model combines the spatial lag and the spatial error models, giving us a spatially lagged dependent variable as well as a spatially lagged error term:

(17) $\quad y = \rho W_1 y + X\beta + \mu$

$\quad \mu = \lambda W_2 \mu + \varepsilon$

$\quad \varepsilon \sim N(0, \sigma^2 I)$

where $W_1 = W_2$ or $W_1 \neq W_2$. This is sometimes known as the Kelejian-Prucha model, as they advocate this model throughout their work (Kelejian and Prucha, 1998).

- Input parameters
  - Dataset Input – the dataset that contains the variables to be used in the spatial regression.
  - Key Column – the key column of the dataset.
  - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
  - Dependent Variable – the dependent variable of the regression equation.
  - Independent Variables – the independent variable(s) of the regression equation.

- Expected outputs
  - Text output with the following:
    - The estimate for the coefficient of the spatially lagged error term, λ (lambda), its standard error, Z-score and *p*-value
    - The estimates for the regression coefficients, their standard error, Z-score and *p*-value
    - Log Likelihood (LL)
    - Sigma squared (σ^2)
    - Sigma (σ)
    - Akaike Information Criterion (AIC)
    - Likelihood Ratio (LR) test result
    - Asymptotic coefficient covariance matrix for (σ, ρ, λ and β)

- Dataset output with the following columns of the same length as the input variable
  - Key column
  - Dependent variable
  - Independent variable(s)
  - Residuals

6.2.6 The spatial Durbin model

This model incorporates spatially lagged independent variables, while excluding a spatially lagged error term:

(18) $y = \rho W y + X\beta + W X\theta + \varepsilon$

$\varepsilon \sim N(0, \sigma^2 I)$

where *WX* is the spatially lagged independent variables and *θ* its vector of coefficients.

- Input parameters
  - Dataset Input – the dataset that contains the variables to be used in the spatial regression.
  - Key Column – the key column of the dataset.
  - Spatial Weight Matrix – the spatial weight matrix to be used, probably derived from one of the methods above.
  - Dependent Variable – the dependent variable of the regression equation.
  - Independent Variables – the independent variable(s) of the regression equation.

- Expected outputs
  - Text output with the following:
    - The estimate for the coefficient of the spatially lagged error term, ρ (rho), its standard error, Z-score and p-value
    - The estimates for the regression coefficients, their standard error, Z-score and p-value
    - Log Likelihood (LL)
    - Sigma squared (σ^2)
    - Asymptotic coefficient covariance matrix for (σ, λ and β)
    - Sigma (σ)
    - Akaike Information Criterion (AIC)
    - Likelihood Ratio (LR) test result
    - Wald test result and p-value
    - LM test on the residuals of the spatial Durbin regression
  - Dataset output with the following columns of the same length as the input variable
    - Key column
    - Dependent variable
    - Independent variable(s)
    - Residuals

6.3   Other diagnostic tests

We saw previously some diagnostic tests that can be used to determine whether there is spatial dependence in the data and also to help point the user in the direction of the most appropriate spatial econometric model to use. Those diagnostic tests use the results from an

OLS regression. There are other diagnostic tests that require one or other of the spatial econometric models to be estimated and use its results to similarly assist in determining whether there is spatial dependence in the data. These are a Likelihood Ratio test and a Wald Test. As the test requires the spatial econometric model to have been estimated, the results for these diagnostic tests are contained in the results for the specific spatial econometric model above, but are more fully explained here.

The Likelihood Ratio (LR) test compares the Log Likelihood of the OLS model to the Log Likelihood of the specific spatial econometric model in the following way:

(19) $LR = -2LL_{null} + 2LL_{alter}$

where LL_{null} is the Log Likelihood of the null model (OLS) and LL_{alter} is the Log Likelihood of the alternative model. This statistic is asymptotically chi-squared distributed with the number of degrees of freedom being the number of additional parameters in the alternative model.

The Wald test uses the maximum likelihood estimate of the extra parameter (ρ or λ) and tests how far they are from zero in standard errors. Like the LR test this statistic is asymptotically chi-squared distributed with the number of degrees of freedom being one. If the test fails to reject the null hypothesis, this suggests that not including the spatially lagged dependent variable or the spatially error term will not substantially harm the fit of that model, since a predictor with a coefficient that is very small relative to its standard error is generally not doing much to help predict the dependent variable.

A final diagnostic test whose results rely on the estimation of a spatial econometric model is a Lagrange Multiplier test for spatial error autocorrelation in the residuals of the spatial econometric model. This is available for the spatial lag model and the spatial Durbin model and tests whether the corresponding model removes the spatial autocorrelation from the residuals. This is similar to the Lagrange Multiplier tests. It is asymptotically chi-squared distributed with one degree of freedom.

7.   Spatial Indices

Spatial indices are used to investigate characteristics about particular regions across larger areas and can be used to compare regions to each other. They typically give information on a particular attribute of a region compared to other regions in the dataset. Spatial indices are not concerned with location of regions compared to other regions, but indeed may provide insight into the locality of certain characteristics.

7.1   Diversity Index

A diversity index measures the degree of specialisation or alternatively the degree of diversity across attributes within a spatial unit. This allows spatial units to be compared as to the mix of the characteristics being measured.

The diversity index is calculated as follows (Tidswell and Barker, 1971):

(20) $SDI = \sqrt{p_1{}^2 + p_2{}^2 + \cdots + p_n{}^2}$

where p_n is the variable of interest as a proportion of the total of the n variables, for the spatial unit. A common use of the diversity index in economics is to measure the industrial diversity in regions, where p_n for a particular region becomes employment in industry n as a proportion of total employment in the region.

The diversity index for a particular region can range from zero to one, where a score approaching zero indicates an increasing degree of specialisation and a score approaching one indicates an increasing degree of specialisation. Determination of whether a region has a high or low diversity index is done by comparing the diversity index scores across all regions.

- Input parameters
  - Dataset Input – the dataset that contains the variable to be tested.
  - Key Column – the key column of the dataset
  - Variables – the variables that are to be included in the calculation of the index.

- Expected outputs
  - Text output referring the user to the dataset output
  - Dataset output with the following columns of the same length as the input variable:
    - Key column
    - All variables
    - Diversity Index column

7.2 Location Quotient

A location quotient compares the proportion of a characteristic in one region to the proportion of that characteristic that is present across all regions.

The location quotient is calculated as:

$$(21) \quad LQ = \frac{p_i}{p} \Big/ \frac{P_i}{P}$$

where p_i is the variable of interest i for the region, p is the total of all variables for the region; P_i is the variable of interest i for the entire area and P is the total of all variables for the entire area. A common use of location quotient is to determine the proportional employment in an industry for a region compared to the national benchmark. In this case p_i would be the employment level in industry i for the region, p would be total employment in the region, P_i the employment level in industry i for the nation and P the total employment of the nation.

The location quotient score for a region can theoretically be anything over 0 (0 implying the variable of interest is 0 for the region). The higher the score implies the proportion of the characteristic in the reference region is greater compared to the characteristic in the entire area (nation). As well as the score, location quotients are placed into intervals to give the user a quick view as to where each region's location quotient lies. In the Tools, the user is required to determine these ranges.

- Input parameters
  - Dataset Input – the dataset that contains the variable to be tested.
  - Key Column – the key column of the dataset.
  - Region Variables – the column(s) in the dataset that corresponds to the variable (e.g. particular industry of employment) that is to be tested for each region (e.g. regional industry employment).
  - Region Total – the column in the dataset that corresponds to the total for each region (e.g. total regional employment).
  - Number of Intervals – the location quotient result for each region is placed into an interval range. This is the number of interval upper limits applicable.
  - Upper Limits of Intervals – this is the upper limits of each of the intervals. Enter as a comma-separated string. The number of upper limits must

correspond with the Number of Intervals input parameter. Scores above the highest upper limit are placed in an interval with the upper limit of the highest location quotient.

- Expected outputs
  - Text output referring the user to the dataset output
  - A dataset with the following columns of the same length as the input variables
    - Key column
    - For each industry:
      - Location quotient
      - Location quotient range lower limit
      - Location quotient range upper limit
      - Location quotient range as string (for the purpose of choropleth mapping if required)

7.3   Shift Share

Shift Share analysis is a technique used to decompose the changes in a particular variable of interest to determine the influence of decomposed parts into the change.

It is commonly used to decompose regional employment growth into three components: (a) a national share (growth) effect, *NS*, being the part of change attributable to overall national trends; (b) an industry mix (proportional) effect, *IM*, being the part of change attributable to the industrial composition or mix of the region; and a (c) regional share (differential or competitive effect), *RS*, being the part of change attributable to regional advantage or competitiveness (Stimson *et al.*, 2002).

The three components are calculated as follows:

$$(22) \quad NS_i = e_{i,t-1}\left(\frac{E_t}{E_{t-1}} - 1\right)$$

$$(23) \quad IM_i = e_{i,t-1}\left(\frac{E_{i,t}}{E_{i,t-1}} - \frac{E_t}{E_{t-1}}\right)$$

$$(24) \quad RS_i = e_{i,t-1}\left(\frac{e_{i,t}}{e_{i,t-1}} - \frac{E_{i,t}}{E_{i,t-1}}\right)$$

Where e_i and E_i respectively are regional and national employment in industry i; e and E respectively are regional and national total employment in all industries; and t-1 is the initial period and t the end period of the analysis.

The total shift share for a region is then the sum of each of the components:

$$(25) \quad \Delta e_i = e_{i,t} - e_{i,t-1} = NS_i + IM_i + RS_i$$

- Input parameters
  - Dataset Input – the dataset that contains the variable to be tested.
  - Key Column – the key column of the dataset.
  - Regional Industry Employment t-1 – the columns in the dataset that correspond to regional industry employment at time t-1 for each industry.
  - Regional Industry Employment t – the columns in the dataset that correspond to the regional industry employment at time t for each industry.

- Expected outputs
  - Text output referring the user to the dataset output

- A dataset with the following columns of the same length as the input variables
  - Key column
  - For each industry:
    - National Share (NS)
    - Industry Mix (IM)
    - Regional Share (RS)
    - Total Shift (TS)

8. Measures of concentration

The following measures of concentration can be used to measure the overall similarity or disparity between different regions within a system. The measures are comparable with themselves only.

8.1 Theil Index

The Theil Index is a common measure of concentration and dispersion. It has historically been used to measure income inequality, but in effect can be used to measure the dispersion of any variable across regions relative to the whole.

The Theil index is calculated as follows (Akita and Kawamura, 2002):

$$(26) \quad T = \sum_i \left(\frac{y_i}{Y}\right) \ln \left(\frac{y_i/Y}{n_i/N}\right)$$

Where y_i is the level of income for region i, Y is total income in all regions, n_i is the population in region i, and N is the total population in all regions.

To use for labour measures, for example unemployment, replace y_i and Y with u_i and U respectively, where u_i is unemployment in region i and U is total unemployment in all regions. Then replace n_i and N with lf_i and LF respectively, where lf_i is labour force in region i and LF is total labour force in all regions. This is equivalent to saying the Theil Index compares the distribution across the population by summing, across regions, the weighted natural logarithm of the ratio between each region's unemployment share and that of the population.

The Theil Index can range from zero to $\ln(n)$. It is common to standardise the Theil Index so it can range between zero and one, which is done by dividing by $\ln(n)$. A Theil Index of zero indicates perfect equality and every region has the same unemployment – mean unemployment for the population. Conversely, a standardised Theil Index value of one represents a state of perfect inequality, where one region has all the unemployment.

- Input parameters
  - Dataset Input – the dataset that contains the variable to be used in calculating the Theil Index.
  - Numerator – the variable that takes the value of the numerator, or y in the above formula.
  - Denominator – the variable that takes the value of the denominator, or n in the above formula.

- Expected outputs
  - Text output with the following:
    - The Theil Index value
    - The standardised Theil Index value

8.2   Gini Coefficient

Similar to the Theil Index, the Gini Coefficient is a common measure of inequality of a distribution, also predominantly used to measure income distribution.

The Gini Coefficient compares the area under a Lorenz curve to the area under a perfect distribution line. It is defined as the ratio of the area between the Lorenz curve and the diagonal of perfect equality, to the area of the triangle below this diagonal as shown in the diagram below. The Lorenz curve is a graphical representation of the cumulative distribution function of some variable, generally income.

Source: http://www3.nccu.edu.tw/~jthuang/Gini.pdf
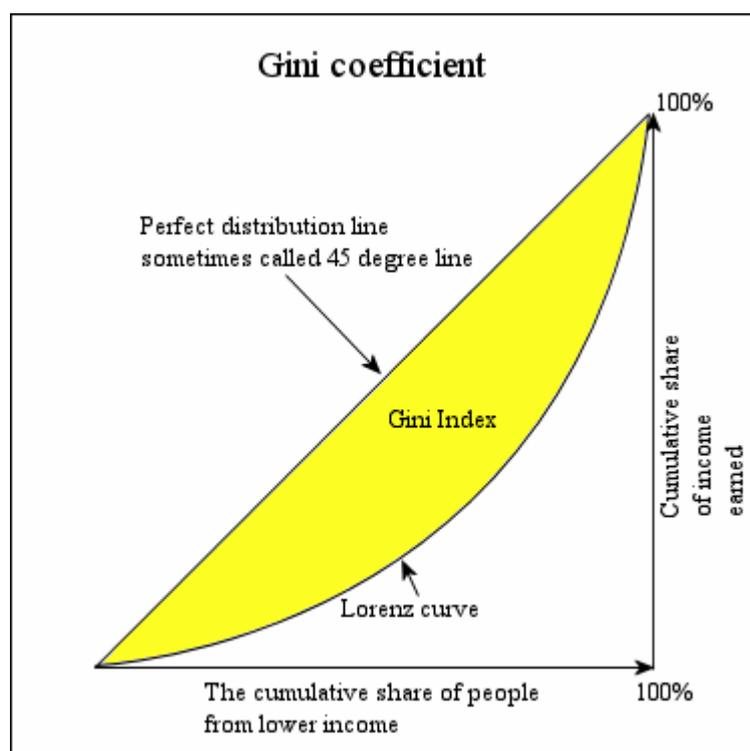
The formula for the Gini Coefficient from Sen as cited in Anand (1983):

(27) $G = \frac{1}{n}\left(n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n}y_i}\right)$

where the values y_i, $i = 1$ to n are the income levels indexed in non-decreasing order and n is the population size.

For a random sample the numerator is the sample size less one. Hence, for a measure of inequality of unemployment across regions the Gini Coefficient becomes:

(28) $G(S) = \frac{1}{r-1}\left(r + 1 - 2\frac{\sum_{i=1}^{r}(r+1-i)u_i}{\sum_{i=1}^{r}u_i}\right)$

where u_i, $i = 1$ to r are the unemployment rates indexed in non-decreasing order and r is the number of regions. This is a consistent estimator of the Gini Coefficient, though not an unbiased one.

As the tool is set up to deal with variables that are counts from the Census, the variable u_i requires a numerator and a denominator to calculate the unemployment rate for each region. Hence the numerator would be the count of unemployed for each region and the denominator would be the count of the labour force for each region. If the tool is to be used with a variable that doesn't require adjustment into this form it is suggested to enter the variable as the numerator and the denominator a variable consisting of a column of 1s.

The Gini Coefficient can range from a value of zero to one. Where there is perfect equality, the Gini Coefficient is zero and it would imply a Lorenz curve that follows the perfect distribution line. A Gini Coefficient of one implies perfect inequality where one region has all the unemployment for example.

- Input parameters
    - Dataset Input – the dataset that contains the variables to be used in calculating the Gini Coefficient.
    - Numerator – the variable that takes the value of the numerator for the u_i variable in the above formula.
    - Denominator – the variable that takes the value of the denominator for the u_i variable in the above formula.

- Expected outputs
    - Text output with the following:
        - The Gini Coefficient value for a population
        - The Gini Coefficient value for a sample

8.3 Dissimilarity Index

The Dissimilarity Index is a measure of the evenness with which two mutually exclusive groups are distributed across geographic regions. It has predominantly been used to measure the degree to which ethnic groups are distributed within regions across a larger area, or segregation.

The Dissimilarity Index is defined as follows:

$$(29) \quad D = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{g1_i}{G1} - \frac{g2_i}{G2} \right|$$

where $g1_i$ is the population of group 1 in region i, $g2_i$ the population of group 2 in region i, $G1$ is the total population of group 1 across all regions and $G2$ the total population of group 2 across all regions. In calculating the Dissimilarity Index all regions that comprise the larger area must be included in the dataset.

The Dissimilarity Index can have values between zero and one. A Dissimilarity Index of zero indicates the populations are evenly distributed across all regions, while a value of one means there is complete segregation between the two groups.

- Input parameters
    - Dataset Input – the dataset that contains the variable to be used in calculating the Dissimilarity Index.
    - Variable One – the variable that takes the value of $g1$ in the above formula.
    - Variable Two – the variable that takes the value of $g2$ in the above formula.

- Expected outputs
    - Text output with the Dissimilarity Index value

9.   Concordance e-Tool

As part of this project, there is a concordance e-tool set up accessible through the CofFEE website. This allows users to convert data from one spatial unit into another, using the standard ABS spatial units and concordance files. The website can be accessed at the following address: e1.newcastle.edu.au/coffee/concordance/ and there is also a link on the AURIN website. As this is on a different domain to the AURIN portal and the setup is very different to the AURIN portal, there is a separate help file for this e-tool. There is a link to the help file at the address above or it can be downloaded from here: http://e1.newcastle.edu.au/coffee/concordance/R-gui/docs/CofFEE_AURIN_Concordance_e-Tool_Help_File.pdf.

10.   Conclusion

This help file has described the Spatial Statistics and Econometric e-Tools section of the AURIN website. The Tools provide users with an interface where they can carry out spatial statistics and econometric analysis. The Tools allows users to upload their own data or utilise the many datasets already available on the AURIN portal and conduct analyses without requiring knowledge of computer languages or commands. It is hoped this capability of the Tools encourages prospective users who may not have accessed this area of statistics and econometrics to discover the potential of it.

In discussing each of the components that are available in the Tools we have placed them in a theoretical context, explaining how they fit into the complete spatial analysis. For each, this has been followed by a practical outline of the requirements and results of each in their implementation in the Tools. It cannot be overstated however that the success of their use relies on the user having a sound understanding of the principles behind them and knowledge of how and when to apply each.

Following this help file is a tutorial which presents an example of using the Tools applying data that is on the AURIN portal.

References

Akita, T. and Kawamura, K. (2002) 'Regional income inequality in China and Indonesia: A comparative analysis', Paper presented at the 46th Congress of European Regional Science Association (ERSA), Dortmund, August.

Anand, S. (1983) *Inequality and Poverty in Malaysia: Measurement and Decomposition*, A World Bank Research Publication, Oxford University Press, New York.

Anselin, L. (1988a) *Spatial econometrics: Methods and models*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (1988b), Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity, Geographical Analysis 20, 1-17.

Anselin, L. (1992) *SpaceStat Tutorial*, University of Illinois, Urbana-Champaign.

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'. pp. 111–125 in Fischer, M. M., Scholten, H.J. and Unwin, D. (eds) *Spatial Analytical Perspectives on GIS*, Taylor & Francis, London.

Anselin, L. and Bera, A. (1998) 'Spatial dependence in linear regression models with an introduction to spatial econometrics', in Ullah, A. and Giles, D. (eds.) *Handbook of applied economic statistics*, Marcel Dekker, New York, 237–89.

Anselin, L., Bera, A.K., Florax, R. and Yoon, M.J. (1996) 'Simple diagnostic tests for spatial dependence', *Regional Science and Urban Economics*, 26, 77-104.

Barff, R.A. and Knight, P.L (1988) 'Dynamic shift-share analysis', *Growth and Change*, 19, 1-10.

Bera, A. and Yoon, M. (1993) 'Specification testing with locally misspecified alternatives', *Econometric Theory*, 9, 649-658.

Bill, A. (2005) 'Neighbourhood inequality – do small area interactions influence economic outcomes', paper presented at the *State of Australian Cities National Conference*, Brisbane, December.

Bivand, R. (2013) 'Spatial dependence: weighting schemes, statistics and models', *CRAN project*, downloaded from http://cran.r-project.org/web/packages/spdep/spdep.pdf, on 28 March 2013.

Bivand, R., Pebesma, E.J. and Gomez-Rubio, V. (2008) *Applied Spatial Data Analysis with R*, Springer, New York.

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., Ord, J. K. (1981) *Spatial Processes*, Pion, London

Cressie, N. (1993) *Statistics for Spatial Data*, Wiley, New York.

Debelle, G. and Vickery, J. (1999) 'Labour Market Adjustment: Evidence on interstate labour mobility', *Australian Economic Record*, 77(238), September, 252-269.

Elhorst, J.P. (2010) 'Applied spatial econometrics: raising the bar', *Spatial Economic Analysis*, 5(1), 9-28.

Getis, A. (1995) 'Spatial filtering in a regression framework: examples using data on urban crime, regional inequality, and government expenditures', in Anselin, L. and Florax, R. (eds.) *New Directions in Spatial Econometrics*, Springer, Berlin.

Getis, A. and Ord, K. (1992), 'The Analysis of Spatial Association by Use of Distance Statistics', *Geographical Analysis*, 24, 189-206.

Getis, A. and Ord, J.K. (1996) 'Local spatial statistics: an overview', in Longley, P. and Batty, M. (eds.) *Spatial Analysis: Modelling in a GIS Environment*, Geoinformation International, Cambridge, 261-277.

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

Hubert, L. and Arabie, P. (1991), 'The Assesment of Spatial Autocorrelation through Constrained Multiple Regression', *Geographical Analysis*, 13:225.

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 Institute.

Kelejian, H.H. and Prucha, I.R. (1998) 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances', *Journal of Real Estate Finance and Economics*, 17(1), 99-121.

LeSage, J.P. (1998) *Spatial Econometrics*, University of Toledo, Ohio.

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

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

Moran, P.A.P. (1948) 'The interpretation of statistical maps', *Journal of the Royal Statistical Society, Series B*, 10(2), 243-51.

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.

Openshaw, S. (1984) 'The modifiable areal unit problem', *Concepts and Techniques in Modern Geography*, 38(41), GeoBooks, Norwich, England.

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

Sawanda, M., (2005), 'Global Spatial Autocorrelation – Moran's I, Geary's C and the General Cross-Product Statistic', (mimeo), University of Ottawa, Ottawa.

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

Stimson, R.J., Stough, R.R. and Roberts, B.H. (2002) *Regional Economic Development: Analysis and Planning Strategy*, Springer-Verlag, Berlin.

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.

Tidswell, W.V. and Barker, S.M. (1971) *Quantitative Methods: An Approach to Socio-Economic Geography*, University Tutorial Press Ltd., London.

Tiefelsdorf, M. (2000) *Modelling Spatial Processes – The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I*, Springer, New York.

Tiefelsdorf, M. (2002) 'The saddlepoint approximation of Moran's I's and local Moran's I i's reference distribution and their numerical evaluation', *Journal of Geographical Analysis*, 34, 187-206.

Tobler, W. (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.

Venables, W.N., Smith, D.M. and the R Core Team (2013) An Introduction to R, downloaded from http://cran.r-project.org/manuals.html on 28 March 2013.

Credits

Centre of Full Employment and Equity

Project Leader

Professor William Mitchell, Director of Centre of Full Employment and Equity (CofFEE), University of Newcastle

Project Team

Mr Michael Flanagan, Research Officer, CofFEE, University of Newcastle

Dr Tung-Kai (Paul) Shyy, Research Fellow, e-Research Group, School of Information Technology and Electrical Engineering, The University of Queensland

Ms Melinda Hannan, Administration Officer, CofFEE, University of Newcastle