Exploratory Spatial Data Analysis

S. Dall'erba, University of Arizona, Tucson, AZ, USA

Glossary

LISA Statistics Indication of the extent to which there is a significant spatial clustering of similar values around a specific observation.

Moran's *I* Statistics that measure global spatial autocorrelation. It corresponds to the degree of linear association between the value of a variable at one location and the spatially weighted average of neighboring values.

Moran Scatterplot Graphical display of the value of a variable at one observation (on the horizontal axis) against the standardized spatial weighted average (average of the neighbors' value) on the vertical axis.

Spatial Autocorrelation Correlation between the value of an observation and the value of neighboring observations.

Spatial Heterogeneity Variability or difference between the values of observations over space. This variation may follow specific geographical patterns such as east and west, or north and south.

Spatial Weight Matrix Matrix whose cells display the degree of connection between all the observations of our sample.

Introduction

Exploratory spatial data analysis (ESDA) has greatly benefited from the integration of spatial statistical methods to the geographic information science (GIS) environment. A substantial collection of spatial data analysis software is currently available to the practitioner, which should sustain the level of success this field has experienced over the last few years. ESDA allows considering the spatial context in which most geographically referenced events take place. These events cannot be viewed as disjoint and spatially independent, otherwise conclusions about the data and, more importantly, theoretical considerations regarding the social processes that occur across space would be overlooked.

Traditionally, there are two spatial effects that ESDA is able to highlight. They are called spatial autocorrelation (or spatial dependence) and spatial heterogeneity. Spatial autocorrelation can be positive or negative. Positive spatial autocorrelation refers to the idea that nearby observations tend to display similar attributes. It refers to a geographic concentration, similarity, or clustering of a particular phenomenon. Spatial autocorrelation arises from different factors. As stated in Tobler's 'first law of geography', nearby locations are more likely to have more in common than locations that are more distant. Spatial autocorrelation may also come from the fact that the data are affected by processes spread over different locations and/or from a variety of measurement problems, such as a mismatching between the administrative boundaries used to organize the data and the actual boundaries of the processes under study. Spatial heterogeneity means that the value of observations is not homogeneous over space. This variation may follow specific geographical patterns, such as east and west, or north and south, also called spatial regimes or spatial clubs. Both spatial effects are related since the same influences captured by spatial autocorrelation can account for spatial heterogeneity.

ESDA is a step-by-step process which often starts with aspatial tools, such as choropleth maps and box plots that display the distribution of a variable and identify outliers. However, both tools are considered aspatial because they do not pay attention to the spatial linkages that influence the distribution of a variable. In order to fill this gap, we need to turn to the definition of spatial interactions via a spatial weight matrix and then to associated spatial ESDA tools. Practitioners often choose first a Moran's *I* statistics for global spatial autocorrelation. It reveals the degree of linear association between the value of a variable at one location and the spatially weighted average of neighboring values. A more detailed picture of the spatial distribution is then offered by Moran's scatterplot, local indicators of spatial association (LISA), and Getis–Ord statistics which measure local spatial autocorrelation and, eventually, reveal spatial heterogeneity. These last three tools compute, for 'each' observation, the degree of spatial clustering of similar values around that observation. While all these steps are performed and developed below on area data, it is important to note that ESDA can be applied to other types of observations such as point data, flow data, etc.

Nonspatial Tools

Choropleth Map

The first step of an ESDA is to perform a so-called choropleth map, which simply displays the distribution of the variables under study. For the sake of clarity, we use the example of the distribution of per capita gross domestic product (GDP) across the 145 regions of 12 countries of the European Union in 1989. Their distribution, relative to the European average, is shown in Figure 1.

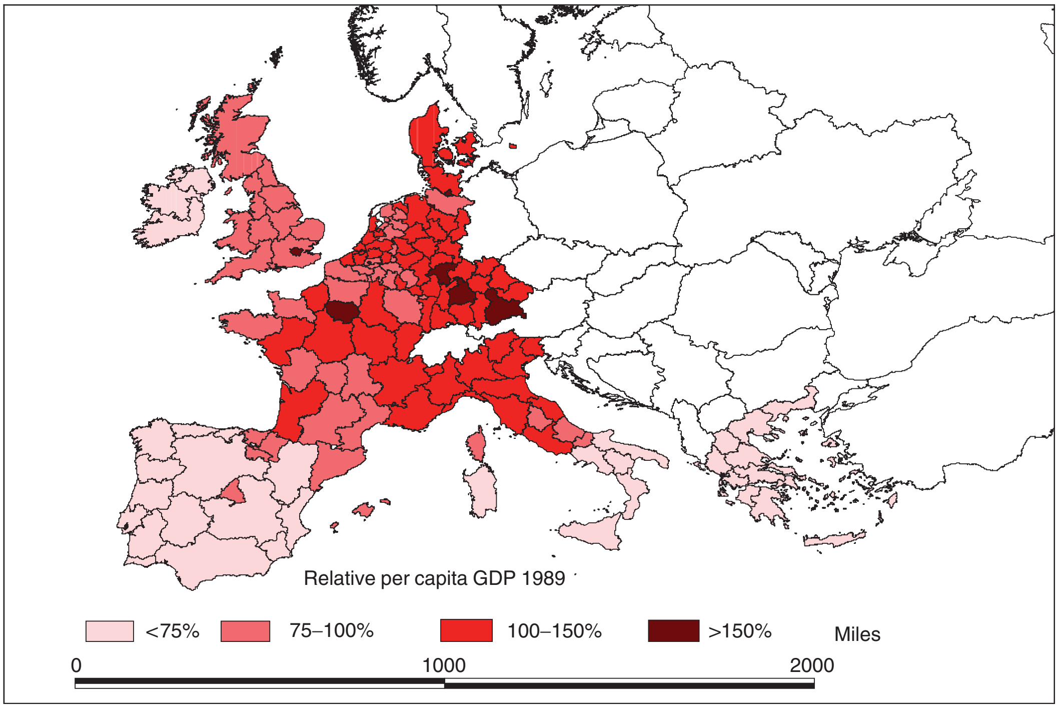

Figure 1 Choropleth map: relative distribution of per capita GDP across the regions of Europe in 1989 (in 1990 Euro prices). This figure has been realized using ArcView GIS 3.2 (Esri). Reprinted from Dall'erba, S. (2005). Distribution of regional income and regional funds in Europe 1989–1999: An exploratory spatial data analysis. *Annals of Regional Science* 39, 121–148.

Europe is marked by a clear core–periphery pattern, with the core (the darker color) composed of the richest regions, whereas the peripheral regions (Spain, Portugal, Greece, and the southern part of Italy) are the poorest ones (per capita GDP is below 75% of the average). Note that the choice of four categories is simply due to the political concerns typical of the case under study. Intervals of the categories could also have been chosen to be equal or according to the quantiles of the distribution of our variable. While a choropleth map allows visualizing the distribution of a variable, it does not tell us much about the persistence over time of this distribution nor the significant presence of the spatial effects we introduced earlier. Before we go any further, we need to say a few words on the spatial weight matrices that will be used in the spatial analysis.

Visualizing Extreme Values

Like the choropleth map above, the box plot is another basic tool that depicts the aspatial distribution of a variable. A box plot displays the median, first, and third quartile of the distribution (respectively the 50%, 25%, and 75% points in the cumulative distribution) as well as outliers. An observation is called outlier when its value is above or below a given multiple of the difference between the first and third quartile. The standard multiples are 1.5 or 3. Here is the corresponding formula with the case of 1.5 as a multiple: outlier $< Q_1 - 1.5*(Q_3 - Q_1)$ $> Q_3 + 1.5*(Q_3 - Q_1)$, where Q_1 and Q_3 are respectively first and third quartile of the distribution. Figure 2 displays an example with the same variable as above and a multiple of 1.5.

The bar in the middle of the dark area corresponds to the median while the upper (resp. lower) part of the dark area is the third (resp. first) quartile of the distribution. The thin line on the upper part of the figure is called the hinge, here corresponding to the default criteria of 1.5 times the difference between the first and third quartile. The figure indicates that five regions display some extreme values of per capita GDP. They are the outliers: Hamburg and Damrstadt (Germany), Ile-de-France (France), Brussels (Belgium), and Luxembourg. They all belong to the core regions of Europe.

Visualizing outliers can be done by linking the box plot to a box map or by representing the distribution according to the standard deviation of the variable. In the latter case, outliers are observations outside of the standard $[m - 2\sigma, m + 2\sigma]$ range, where m is the mean and σ the standard deviation.

Spatial Weight Matrix

A spatial weight matrix is a matrix whose cells display the degree of connection between all the observations of our sample. Traditionally, connections are based on pure

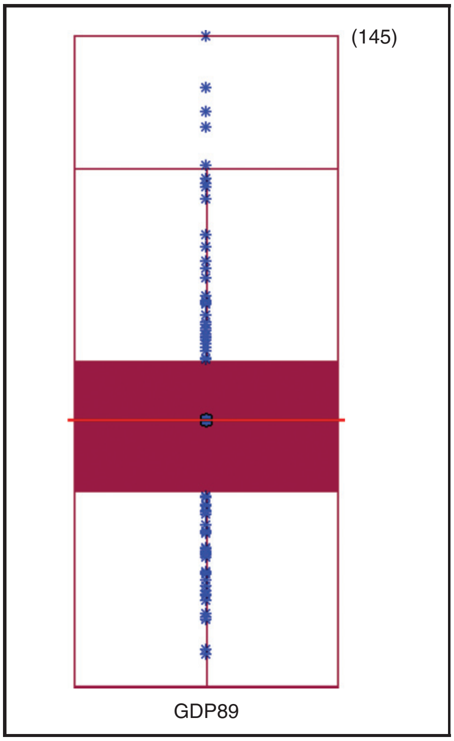

Figure 2 Box plot of the distribution of European regional per capita GDP in 1989. This figure has been created in Geoda.

geographical distance as the exogenous character of distance is unambiguous. This aspect becomes necessary when, at a later stage, spatial econometric regressions are run. (Note that in the case of per capita GDP, the true connections between regions are not based on distance but on factors such as interregional migration, trade flows, knowledge spillovers, etc. However, such data do not exist for the European regions yet.) We base our spatial weight matrices on the great circle distance for dependence to be considered in any direction. More precisely, we base our weight matrices on the $k = 10$, 15 nearest neighbors. The form of the spatial weight matrix is the following:

$$\begin{cases} w_{ij}(k) = 0 \text{ if } i = j \\ w_{ij}(k) = 1 \text{ if } d_{ij} \le D_i(k) \quad \text{for } k = 10, 15 \\ w_{ij}(k) = 0 \text{ if } d_{ij} > D_i(k) \end{cases} \qquad [1]$$

where d_{ij} is the great circle distance between centroids of regions i and j. $D_i(k)$ is the critical cutoff distance defined for each regions i, above which interactions are assumed to be negligible. In other words, $D_i(k)$ is the k^{th} order smallest distance between regions i and j such that each region i has exactly k neighbors. In the spatial weight matrix above, the weights (1 or 0) capture the spatial relationships between geographic observations. It is worth mentioning that in the European context, the minimum number of nearest neighbors that guarantees international connections between all regions is $k = 7$. Different kinds of weight matrices can be built (great circle distance with a cutoff, contiguity matrices, share of common borders, etc.), and their choice is always somewhat arbitrary. In the case of a k-nearest neighbors weight matrix, we recommend limiting the number of neighbors to 10% of the total number of observations. It is common practice to use the inverse of the distance $(1/d)$ or of the squared distance $(1/d^2)$ in order to reflect the greater degree of connection between closer observations. Another common practice consists in normalizing the outside influence upon each region by row standardizing each matrix (i.e., the row elements of the matrix are divided by the sum of the elements by row) so that each row sum is one. Row standardization means that it is relative and not absolute distance that matters.

Global Spatial Autocorrelation

Join Count Statistics

The simplest measures of spatial autocorrelation are for qualitative (binary) variables that take the value 0 or 1. These measures are called join count statistics because they measure the number of times a join corresponds to a similar or dissimilar value in the neighboring location. Because areas with observations 1 (resp. 0) are refereed as black (resp. white), there are three types of join counts:

$$\begin{aligned} \mathrm{BB} &= \tfrac{1}{2}\sum_i \sum_j w_{ij} x_i x_j \\ \mathrm{BW} &= \tfrac{1}{2}\sum_i \sum_j w_{ij} \left(x_i x_j\right)^2 \\ \mathrm{WW} &= \tfrac{1}{2}\sum_i \sum_j w_{ij} (1 - x_i)(1 - x_j) \end{aligned} \qquad [2]$$

There is positive spatial autocorrelation if the BB or WW joins dominate; there is negative spatial autocorrelation if the BW joins are more numerous. Inference is based on the approaches which are listed below for Moran's I and Geary's c.

Moran's *I* and Geary's *c*

Moran's I and Geary's c are the best-known statistics to measure global spatial autocorrelation of a quantitative variable. They correspond to the degree of linear association between the value of a variable at one location and the spatially weighted average of neighboring values. Formally, the Moran's I is given by

$$I = \frac{N}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \qquad [3]$$

where N is the number of observations, w_{ij} is the degree of connection between the spatial units i and j, and x_i is the variable of interest in region i and

$$S_0 = \sum_i \sum_j w_{ij}$$

that is, the sum of all the weights. If the spatial weight matrix is standardized, $N = S_0$, the Geary's c is given by

$$c = \frac{N-1}{2S_0} \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \qquad [4]$$

with the same notation as above.

Moran's I values larger (smaller) than the expected value $E(I) = -1/(n-1)$ indicate positive (negative) spatial autocorrelation. In our case, with $k = 10$ neighbors, $I = 0.481$ (calculations have been performed in SpaceStat) while $E(I) = -1/144 = -0.00694$. This indicates the presence of positive spatial autocorrelation, that is, in our case study, wealthy regions tend to be geographically clustered and poor regions tend to be close to each other as well. The distribution of the regional per capita GDP in Europe is therefore certainly not random. This result is confirmed by the Geary's c coefficient (0.511) which is smaller than its expected value, 1. Had the c coefficient been larger than 1, we would have concluded that a negative spatial autocorrelation exists. This would have suggested a chess board pattern of spatial dissimilarity that is rarely found in spatially referenced data. It is good practice to perform both statistics with various weight matrices and over several years (if available) to confirm your results.

Significance tests for the I and c statistics are often based on a standardized z-value whose mean and standard deviation can be calculated according to different approaches: normal distribution, the randomization assumption, or the permutation approach. For instance, the results above were calculated using a permutation approach with 10 000 permutations. It means that 10 000 resampled datasets were automatically created for which the I (or c) statistics are computed. The value obtained for the actual dataset has then been compared to the empirical distribution obtained from these resampled datasets.

While the previous spatial statistics can detect global spatial autocorrelation, they are not able to identify local patterns of spatial association, such as local spatial clusters or local spatial outliers. In addition, positive global spatial autocorrelation does not indicate whether one is dealing with a cluster of high or low values. The identification of observations that belong to a clustering of high (low) values is based on the results of a Moran scatterplot.

Local Spatial Autocorrelation

Moran Scatterplot

The idea of the Moran scatterplot is to display the value of an observation (on the horizontal axis) against the standardized spatial weighted average (average of the neighbors' value, also called spatial lag) on the vertical axis. Expressing the variables in standardized form (i.e., with mean 0 and standard deviation equal to 1) allows the assessment of both the global spatial association, since the slope of the line is the Moran's I coefficient, and local spatial association (the quadrant in the scatterplot). As shown in **Figure 3** the Moran scatterplot is therefore divided into four different quadrants corresponding to the four types of local spatial association between an observation and its neighbors:

1. Quadrant I (on the top right corner) displays the observations with a high value (above average) surrounded by observations with a high value (above average). This quadrant is usually noted as HH. In our case study, this would correspond to rich regions surrounded by rich regions.
2. Quadrant II (on the top left corner) shows observations with a low value surrounded by observations with high values. This quadrant is usually noted as LH. This would correspond to poor regions surrounded by rich regions.
3. Quadrant III (on the bottom left) displays observations with low values surrounded by observations with low values, and is noted as LL. This corresponds to poor regions surrounded by poor regions.
4. Quadrant IV (on the bottom right) shows observations with high value surrounded by observations with low values. It is noted as LL. This corresponds to rich regions surrounded by poor regions.

Figure 4 displays the Moran scatterplots of regional per capita GDP for 1989, with $k = 10$ nearest neighbors. Regions located in quadrants I and III refer to positive spatial autocorrelation, the spatial clustering of similar values, whereas quadrants II and IV represent negative spatial autocorrelation, the spatial clustering of dissimilar values. Positive global spatial autocorrelation, which was previously detected by the value of Moran's I, is reflected by (1) the positive slope coefficient of the linear regression of the spatial lag value on z, (a slope of one, passing through the origin, would indicate perfect spatial association) and (2) the fact that most of the regions are located in quadrants I and III. Those are regions respectively located in the core and the periphery of regions. Conversely, HL- and LH-type regions are often regions located at the border between peripheral and core regions. Overall, the results of the Moran scatterplot indicate that the spatial distribution of regional income is more complicated than the simple core–periphery framework previously detected in the choropleth map.

Moran scatterplot is also a useful tool for the detection of spatial heterogeneity. Indeed, from **Figure 4** it is clear that most European regions are either in the HH or in the LL quadrants which means that the distribution

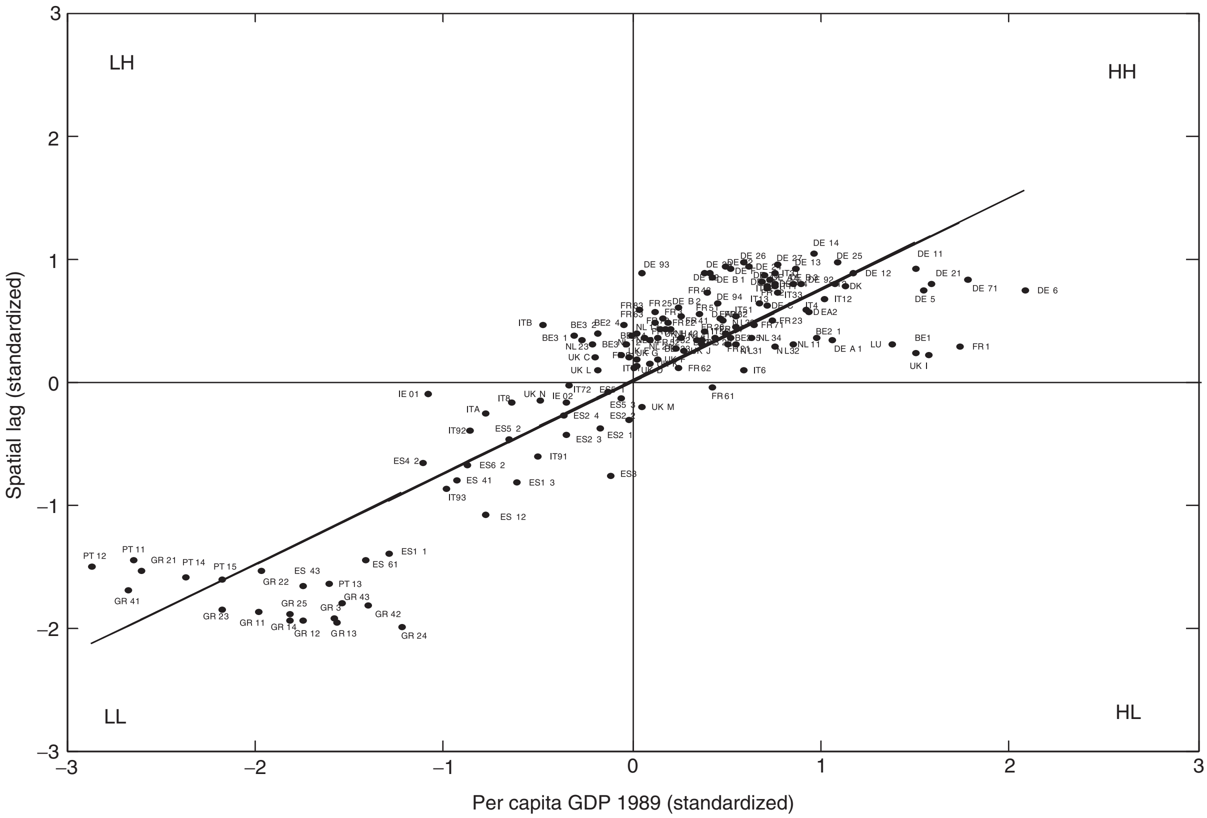

Figure 3 Moran scatterplot of regional per capita GDP in 1989. This figure has been performed in Matlab. Reprinted from Dall'erba, S. (2005). Distribution of regional income and regional funds in Europe 1989–1999: An exploratory spatial data analysis. *Annals of Regional Science* 39, 121–148.

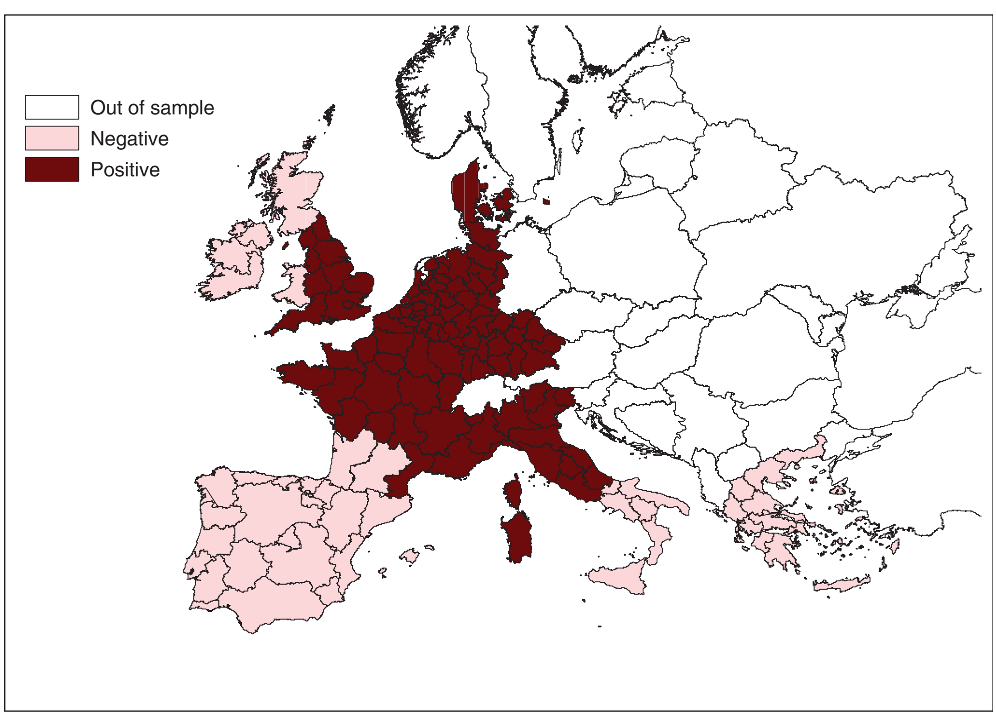

Figure 4 Getis–Ord statistics on the per capita GDP in 1989. This figure has been realized using ArcView GIS 3.2 (Esri) and the Getis–Ord statistics have been computed in SpaceStat.

of the variable studied (per capita GDP in 1989) is not homogeneous over space. Further tools are used below to detect whether these two spatial clubs are consistent.

Finally, note that Figure 4 corresponds to a so-called univariate Moran scatterplot because we use the same variable on the x- and the y-axis. A multivariate Moran scatterplot can also be done. It corresponds to a case where the y-axis is calculated on another variable (a different year, a different variable or both) than the variable used in the x-axis. The case where the two axes are based on different years is referred to as spatial–temporal dependence.

Getis–Ord statistics

The G statistics developed by Getis and Ord provide an alternate measure of global spatial association (G) and observation specific (G_i and G^*) measures of local spatial association. These last statistics detect the presence of local spatial autocorrelation: a positive value of this statistic for observation i indicates a spatial cluster of observations with high value of the variable of interest, whereas a negative value of the G_i or G_i^* statistics indicates a spatial cluster of observations with low value around observation i.

Formally, the G_i statistics can be written as follows:

$$G_i = \frac{\sum_{j \neq i} w_{ij}(d) x_j}{\sum_{j \neq i} x_j} \qquad [5]$$

where $w_{ij}(d)$ are the elements of the matrix whose distance cutoff is d. The matrix used for the calculation of this statistic does not necessarily have to be symmetric and binary. The difference between G_i and G_i^* is that the latter statistics provide a measure of spatial clustering that includes the observation under consideration, while G_i does not. In other words, $j = i$ is included in the sum (in the denominator) in G_i^* statistics.

These statistics imply a two-way split of the sample which can be useful for the determination of spatial regimes or spatial clubs. Applied to our case study, the two clubs (rich surrounded by rich and poor surrounded by poor) are represented in Figure 4. Spatial heterogeneity, previously detected in the Moran scatterplot, seems to be consistent with the Getis–Ord statistics as well.

Note, however, that we did not pay attention to the significance level of the statistics for Figure 4. If we had included the significance level, 68 of the 145 statistics would have been nonsignificant.

Local Indicator of Spatial Association

The Moran scatterplot displayed above does not give any indication on the 'significance' of spatial clustering. We therefore need to calculate LISA statistics for each observation to obtain an indication of the extent to which there is a significant spatial clustering of similar values around a specific observation. LISA statistics, developed by Anselin, are used for the detection of significant local spatial clusters (also called 'hot spots') as well as for the diagnostics of local instability, significant outliers, and spatial regimes. The local Moran's statistics for each region i can be formalized in the following way:

$$I_i = \left(\frac{x_i}{m_0}\right) \sum_j w_{ij} x_j \quad \text{with} \quad m_0 = \sum_i \frac{{x_i}^2}{n} \qquad [6]$$

with x_i (x_j) is the observation in region i (j) (each measured as a deviation from the mean value) and the matrix w is row standardized.

The sum of LISAs for all observations is proportional to a global indicator of spatial association. Since we use a row-standardized matrix, the average of local Moran statistics is also equal to the global Moran's I statistics.

The results from the application of LISA with $k = 10$ nearest neighbors are displayed in Figure 5. The significance level is based on a conditional permutation approach with 10 000 random permutations of the neighboring regions for each observation. The pseudo-significance level is 5%.

Regions revealing a significant spatial association of per capita GDP (HH or LL) are mostly those previously detected as core (HH cluster) and peripheral (LL cluster) regions. Indeed, in the figure, 83% of the local statistics that are significant are either HH- or LL-type, reflecting the global trend of positive spatial association and spatial heterogeneity since the clusters represent different values of per capita GDP. The figure also indicates that some regions display signs of significant negative spatial autocorrelation (LH- or HL-type). As explained above, those are regions with a level of per capita GDP significantly different than the one of their neighbors, such as Madrid in Spain for instance (HL-type).

Finally, it is important to pay attention to the problem of multiple statistical comparisons that the LISA statistics above may present. This problem comes from the fact that the neighborhood sets of two regions may contain common regions. In order to correct for the error on the appropriate significance level due to multiple comparisons, we need to compute a Bonferroni pseudo-significance level (= a 5% pseudo-significance level in our case). The Bonferroni procedure consists in adjusting downward the alpha level of 'overall' significance (5%) by the number of comparisons ($k = 10$) in order to find the 'individual' significance level (= 0.5% or 5% Bonferroni pseudo-significance level).

Software

We seek to provide the reader with a brief review of the software currently available to perform ESDA. We realize this section is doomed to be quickly obsolete given the

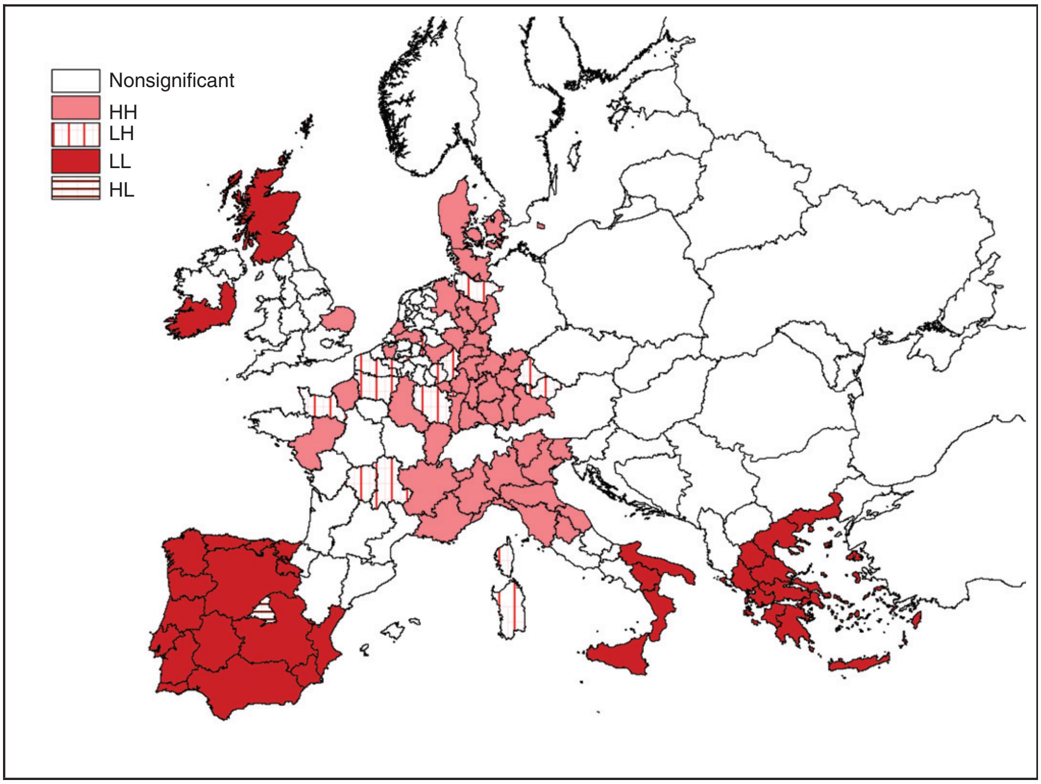

Figure 5 LISA statistics on the per capita GDP in 1989. This figure has been realized using ArcView GIS 3.2 (Esri) and the LISA statistics have been computed in SpaceStat.

speed at which this field evolves. Some of the oldest software includes the 'Spider/Regard' toolboxes which allow 'geographic brushing'; SAGE which adds ESDA capability to the standard ArcInfo GIS environment; and the DynESDA extension for ArcView. ArcGIS Geostatistical Analyst allows exploring the spatial distribution of data, but not necessarily with all the tools described above. Currently, the most complete software is Geoda and SpaceStat. Despite its old-fashioned interface and performance constraints for large datasets, the latter provides very useful capabilities. It requires an ArcView extension for visualization. Geoda is a more recent and user-friendly software that allows both ESDA and some (limited) spatial modeling capabilities. CrimeStat is a Windows-based software that interfaces with most desktop GIS programs. Primarily devoted to fighting crime, it has been very popular with police departments around the country. ArcGIS Desktop 9.0 Spatial Statistics Geoprocessing Toolbox Patch contains performance improvements for cluster and outlier analysis, hot spot analysis, and spatial autocorrelation. Other toolboxes, but which require some programming skills, include the 'tmap' function developed in Stata; the 'spdep' and 'DCluster' functions in R; PySpace and STARS (space–time analysis of regional systems) which are written in Python and, finally, the Matlab Arc_Mat and econometric routines.

Looking Ahead

We live in an exciting period when spatial statistics have tremendously benefited from technical and theoretical advances in GI science. It is very likely that these prowesses will continue in the decades to come. However, with the use of spatial statistics becoming more popular across different fields, there is a clear need to tighten the interaction between the methodologies, the computational capabilities, and the specificities inherent to each discipline. For instance, many disciplines deal with linkages that are not based on pure geographical distance but, instead, rely on more sophisticated concepts of space (such as economic, social, or political distance). It is not an easy task though since it would require a rethinking of the current statistical methods. Another aspect to be developed focuses on the exploration of patterns of flow data, since reducing transportation costs and more generally technological improvements have tremendously increased mobility. In addition, previous ESDA tools do not necessarily allow the user to capture the full extent of a variable distribution as the value in one observation is often compared to the average of the sample or neighboring observations. With the size of the samples becoming bigger, the chances for a distribution to be multimodal (i.e., displaying signs of multiple clusters) increase. In this case, we recommend complementing an

ESDA with the increasingly popular tools such as density functions, Markov chains, stochastic kernels, and their spatial version. Finally, the outcome of an ESDA will never be more than a detailed description of the distribution of a variable over space. Its capacity to measure the relationships between several variables is very limited. The natural subsequent step is therefore a thorough estimation of the workings of the system under study. This could be done, for example, using spatial econometric techniques where the interaction between variables and spatial effects are formally tested and measured.

See also: Spatial Autocorrelation; Spatial Clustering, Detection and Analysis of; Statistics, Descriptive; Statistics, Spatial.

Further Reading

Anselin, L. (1995). Local indicator of spatial association – LISA. *Geographical Analysis* 27, 93–115.

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. In Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. (eds.) *Geographic Information System: Principles, Techniques, Management and Applications*, pp 251–264. New York: John Wiley.

Anselin, L. (2005). *Exploring Spatial Data with Geoda: A Workbook*. Urbana, IL: University of Illinois at Urbana-Champaign.

Anselin, L., Florax, R. G. J. M. and Rey, S. (2004). *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Heidelberg: Springer-Verlag.

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

Dall'erba, S. (2005). Distribution of regional income and regional funds in Europe 1989–1999: An exploratory spatial data analysis. *Annals of Regional Science* 39, 121–148.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2000). *Quantitative Geography, Perspectives on Spatial Data Analysis*. Thousand Oaks, CA: Sage Publications.

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24, 189–206.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.

Le Gallo, J. (2004). Space–time analysis of GDP disparities among European regions: A Markov chains approach. *International Regional Science Review* 27, 138–163.

Le Gallo, J. and Dall'erba, S. (2006). Evaluating the temporal and spatial heterogeneity of the European convergence process, 1980–1999. *Journal of Regional Science* 46, 269–288.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27, 286–305.

Rey, S. J. and Montouri, B. D. (1999). US regional income convergence: A spatial econometric perspective. *Regional Studies* 33, 143–156.

Upton, G. and Fingleton, B. (1985). *Spatial Data Analysis by Example*. New York: Wiley.

Relevant Websites

http://www.csiss.org
Center for Spatially Integrated Social Science.

http://www.esri.com
ESRI: GIS and mapping software, ArcGIS Geostatistical Analyst.

http://www.geoda.uiuc.edu
Geoda.

http://www.icpsr.umich.edu
Inter-University Consortium for Political and Social Research: CrimeStat.

http://www.spatial-econometrics.com
Matlab spatial toolboxes.

http://www.spatstat.org
Open source software for spatial statistics.

http://regionalanalysislab.org
Regional Analysis Laboratory: STARS.

http://www.terraseer.com
Terraseer, Inc.: SpaceStat.

http://www.shef.ac.uk
The University of Sheffield: SAGE.