Mathematics and Statistics 8(2): 82-86, 2020
DOI: 10.13189/ms.2020.080202 http://www.hrpub.org

From Exploratory Data Analysis to Exploratory Spatial Data Analysis

Patricia Abelairas-Etxebarria*, Inma Astorkiza

Department of Applied Economics V, University of the Basque Country, Spain

Received November 18, 2019; Revised January 10, 2020; Accepted January 16, 2020

Abstract The Exploratory Data Analysis raised by Tuckey [19] has been used in multiple research in many areas but, especially, in the area of the social sciences. This technique searches behavioral patterns of the variables of the study, establishing a hypothesis with the least possible structure. However, in recent times, the inclusion of the spatial perspective in this type of analysis has been revealed as essential because, in many analyses, the observations are spatially autocorrelated and/or they present spatial heterogeneity. The presence of these spatial effects makes necessary to include spatial statistics and spatial tools in the Exploratory Data Analysis. Exploratory Spatial Data Analysis includes a set of techniques that describe and visualize those spatial effects: spatial dependence and spatial heterogeneity. It describes and visualizes spatial distributions, identifies outliers, finds distribution patterns, clusters and hot spots and suggests spatial regimes or other forms of spatial heterogeneity and, it is being increasingly used. With the objective of reviewing the last applications of this technique, this paper, firstly, shows the tools used in Exploratory Spatial Data Analysis and, secondly, reviews the latest Exploratory Spatial Data Analysis applications focused on different areas in the social sciences particularly. As conclusion, it should be noted the growing interest in the use of this spatial technique to analyze different aspects of the social sciences including the spatial dimension.

Keywords Exploratory Data Analysis, Exploratory Spatial Data Analysis, Last Applications

1. Introduction

Exploratory Data Analysis (EDA) has been a widely used technique to analyze different aspects, in particular, of the social sciences. It tries to find behavioral patterns of the variables of the study, establishing a hypothesis with the least possible structure (Tukey [19]). On the one hand, the box plots give information on the scope or range of the distribution of variables, quartiles, the central tendency of the distribution (mean and median), dispersion of the distribution (interquartile range and standard deviation) and atypical values or outliers. On the other hand, the scatter plots and the correlation matrix report the association between the variables. However, the EDA ignores the spatial dimension and the location of observations and, therefore, the *spatial effects* (spatial dependence and spatial heterogeneity).

The Exploratory Spatial Data Analysis includes those spatial effects and, in recent times, it is being increasingly used. This paper reviews the last applications of this technique showing, first, the tools Exploratory Spatial Data Analysis uses and, second, reviewing the latest Exploratory Spatial Data Analysis applications focused on different areas in the social sciences.

2. Exploratory Spatial Data Analysis

Exploratory Spatial Data Analysis (ESDA) is a sub-discipline of general EDA and is concerned with the presence of *spatial effects*. ESDA can be defined as a set of techniques that describe and visualize spatial distributions, identify *outliers*, find distribution patterns, *clusters* and *hot spots* and suggest spatial regimes or other forms of spatial heterogeneity. Exploratory Spatial Data Analysis is descriptive and is the step prior to confirmatory analysis in which the econometric model is estimated.

Exploratory Spatial Data Analysis techniques include two spatial effects: spatial dependence and spatial heterogeneity. Spatial dependence or spatial autocorrelation arises from the existence of a relationship between what happens in a particular place and what happens in another point considered a neighbor (Cliff [6]; Paelinck [14]; Anselin [4]). It is based largely on the "first law of geography" formulated by Tobler [17] in 1979: *"Everything is related to everything else, but near things are more related to each other than*

distant things". Thus, the value taken by a variable at a point is given, in addition to internal constraints, by the value taken by that same variable at a neighboring point. This means the sample observations are not independent and that this interrelation has to be treated with spatial techniques. Following this idea, Andrienko [2] shows exploratory data analysis and, in particular, exploratory analysis of spatial and temporal data.

The spatial dependence can be positive or negative. In the first case the values of the variable in one place and at one neighboring point are similar, that is to say, the existence of a phenomenon in a particular place means that same phenomenon is extended to the rest of the places that surround it. In the case of negative spatial autocorrelation these values are dissimilar, or said another way, the existence of a phenomenon in one place prevents or hinders the emergence of this phenomenon in a neighboring point (Moreno [13]).

The existence of the first of these spatial effects, spatial autocorrelation, could be likened to what happens in the temporal context. When autocorrelation appears in this context, it does so unidirectionally, meaning the observations have some kind of temporal relationship between them, but in a single direction. In the spatial case, however, dependence between observations, or autocorrelation, is more complex since it is multi-directional, i.e., each of the observations may be related to several neighboring observations and also be related to each of them in a different way.

The second of the spatial effects, spatial heterogeneity, arises when working with heterogeneously related spatial units. Spatial heterogeneity can be caused by two factors: structural instability and heteroskedasticity. When it is due to the first factor, the variable is not stable in space and its behavior varies depending on its location. In models in which this effect appears, the functional form and the parameters may vary by geographic location. In regard to the second factor, the existence of heteroskedasticity, we can say it comes from the omission of variables or from specification errors that make the regression term present a non-constant variance.

Unlike what occurs with spatial autocorrelation, the existence of spatial heterogeneity could be treated with procedures of general econometrics; but, at the moment in which spatial dependence appears in the model, these techniques are no longer adequate and it becomes necessary to adapt the statistics in order to introduce the detected spatial autocorrelation.

As explained above, the relationships that appear in the spatial context are more complex than those in the time context as they are multi-directional and every observation may be related to several nearby observations. Spatial econometrics has solved this problem by creating the so-called Weight Matrix (W), which is a square matrix of identical size to the number of observations. The elements forming the matrix W capture the way in which each observation is related to the rest, that is to say, the element w_{ij} explains the intensity of the relationship between observation i and observation j (The main diagonal is by definition assumed equal to 0):

$$W = \begin{bmatrix} 0 & w_{12} & .. & w_{1n} \\ w_{21} & 0 & .. & w_{2n} \\ : & : & .. & : \\ w_{n1} & w_{n2} & .. & 0 \end{bmatrix}$$

Although there is no consensus on a definition of the spatial weights between observations, the most frequent approach is the use of physical contiguity of the first order, where w_{ij} is equal to 1 if the observations are neighbors and 0 if they are not. In relation to physical adjacency, and assuming the existence of a regular grid, this work follows the contiguity criterion based on the chess movement of the '*queen*' (queen-based contiguity) to identify neighboring municipalities. So, neighbors of the municipality *i*, will be those municipalities that share any of the 8 sides or vertices with *i*. There are other matrices that also follow the physical proximity criterion, but they are based on the distance between observations. In these, spatial weights have to do with the distance that separates the observations. Anselin [3] proposes the use of an inverse matrix of squared distances, so that the intensity of the interdependence between observations decreases with the distance that separates them. Other ways to specify the weight matrices have been based not so much on physical proximity criteria but, for example, on economic distances or the level of trade between regions.

It should be noted that it is common to standardize the weight matrix by rows, by dividing each element w_{ij} by the total sum of the row in which it is located, in such a way that the sum of the row is equal to the unit. This facilitates equal weighting of the total influence each observation receives from its neighbors, regardless of the total number of neighbors it has.

ESDA can be divided into three parts: the visualization of the spatial distribution, the study of global spatial autocorrelation or dependence and the study of spatial dependence at the local level.

2.1. Spatial Visualization Techniques

One of the ways to perform Exploratory Spatial Data Analysis begins with the visualization of the spatial distribution of the variable under study on a map. There we begin to observe whether or not the variable presents some degree of spatial association. Next, we try to identify spatial outliers using the Box Map, which consists of a geographical representation of the Box Plot.

2.2. Statistics of Global Spatial Autocorrelation

For the analysis of the global autocorrelation, a statistic is

calculated that looks for the existence of a relationship between the similarity in the localization and the similarity between the values of the variable: Moran's I statistic. This statistic summarizes the general profile of spatial dependence existent in a variable by checking the presence of general spatial structures in the distribution of the variable throughout the area under study.

Moran's I statistic (Moran [12]) is expressed as:

$$I = \frac{N}{S_0} \frac{\sum_{ij}^{N} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \quad (1)$$

where x_i takes the value of the variable in place i,$\bar{x}$ is the sample mean, w_{ij} are the weights of the matrix W, S_0 is the sum of all w_{ij} and N is the sample size. According to Cliff [6], when the sample size is large enough, Moran's standardized I follows a normal asymptotic distribution $N(0.1)$:

$$Z(I) = \frac{I - E(I)}{\left[V(I)\right]^{1/2}} \approx N(0,1) \quad (2)$$

where E(I) and V(I) are the expectation and the variance, respectively.

An insignificant value of Z(I) accepts the null hypothesis of no spatial autocorrelation and a significant value indicates the existence of spatial autocorrelation. If the value is positive, it indicates positive spatial autocorrelation and if the value is negative, spatial autocorrelation is negative.

In this analysis of global spatial autocorrelation Moran's Scatterplot is also used. Moran's Scatterplot represents, on the x-axis, the values of the observed variable, and on the ordinate axis, the "spatial lag" of that same variable. If data the points are concentrated in the first and third quadrant there is spatial autocorrelation (similar values) and if they are concentrated in the second and fourth quadrant there is negative autocorrelation (different values).

2.3. Statistics of Local Spatial Autocorrelation

The previous statistical analyzes all the sample observations as a whole, but it is not sensitive to situations of instability in the spatial distribution. For the analysis of spatial autocorrelation at a local level, we use other statistics called Local Indicators of Spatial Association (LISA): Local Moran's I and Local Getis-Ord G. These local autocorrelation tests are useful for detecting clusters and for finding regions that contribute in a special way to the global spatial autocorrelation. Following Moreno [13], and Chasco [5], we have:

Moran's local I statistic (Anselin [3]) which is expressed as follows:

$$I_i = \frac{z_i}{\sum_j z_j^2 / N} \sum_{j \epsilon J_i} w_{ij} z_j \quad (3)$$

where z_i is the value of the normalized variable at point i and J_i is the set of neighboring observations at i. It is assumed that the standardized I_i is asymptotically distributed as a normal $N(0.1)$ under a null hypothesis of no spatial autocorrelation. A significant positive value indicates the presence of a *cluster* of similar values and a negative value the presence of a cluster of dissimilar values.

As in the global analysis, there are graphs associated with these statistics, such as significance maps, that will help with the interpretation of the obtained results.

3. Exploratory Spatial Data Analysis Last Applications

In recent years, a considerable number of authors from different disciplines, particularly in social sciences, have applied this methodological tool to describe the variables of interest. They all have the common characteristic of studying a geographically localized phenomena. Below, several papers are presented, all of them have applied the ESDA as a tool to analyze their different subjects of study.

Ye [11] applies the techniques that encompass ESDA to study the dynamics of spatial patterns and structural indicators of homicide rates in Chicago, including the temporal dimension. The ESDA reveals spatially complex phenomena that it would otherwise not be captured.

Li [10] employs GIS and ESDA techniques to study the clustering behavior of interactions and daily activities of children. The use of these techniques allows them the simultaneous analysis of the social and spatial dimension. Guo [7] seek to examine the dynamic spatial patterns between a term coined “coupling”, the interaction between the urban man-made environment and the ecological environment, and the policies applied in the Huai River basin. To do this, among other techniques, they apply the ESDA that allows them to explore changes at a spatial level between urbanization and the environment in that enclave. Kaygalak [8] tries to identify clusters and agglomerations of industries throughout Turkey at the provincial level through the use of the ESDA. In addition, these techniques allow them to identify the existing spatial dependence between different provinces.

Agovino [1] applies cluster analysis and ESDA to analyze the process of waste management in 103 Italian provinces with the aim of improving such management. In this way, they describe the spatial distribution of the variable and identify the processes of spatial autocorrelation, spatial clusters and existing atypical observations.

Gutierrez [8] discuss the worrisome evolution of evictions in Spain in the current crisis from a spatial perspective. They try to find spatial clusters of that variable and relationships with other variables. To do this, they apply ESDA techniques to the purchases of second homes acquired by banks as a proxy variable for the number of

evictions, since the data of this variable at the local level are not available.

Tang [16] studies the development of China's provincial industrial circular economy including the dimensions time and space. Appling ESDA, they conclude that the industrial circular economy of China presents spatial correlation and regional differences and that the spatial agglomeration pattern of this industrial circular economy is roughly the same that the economic development.

Tu [18] studies the spatial patterns of Low Birth Weight prevalence and the existence of spatial clusters in the State of Georgia of USA at county and census tract levels. Using ESDA, conclude that there is geographical variation in Low Birth Weight prevalence in Georgia in 2000. Also, the study concludes that a significant disparity between white and black racial subgroups exists as the literature reviewed has showed.

Recently, Salvati [15] has used this ESDA technique to compare the spatial pattern of prices in the economic expansion and in recession in Rome, analyzing demography, land-use and territorial factors. Their results show that crisis influenced considerably the spatial structure of housing prices in Rome.

4. Conclusions

The use of the Exploratory Data Analysis, that seeks patterns of behavior of the study variables, has been very useful for analyzing multiple subjects. Nevertheless, in recent times, the spatial dimension of the variables of study has become more and more important. Because of that, the application of the Exploratory Spatial Data Analysis has been expanding in the study of various areas to include in the research the spatial effects, the spatial autocorrelation and the spatial heterogeneity. This type of analysis gives the study a dimension that was previously not taken into account.

As it can be seen in the review of the literature of recent years, the Exploratory Spatial Data Analysis is being used in lots of research in the context of the social sciences, specially. The inclusion of the spatial dimension in the Exploratory Data Analysis gives to the analysis a perspective that, in some cases, modifies the results, because not to include the spatial effects in the analysis could give biased or incomplete results

Fortunately, in recent years, spatially geo-referenced data is widespread and increasingly accessible, which helps the development of such analyzes.

Acknowledgements

We are very grateful to experts for their appropriate and constructive suggestions to improve this template.

REFERENCES

[1] M. Agovino, M. Ferrara, A. Garofalo. An exploratory analysis on waste management in Italy: A focus on waste disposed in landfill. Land Use Policy, Vol. 57, 669–681. https://doi.org/10.1016/j.landusepol.2016.06.027

[2] N. Andrienko, G. Andrienko. Exploratory Analysis of Spatial and Temporal Data. A Systematic Approach, Springer. 2005.

[3] L. Anselin. Estimation Methods for Spatial Autoregressive Structures. Ithaca, NY: Cornell University, Regional Science Dissertation and Monograph Series. 1980.

[4] L. Anselin. Spatial Econometrics: Methods and Models, Dordrecht: Kluwer Academic. 1988.

[5] C. Chasco Irigoyen. Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales, Madrid: Consejería de Economía e Innovación Tecnológica, Comunidad de Madrid. 2003.

[6] A. Cliff, J. Ord. Spatial autocorrelation, London: Pion. 1973.

[7] Y. Guo, H. Wang, P. Nijkamp, J. Xu. Space-time indicators in interdependent urban-environmental systems: A study on the Huai River Basin in China. Habitat International Vol. 45, 135-146. https://doi.org/10.1016/j.habitatint.2014.06.030

[8] A. Gutiérrez, X. Delclòs. The uneven distribution of evictions as new evidence of urban inequality: A spatial analysis approach in two Catalan cities. Cities Vol. 56, 101–108. http://dx.doi.org/10.1016/j.cities.2016.04.007

[9] I. Kaygalak, N. Reid. The geographical evolution of manufacturing and industrial policies in Turkey. Applied Geography Vol. 70, 37-48. https://doi.org/10.1016/j.apgeog.2016.01.001

[10] X. Li, W.A. Griffin. Using ESDA with social weights to analyze spatial and social patterns of preschool children's behavior. Applied Geography Vol. 43, 67-80. https://doi.org/10.1016/j.apgeog.2013.06.003

[11] X. Ye, L. Wu. Analyzing the dynamics of homicide patterns in Chicago: ESDA and spatial panel approaches. Applied Geography Vol. 31, 800-807. https://doi.org/10.1016/j.apgeog.2010.08.006

[12] P. Moran. The interpretation of statistical maps. Journal of the Royal Statistical Society, Vol. 10B, 243-251

[13] R. Moreno, E. Vayá. Técnicas econométricas para el tratamiento de datos espaciales: La econometría espacial. Barcelona: Edicions Universitat de Barcelona. 2000.

[14] J.H.P. Paelinck, L.H. Klaassen. Spatial Econometrics, Farnborough: Saxon House. 1979.

[15] L. Salvati, M.T. Ciommi, P. Serra, F.M. Chelli. Exploring the spatial structure of housing prices under economic expansion and stagnation: The role of socio-demographic factors in metropolitan Rome, Italy. Land Use Policy, Vol. 81, 143-152.

[16] J.Tang, M.Tong, Y. Sun, J. Du, N. Liu A spatio-temporal perspective of China's industrial circular economy development. Science of the total environment. In Press.

[17] W. Tobler. Cellular geography, Philosophy in Geography, 379-386. 1979.

[18] W. Tu, S. Tedders, J. Tian. An exploratory spatial data analysis of low birth weight prevalence in Georgia. Applied Geography, Vol. 32, 195-207.

[19] J.W. Tukey. Exploratory data analysis, Addison-Wesley, Reading, Mass. 1977.