Review

Basic Spatial Exploratory Spatial Data Analysis in GeoDa Software

Coro Chasco *

Posted Date: 27 March 2025

doi: 10.20944/preprints202503.2044.v1

Keywords: ESDA; thematic maps; continuous variables; discrete variables; econometric models

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2025 doi:10.20944/preprints202503.2044.v1

Review

Basic Spatial Exploratory Spatial Data Analysis in GeoDa Software

Coro Chasco

Department of Applied Economics, Universidad Autónoma de Madrid, C/ Francisco Tomás y Valiente, 5, 28049 Madrid, Spain; coro.chasco@uam.es

Abstract: This paper provides an in-depth examination of Exploratory Spatial Data Analysis (ESDA), an extension of the exploratory data analysis concepts introduced by John Tukey in the 1970s. ESDA is designed to uncover hidden patterns and relationships in spatial data that are not apparent through traditional statistical methods. It incorporates techniques such as spatial visualization, dynamic linking, and spatial autocorrelation analysis to investigate spatial distributions and relationships across different datasets. By employing maps, histograms, scatter plots, and other graphical representations, ESDA enables the interactive exploration of spatial dependencies and heterogeneities. The integration of geographic information systems (GIS) tools like Luc Anselin's GeoDa software enhances the ability of ESDA to handle complex spatial data structures and to visualize geographical data effectively. This paper discusses various ESDA methodologies tailored to analyze continuous, discrete, and spatial-temporal data, emphasizing their critical role in the development of econometric models. Additionally, it underscores the importance of spatial context in statistical analysis, advocating for the consideration of both physical and socio-cultural dimensions of space in understanding patterns and processes that influence human behavior and environmental dynamics. The detailed review also covers the use of ESDA in multiple disciplines, illustrating its versatility and effectiveness in providing insightful analyses that inform better decision-making in fields ranging from urban planning to environmental science.

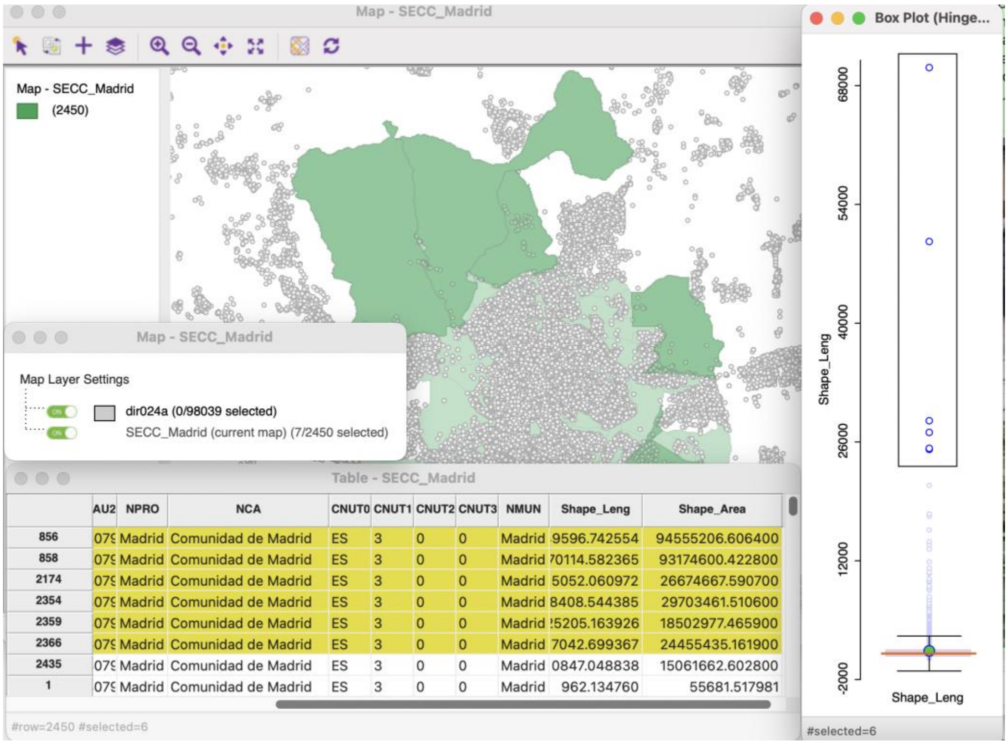
Keywords: ESDA; thematic maps; continuous variables; discrete variables; econometric models

1. What Is Exploratory Spatial Data Analysis (ESDA)

Exploratory spatial data analysis (ESDA) was developed from a-spatial exploratory data analysis (EDA). EDA is generally viewed to have originated in the 1960s as a reaction by computationally oriented statisticians to the primary focus on mathematics and modeling in their discipline. It was first comprehensively outlined in the classic book by John Tukey (Tukey 1977). His presentation stressed the value of investigating the raw data by means of a range of graphic devices, several of which were developed by Tukey himself, like the box plot.

ESDA is widely used in spatial statistics, spatial econometrics and geostatistics (see Bivand 2010). It is a set of techniques and methods used to analyze and visualize spatial data in order to uncover patterns, trends, and relationships that might not be immediately apparent. Common techniques used in ESDA include spatial visualization (e.g., thematic maps) and spatial autocorrelation analysis. ESDA allows the user to manipulate various 'views' of the data: maps, histograms, box plots, scatter plots, etc. The map is just another view, integrated into the overall scheme (Figure 1). It is a dynamic linking and brushing of various views (Anselin 2023): map, box plot, frequency histogram, data table, etc.

Why incorporate maps and, in general, tools from geography into exploratory analysis? The answer could be found in the French philosopher and sociologist Henri Lefebvre, who affirmed in his work "*The Production of Space*": "*space matters*" (Lefebvre 1992). The idea behind this statement emphasizes the importance of understanding space not just as a physical entity but also as a social and cultural construct that plays a significant role in shaping human experiences and interactions. In fact, social phenomena are often not independent of the geographical space in which they occur. For example, in cities one can find certain neighborhoods with high-income households surrounded by neighborhoods with high-income households, and vice versa: spatial concentrations of neighborhoods with lower-income families. These conditions are framed by Tobler's First Law of Geography: "*Everything is related to everything else, but near things are more related than distant things*" (Tobler 1970, p 236).



Map Layer Settings

dir024a (0/98039 selected)

SECC_Madrid (current map) (7/2450 selected)

	AU2	NPRO	NCA	CNUT0	CNUT1	CNUT2	CNUT3	NMUN	Shape_Leng	Shape_Area
856	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	9596.742554	94555206.606400
858	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	70114.582365	93174600.422800
2174	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	5052.060972	26674667.590700
2354	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	8408.544385	29703461.510600
2359	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	25205.163926	18502977.465900
2366	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	7042.699367	24455435.161900
2435	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	0847.048838	15061662.602800
1	079	Madrid	Comunidad de Madrid	ES	3	0	0	Madrid	962.134760	55681.517981

#row=2450 #selected=6

Figure 1. Example of linking and brushing different views in ESDA.

Hence, space can play an important role in determining the processes to be modelled. For this cause, one of the reasons why ESDA is relevant is the need to perform an exploratory analysis of the variables included in a spatial econometric model prior to its estimation, especially the dependent variable. Table 1 presents some basic ESDA methods that are suitable for analyzing various types of dependent variables in possible econometric models.

In fact, ESDA has developed different methods depending on the nature of the variable(s) involved: if only one variable is involved, we speak of univariate ESDA, and if two or more variables are involved, we speak of bivariate or multivariate ESDA, respectively. And within each ESDA categories, there are different methods depending on the type of the analyzed variable.

Table 1. ESDA tools suitable for dependent variable econometric models.

Dependent variable	ESDA tool	Econometric model
Continuous	Histogram Box plot Quantile map Natural break map Box map	Spatial linear regression models Spatial Expansion models Geographically Weighted Regres. Trend surface models Spatial ridge and lasso models Spatial Partial Least Squares

Discrete	Bar chart	Spatial count data models
	Unique values	Binary Spatial models
		Spatial logit models
		Spatial probit models
Rates	Raw rate map	Spatial beta regression models
	Excess risk map	Spatial fractional response mod.
	Empirical Bayes smoothed rate map	Spatial logit, probit, tobit mod.
Space-Time	Box plot over time	Spatiotemporal models
	Scatter plot with time lagged vars.	Spatial panel data models
	Thematic map over time	Difference-in-difference models
	Difference in means	
	Difference-in-difference	

In this paper, an introduction to ESDA, its origins and main applications is provided. A first chapter is dedicated to describing some basic ESDA techniques for continuous and discrete variables, focusing on methods like histograms, bar charts, and thematic maps to visualize univariate data distributions. A second chapter presents some basic multivariate ESDA that explores relationships, structures and spatial patterns among multiple variables, employing, among others, scatter plots for continuous data to assess correlations and linear –and nonlinear– relationships, and using tools like the parallel coordinate plot and, for discrete data, the co-location maps, highlighting how these techniques help identify patterns, clusters, and outliers in complex data sets. The Conclusion section, References, and Appendix A close.

2. Basic Univariate ESDA

2.1. For Continuous Variables

In the current section, I focus on techniques to describe the distribution of **one variable at a time** (univariate). This variable, in turn, can be continuous or discrete. A continuous variable is a variable such that there are possible values between any two values. Continuous variables represent measurable amounts (e.g. monetary quantities, water volume or weight).

The ESDA of one continuous variable makes sense in itself, but also as part of the identification phase of a **continuous dependent variable** econometric model. This is the case of the spatial linear regression, trend surface, spatial expansion, spatial ridge, spatial lasso, spatial partial least squares, Geographically Weighted Regression, etc.

Generally speaking, exploratory analysis for continuous variables is a critical initial step in understanding the distributions, trends, and patterns within a dataset. Univariate ESDA seeks representation of **central tendency** and identification of **outliers** of spatial distributions (Anselin 2023, Ch. 4; Chasco and Vallone, 2023). Next, we present three basic representations: histogram, boxplot and thematic maps (quantile map, natural breaks map and box map).

2.1.1. Histogram

It is the representation of the general distribution of the values of a variable. It is also a **discrete representation of the density function of a continuous variable**. The range[1] of the variable is

[1] In descriptive statistics, the **range of a data set** (r) is determined by subtracting the smallest value (sample minimum, x_1) from the largest value (sample maximum, x_n) in the set: $r = x_n - x_1$. The range is given in the same units as the data and serves as a measure of **statistical dispersion**, like the variance and standard deviation.

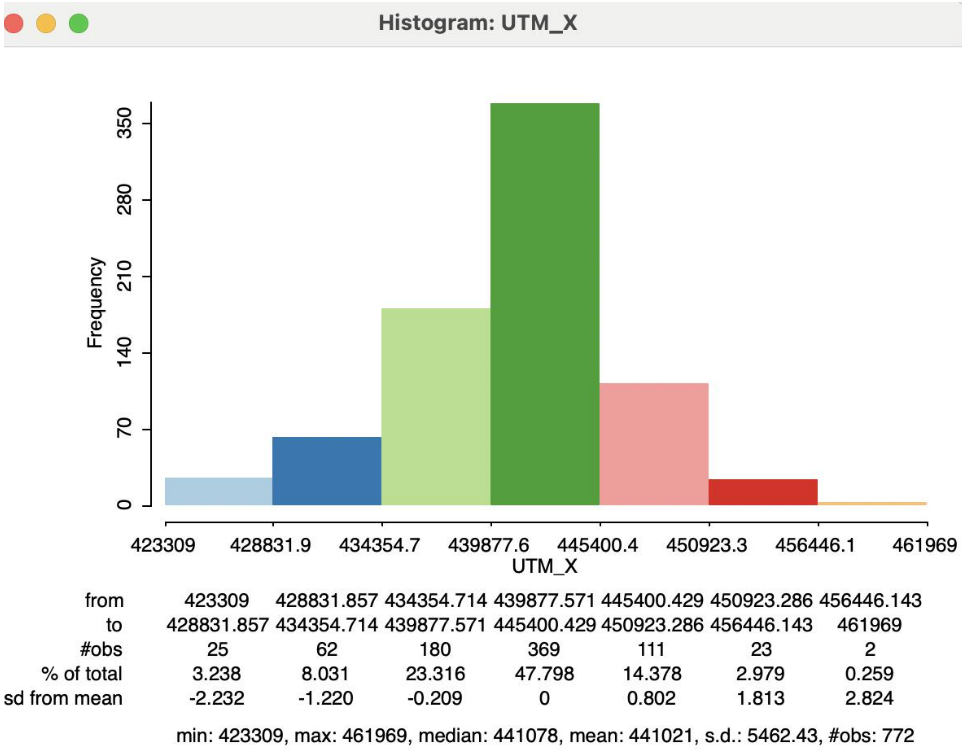
divided into a number of equal intervals (or bins), and the number of observations that fall within each bin is depicted proportional to the height of a bar.

The main challenge in creating an effective visualization is to find a **compromise between too much detail** (many bins, containing few observations) and **too much generalization** (few bins, containing a broad range of observations).

The histogram allows assessing about the **shape of a distribution**, detecting the existence of skewness and asymmetry. This graph is particularly useful for **variables of a continuous nature**, as it makes it easier to visualize their distribution by grouping these values into different categories or classes. In the example presented in Figure 2, a certain normality can be seen in the variable represented. If the standard deviation is multiplied by 2 and this quantity is added/subtracted to/from the mean[2], an interval of approximately (430,000; 452,000) is obtained in which about 95% of the probability mass is found, being the remaining 5% distributed between the two tails of the distribution.[3]

[2] The mean of a variable x (μ_x), often referred to as the arithmetic average, is calculated by dividing the sum of all values of this variable by the sample size: $\mu_x = \sum_{i=1}^{n} x_i/n$.

[3] In statistics, the normal distribution, also known as Gaussian distribution, represents a type of **continuous probability** distribution. The general form of its probability **density function** is $f(x) = (2\pi)^{-(1/2)}\sigma_x^{-1}exp[-(x-\bar{x})^2/2\sigma_x^2]$. Normal distributions are commonly used to **represent real-valued random variables** with unknown distributions due to the **Central Limit Theorem**, which asserts that the average of multiple samples from a random variable with a defined mean and variance will approximate a normal distribution as the sample size increases, assuming certain conditions are met. The normal distribution is **symmetric around the mean** value, which is also equal to the mode and median. It is **unimodal** and the **total area** between the curve and the x-axis, is equal to **one**. Approximately 68% of values from a normal distribution fall within one standard deviation (σ) of the mean, about 95% within two standard deviations, and roughly 99.7% within three standard deviations. This phenomenon is referred to as the 68-95-99.7 (empirical) rule, or the **three-sigma rule**. The most fundamental form of a normal distribution is the **standard normal distribution**, which occurs when the **mean is 0 and the variance is 1**.



from	423309	428831.857	434354.714	439877.571	445400.429	450923.286	456446.143
to	428831.857	434354.714	439877.571	445400.429	450923.286	456446.143	461969
#obs	25	62	180	369	111	23	2
% of total	3.238	8.031	23.316	47.798	14.378	2.979	0.259
sd from mean	-2.232	-1.220	-0.209	0	0.802	1.813	2.824

min: 423309, max: 461969, median: 441078, mean: 441021, s.d.: 5462.43, #obs: 772

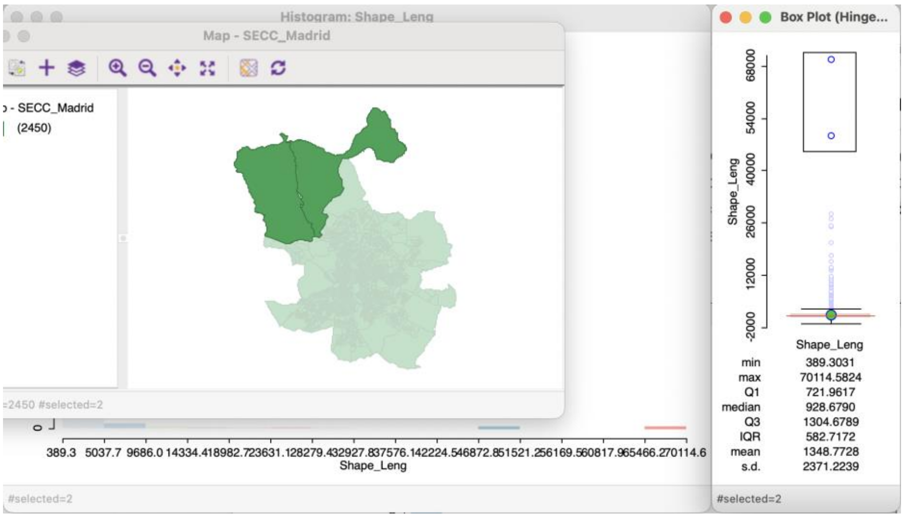
Figure 2. Histogram of the X coordinate distribution of the metro stations in Madrid, 2024.

2.1.2. Box Plot

Method of representation based on the calculation of the **quartiles and the median** of a variable, as well as obtaining the upper and lower quartiles or adjacent values.[4] The 'box' is a rectangle that is constructed in such a way that the lower value of the box is the first quartile (containing 25% of the observations) and the upper value is the third quartile (containing 75% of the observations). The median is highlighted in the middle of the box with a circle and a horizontal line (Figure 3).

The values are obtained by adding/subtracting to the median the product of the values of the third (first) quartile by 1.5 times (or 3 times) the interquartile range. The **outliers** are values located above (or below) the aforementioned peaks, which may not exist (when the variable has values highly concentrated near the median).

[4] In statistics, quartiles are a specific type of quantile that divides the data into **four equal parts** once the data is sorted from lowest to highest. There are three quartiles that segment the data into four groups. The first quartile (Q1), or lower quartile, marks the 25th percentile, meaning 25% of the data falls below this point. The second quartile (Q2) is the median, indicating that half of the data lies below this value. The third quartile (Q3), or upper quartile, is the 75th percentile, where 75% of the data is below this point.



	Shape_Leng
min	389.3031
max	70114.5824
Q1	721.9617
median	928.6790
Q3	1304.6789
IQR	582.7172
mean	1348.7728
s.d.	2371.2239

Figure 3. Histogram and box plot of the shape length of census sections of Madrid, 2024.

2.1.3. Thematic Maps

Standard exploratory data analysis uses tools that allow to observe the distribution of data (such as histograms), but not their **geographical location**. One way to observe the spatial distribution of variables is through the use of thematic maps. A thematic map is a cartographic representation of a spatial variable using symbols and colors that highlight differences in values. When the regions are colored based on the values of a variable, the thematic map is called **"choropleth map"** while a so called **"graduated symbol map"** would map the same data using a symbol sized proportionately to the data amount and placed within each county on the map. Choropleth maps are the most commonly used in ESDA. There are several choropleth maps depending on the way the continuous data is classified or divided into discrete intervals.

1) Quantile map

Quantile maps are thematic maps representing the overall spatial trend of a variable by dividing and grouping the data into categories with equal numbers of observations (Figure 4). The quantiles are values that divide a data sample into a number of categories so that each category (as far as possible) contains an **equal number of observations** (when the number of categories is 4, 5 or 6, we speak of quartiles, quintiles or sextiles, respectively).

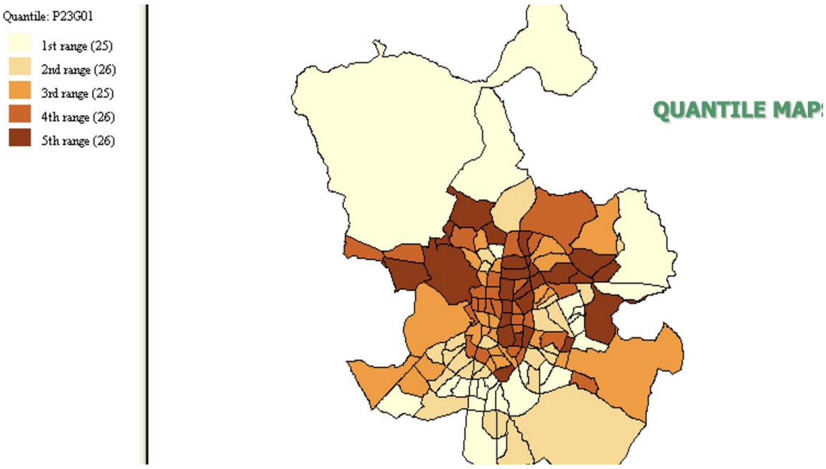
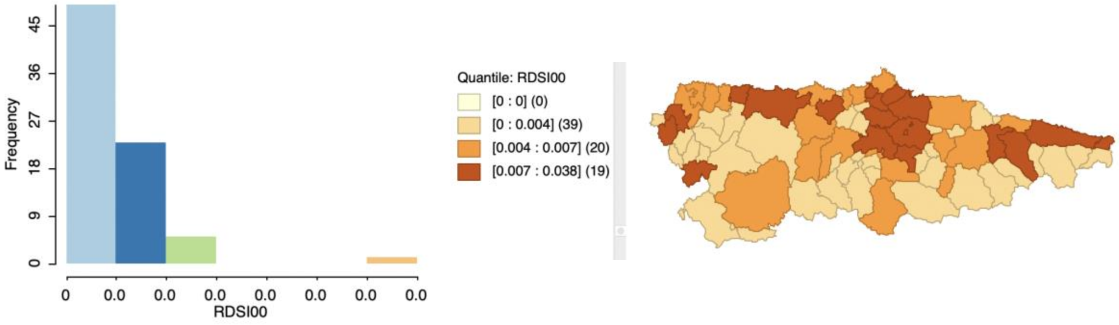


Figure 4. Quantile map of the higher education rate of the Madrid neighborhoods, year 2001.

The quantile map is **not useful** in cases where the variable in question has a distribution **far away from the normal distribution**. It is not useful when the variable is very asymmetric or contains a large number of observations with similar values as there will be **quantiles that cannot be defined**

doi:10.20944/preprints202503.2044.v1

as the same number of observations cannot be assigned to the different groups (Figure 5). In these cases, we must either increase or decrease the number of quantiles, to avoid null values in some of them, or use another map that is more suitable for representing highly non-normal variables.



(a) Histogram (b) Quantile map

Figure 5. Histogram and quantile map of phone lines of the Asturian municipalities.

2) Natural breaks map

A natural breaks map uses a nonlinear algorithm to group observations such that the within-group homogeneity is maximized, following the path-breaking work of Jenks (1977), among others. In essence, this is a clustering algorithm in one dimension to determine the break points that yield **groups with the largest internal similarity, i.e., the smallest internal variance**[5] (Figure 6). The algorithm to obtain the optimal break points is quite complex.

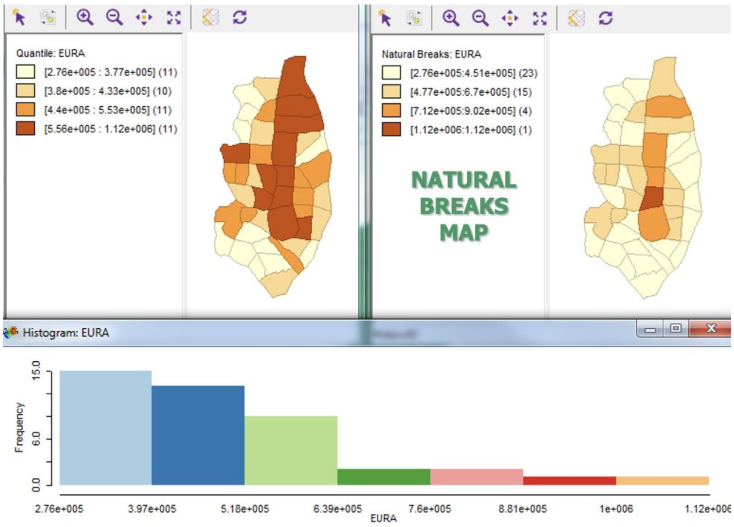

Quantile: EURA
[2.76e+005 : 3.77e+005] (11)
[3.8e+005 : 4.33e+005] (10)
[4.4e+005 : 5.53e+005] (11)
[5.56e+005 : 1.12e+006] (11)
Natural Breaks: EURA
[2.76e+005:4.51e+005] (23)
[4.77e+005:6.7e+005] (15)
[7.12e+005:9.02e+005] (4)
[1.12e+006:1.12e+006] (1)
NATURAL BREAKS MAP
Histogram: EURA
Frequency
0.0
6.0
15.0
2.76e+005
3.97e+005
5.18e+005
6.39e+005
7.6e+005
8.81e+005
1e+006
1.12e+006
EURA

Figure 6. Quartile *vs* natural breaks maps of house price per area in Madrid, year 2008.

The **format of the legend** in the natural breaks map differs slightly from the one used in other map types. Each interval is depicted as half open, with the lower value included - shown by the left square bracket **[** - and the upper value excluded - shown by the right parenthesis **)**. Similarly, the bounds of the lowest and highest category are shown, not the lowest and highest value as before.

3) Box map

[5] In statistics, variance (σ_x^2) measures the **spread of a random variable's values (x_i) around its mean**, calculated as the expected value of the squared differences from the mean: $\sigma_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$. **Standard deviation** (σ_x), derived as the **square root of variance**, quantifies the amount of variation or dispersion in a set of data values. Variance serves as the **second central moment** of a distribution and represents the covariance of the variable with itself.

In the box map, the category bounds are connected to the visualization of the distribution of a variable in a box plot. The point of departure is a quartile map. The four categories are extended to six bins, to separately identify the lower and upper outliers (Figure 7). The definition of outliers is a function of a multiple of the inter-quartile range (IQR), the difference between the values for the 75 and 25 percentiles. As is customary, there are two options for these cut-off values, or hinges in a box plot: 1.5 and 3.0. The box map uses the same convention.

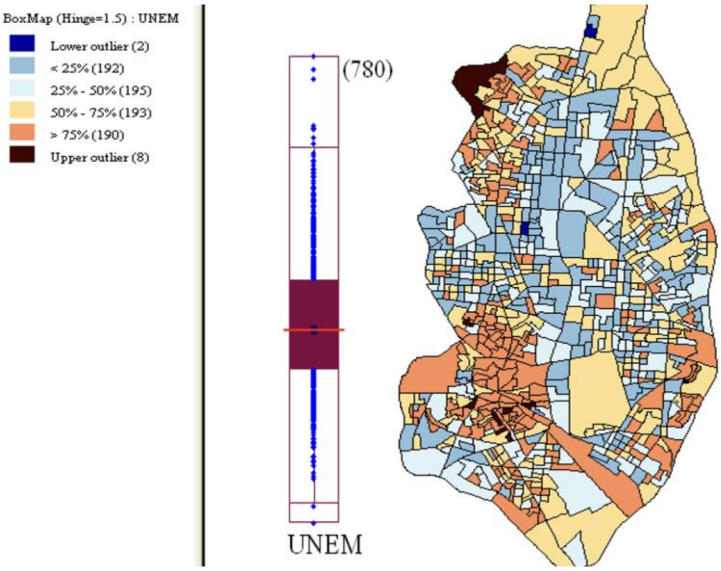


Figure 7. Box map and boxplot of the unemployment rates in Central Madrid.

The box map is used to identify extreme values and/or outliers in a spatial distribution. The outliers are elements of discontinuity in a variable, that is, they are exceptionally low/high values of the variable with respect to the rest of the distribution, which may not be representative of the overall distribution. They could affect the behavior of statistical tests. ESDA often detects, as outliers, values that are simply data entry errors or rare events, for which there is no explanation, in which case it is advisable to remove them, to avoid distortions in the subsequent analysis.

There are some other choropleth maps like the unique-values map, the percentile map, the standard deviation map and the cartogram. In practice, it is important to go beyond using a single map type, and to compare the similarities and differences between the patterns suggested by the various map classifications.

2.2. For Discrete Variables

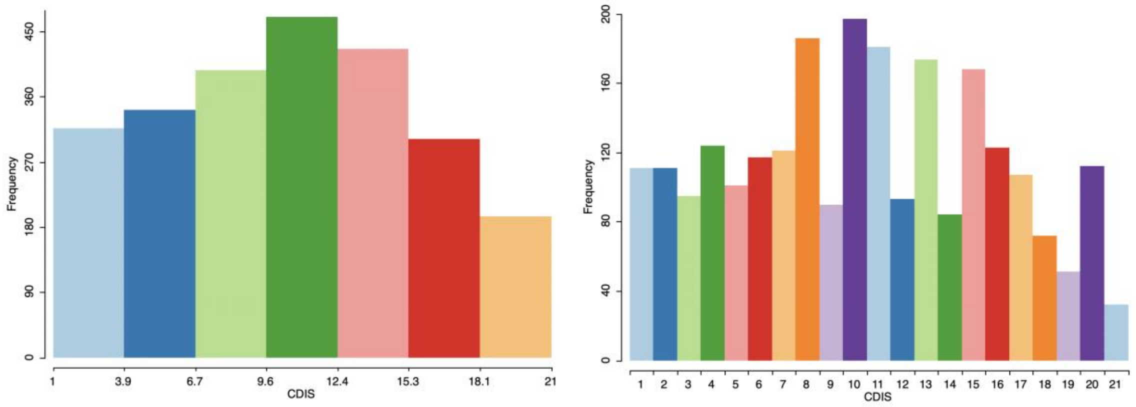
In mathematics and statistics, a quantitative variable may be continuous or discrete if it is typically obtained by measuring or counting, respectively (Wikipedia 2025). A discrete variable within a specific range of real values is defined as one where each permissible value within the range is separated by a positive minimum distance from its nearest allowable neighbor. Common examples are variables that must be integers: categorical, ordinal, or only the integers 0 and 1.

The ESDA of one discrete variable makes sense in itself, but also as part of the identification phase of a discrete dependent variable econometric model. This is the case of qualitative response models: spatial count data models, binary spatial models, spatial logit and spatial probit models, etc.

Next, we present three representations: bar chart, unique values map and co-location map.

2.2.1. Bar Chart

A bar chart is not technically a histogram, since categories do not imply necessarily any order. It is a chart that presents discrete variables with rectangular bars with heights or lengths proportional to the values that they represent (Figure 8). A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured integer value. The bars can be plotted vertically or horizontally.



(a) Histogram (b) Bar chart

Figure 8. Histogram and bar chart of the number of census sections of Madrid by districts.

The histogram divides the categories as if they were not integer but real numbers, even including decimals. The most appropriate in this case, where the categories express districts of Madrid, is the bar chart, which makes it clear that each category corresponds to an integer number. On the vertical axis, the values of the variable: number of census sections present in each district are presented.

2.2.2. Unique Values Map

Unique values maps use different colors or symbols to shade or represent objects with different values for an attribute. In this map, each category representing a variable must be distinguished by a different integer value, though those values are typically not meaningful, and the numerical values do not imply any ordering of the categories. Unique-values Map can be made based on any point, line or area vector layer. The unique values map gives each census tract on the map the color corresponding to the district to which it belongs. The legend indicates, for each district, the number and color assigned, as well as the number of corresponding census tracts.

2.3. Map Classification, Legends and Colors

A map classification is the process of binning the observations from a continuous distribution into discrete categories, each of these corresponding to a different color (or shading) on the map. There have been proposals to visually represent the full continuity of the distribution, e.g., by the use of a full color spectrum (Brewer 1997). However, this quickly becomes impractical for larger data sets. Alternatively, symbols with different sizes and/or colors can be used as well.

The number of categories and how the cut-off or break points are determined are usually based endogenously by exploiting the properties of the underlying distribution. Common methods use, for example, the quantiles of the distribution, in a quantile map, or divide the range of the variable into a number of equal intervals, similar to what is customary for a histogram.

The legend is the way in which the map classification is symbolized, typically positioned next to the map itself. The choice of color can have major impacts on the perception of value and pattern in the map. For example, red colors are typically associated with hot locations, whereas blues are associated with cold. More importantly, red is often used to represent or imply danger, especially in political maps. Regarding the gradation of values represented in the legend bar, three broad categories can be distinguished:

a) For maps where the values represented are **ordered** and follow a single direction, from low to high, a **sequential legend** is appropriate. Such a legend typically uses a **single tone** and associates **higher categories with increasingly darker values**.

b) In contrast, for the maps representing the **extreme values of a variable**, the focus should be on the central tendency (mean or median) and how observations sort themselves away from the center, either in downward or upward direction. An appropriate legend for this situation is a **diverging legend**, which emphasizes the extremes in either direction. It uses **two different tones**, one for the downward direction (typically blue) and one for the upward direction (typically red or brown).

c) Finally, for **categorical data** a qualitative legend is appropriate; that is, no order should be implied (no high or low values) and the legend should suggest the equivalence of categories.

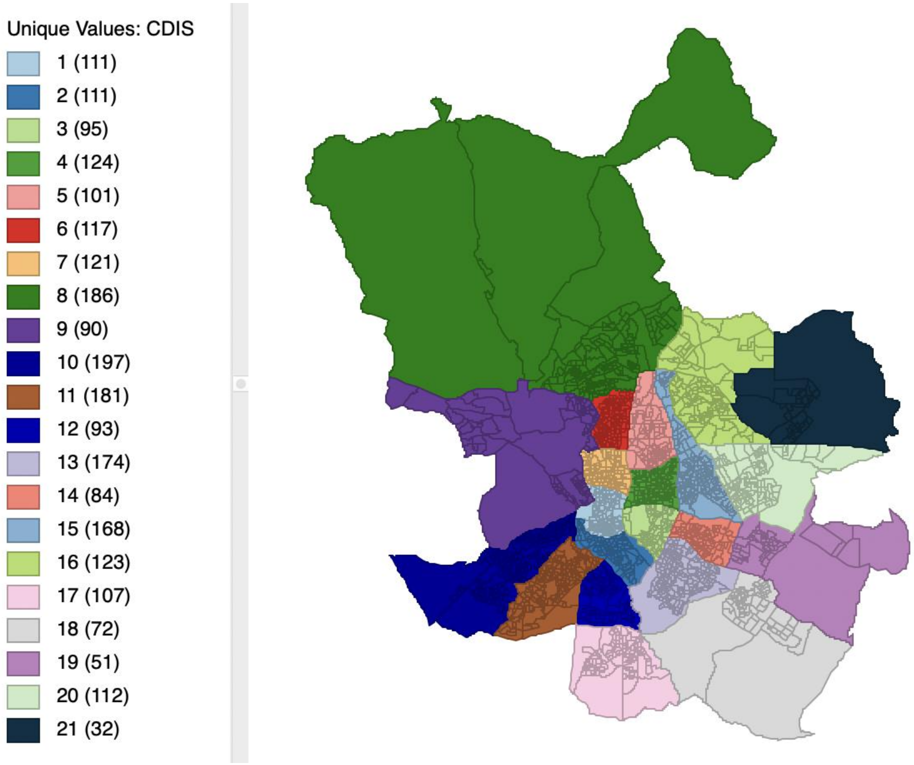


Figure 9. Unique values map of the number of census sections of Madrid by districts, 2024.

3. Basic Multivariate ESDA

Multivariate analysis is based on the principles of multivariate statistics. Typically, it is used to address situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important.

3.1. For Continuous Variables

3.1.1. Scatter Plot

A scatter plot, often employed to evaluate the correlation[6] **between two variables**, consists of a diagram with two perpendicular axes, each representing one of the variables (Anselin 2023). Points

[6] **Pearson's correlation coefficient** is a statistical metric that quantifies the degree to which two variables (x, y) are **linearly associated**, indicating their tendency to move together at a consistent rate. It is defined as the quotient between the covariance of these variables and the product of their respective standard deviation: $r_{xy} = \sigma_{xy}/\sigma_x \cdot \sigma_y = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/\sqrt{(x_i - \bar{x})^2(y_i - \bar{y})^2}$. It is frequently used to depict straightforward relationships **without implying causality**. The correlation coefficient, varying from -1 to 1, indicates the **strength and direction of a linear relationship** between

doi:10.20944/preprints202503.2044.v1

on the plot represent observation pairs (X, Y). Commonly, a linear regression line is drawn through these points to provide a **summary of the relationship between the variables**. The regression[7] is:

$$y = a + bx + u \qquad (1)$$

where a is the intercept, b is the slope, and u is a random error term. The method used to estimate the coefficients involves minimizing the sum of squared residuals, a process known as least squares fit or ordinary least squares (OLS).[8]

The intercept a is the average of the dependent variable (y) when the explanatory variable (x) is zero. The slope shows how much the dependent variable changes on average (Δy) for a one unit change in the explanatory variable (Δx). Hence, the scatterplot can be of great help in identifying relationships between variables in a regression model: both between the explanatory variables x (to study multicollinearity), and between the variable y and each of the x (possible non-linearity, intensity of correlation, etc.).

There are two important issues which must be considered: non-linear regressions and standardization of the x and y variables.

1) Non-linear relationships

The regression technique used on the scatter plot assumes a linear relationship between variables. This approach may not be suitable when the relationship is non-linear, such as in the case of second-degree polynomial or U-shaped relationships or shows complex sub-patterns in the data (Figure 10).

two variables. An absolute value of 1 signifies a perfect linear equation where all observations lie precisely on a line. If the coefficient is +1, y increases with x; if it is -1, y decreases as x increases. A correlation of 0 means there is no linear association between the variables. Correlation analysis has **limitations**. It does not consider the influence of other variables beyond the two examined, nor does it establish causation, and it is ineffective at describing relationships that are curvilinear.

[7] Simple linear regression in statistics is a model that uses **one explanatory variable to predict the dependent variable through a linear function**. It deals with two-dimensional sample points, mapping one independent variable to one dependent variable (typically represented as x and y in a Cartesian coordinate system).

[8] OLS is a method used in simple linear regression to determine the best-fit line by **minimizing the sum of the squares of the differences between observed and predicted values** of the dependent variable, that is the sum of squares of the regression error term: $\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n} e_i^2$. This **optimizatio**n technique seeks to find the line that most closely aligns with the data points, providing unbiased estimates for the intercept ($a = \bar{y} - b\bar{x}$) and slope ($b = \sum_1^n (x_i - \bar{x})\,(y_i - \bar{y}) / \sum_1^n (x_i - \bar{x})^2$) parameters in the model. The slope can also be expressed as: $b = r_{xy}\left(s_y/s_x\right)$. In the case of a simple linear regression with **standardised variables**, the standard deviation of both x, y variables is one and, therefore, the slope will coincide with the correlation coefficient. In this case, since $\bar{x} = 0$, the intercept is equal to the mean value of the dependent variable.

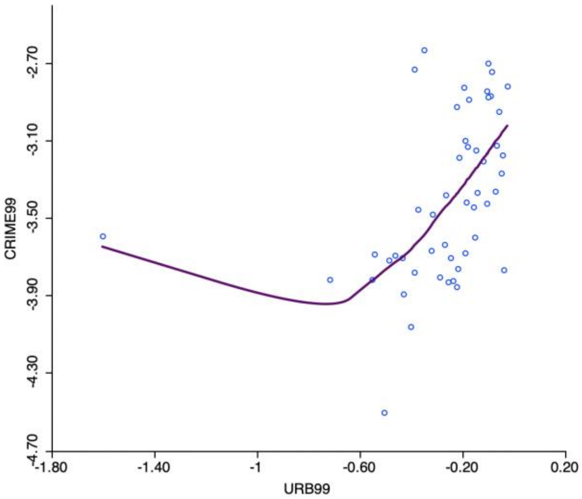


Figure 10. Scatterplot of urbanization level on crime rates in the Spanish provinces, year 99.

An alternative consists of a so-called local regression fit, which is a nonlinear procedure that computes the slope from a subset of observations on a small interval on the X-axis, weighting down values furthest from the center of the interval (like a kernel smoother). As the interval (the bandwidth) moves over the full range of observations, the different slopes are combined into a smooth curve (see, for example, Loader 2004).

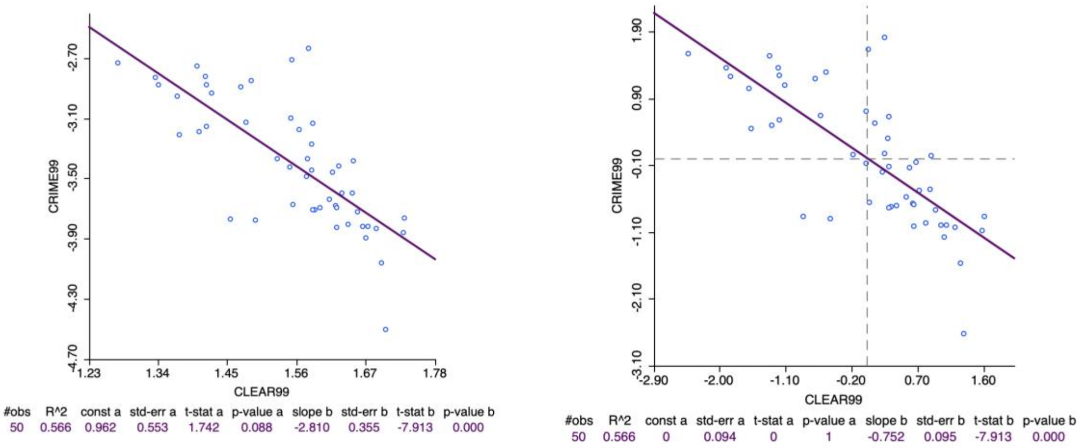
A local regression fit reveals potential nonlinearities in the bivariate relationship or may suggest the presence of structural breaks. Two common implementations are Locally Weighted Scatterplot Smoother (LOWESS) and Local Polynomial Regression (LOESS), which apply different fitting algorithms. GeoDa implements the LOWESS function, which is based on the function called "lowess" in the R stats package. It has the following smoothing parameters: bandwidth, number of iterations, and delta factor.

- The **bandwidth** is the proportion of points in the plot that influence smoothing at each value, so as larger values give more smoothness. For example, the default bandwidth of **0.20** implies that for each local fit (centered on a value for X), about **one fifth of the range of X-values** is considered.
- **Iterations**: are the number of **"robustifying" iterations** which should be performed. Using **smaller values will speed up** the smoothing.
- **Delta factor**: Small values of delta speed up computation because local polynomial fit is only computed for a small amount of data at each data point, filling in the fitted values for the skipped points with linear interpolation.

2) Standardization of the x, y variables

When both y and x are standardized (mean of zero and variance of one), the slope of the regression line matches the correlation coefficient between the two variables, with the intercept being zero (Figure 11). It is important to recognize that while correlation represents a symmetric relationship, regression does not, leading to a different slope if the roles of explanatory and dependent variables are reversed. Like linear regression, the correlation coefficient is also limited to measuring linear relationships between variables, as evidenced by its equivalence to the regression slope among standardized variables.

Hence, the scatter plot or point cloud is a description of the relationship or dependence between two x–y variables. The shape of this point cloud reflects the degree of correlation between the two variables, which may be zero (if the points form a circle), linear (if the points represent an ellipse) or non-linear (if the points take any other shape). The simplest and most useful function in most cases is the line. In this case, the dependence between the variables is measured through the linear correlation coefficient. The statistical significance of this coefficient measure whether the relationship between the two variables is linear or not, and whether there are certain outliers that detract from its representativeness.



(a) Raw data (b) Standardized data

Figure 11. Scatterplots of police clearance on crime rates in the Spanish provinces, year 99.

3.1.2. Scatter Plot Matrix

A scatter plot matrix displays the bivariate relationships between various pairs of variables. The individual scatter plots are organized in a matrix format where each variable alternately appears on the X-axis and the Y-axis (Figure 12).

When utilized with standardized variables (mean zero and variance one), the scatter plot matrix serves as the visual equivalent of a correlation matrix. Typically, these matrices illustrate the linear relationships between variables, although the plots themselves are not symmetrical. Moreover, certain variables may not suitably function as both dependent (on the Y-axis) and explanatory (on the X-axis). For instance, while elevation might influence socio-economic outcomes, it is unlikely that socio-economic factors could reciprocally impact elevation.

3.1.3. Parallel Coordinate Plot

In the parallel coordinate plot (PCP), data points are replaced by data lines, allowing a large number of variables to be considered. This approach was originally suggested by Inselberg (1985), and it has become a main feature in many visual data mining frameworks. In a PCP, each variable is depicted as a parallel axis, with each observation illustrated by a line connecting points across these axes. The addition of axes allows for the easy incorporation of more than three variables. The main visual limitation is the number of observations which can result in an overly congested display of lines (Figure 13).

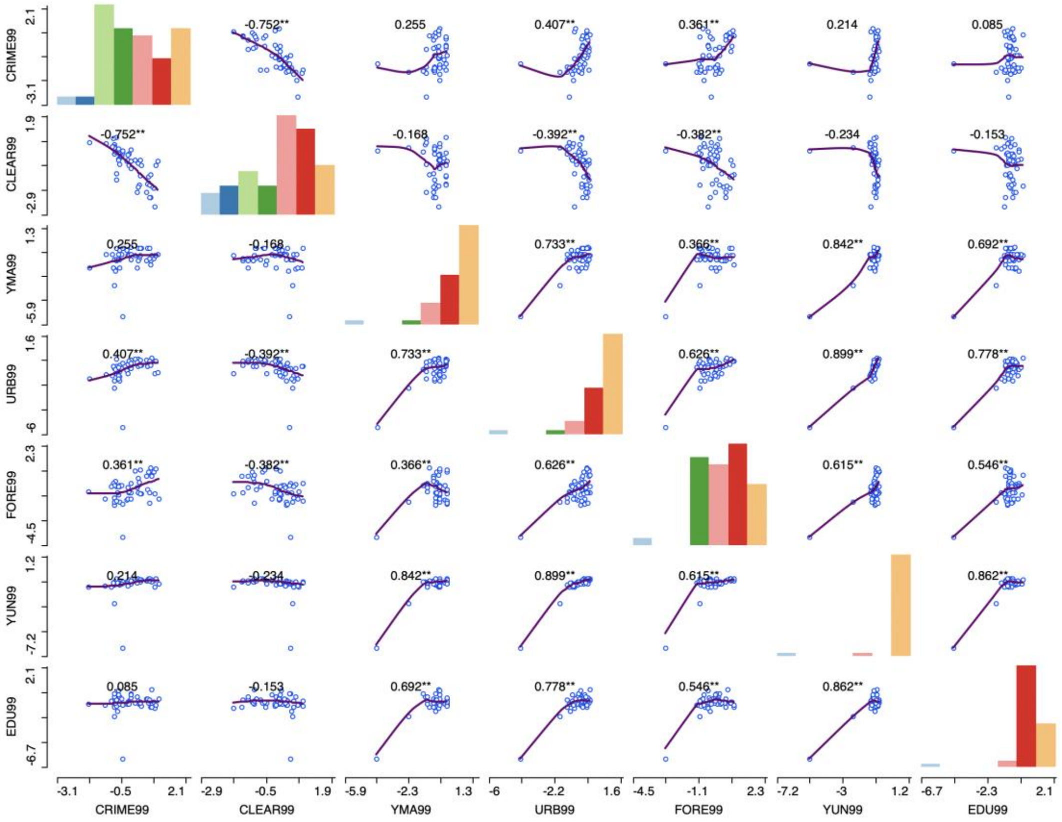
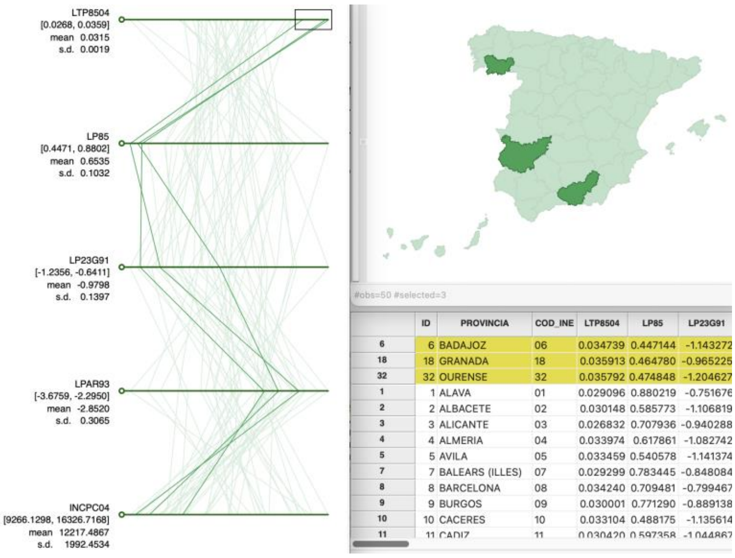


Figure 12. Scatterplots of crime rates and some explanatory variables in the Spain, year 99.



#obs=50 #selected=3

	ID	PROVINCIA	COD_INE	LTP8504	LP85	LP23G91
6	6	BADAJOZ	06	0.034739	0.447144	-1.143272
18	18	GRANADA	18	0.035913	0.464780	-0.965225
32	32	OURENSE	32	0.035792	0.474848	-1.204627
1	1	ALAVA	01	0.029096	0.880219	-0.751676
2	2	ALBACETE	02	0.030148	0.585773	-1.106819
3	3	ALICANTE	03	0.026832	0.707936	-0.940288
4	4	ALMERIA	04	0.033974	0.617861	-1.082742
5	5	AVILA	05	0.033459	0.540578	-1.141374
7	7	BALEARS (ILLES)	07	0.029299	0.783445	-0.848084
8	8	BARCELONA	08	0.034240	0.709481	-0.799467
9	9	BURGOS	09	0.030001	0.771290	-0.889138
10	10	CACERES	10	0.033104	0.488175	-1.135614
11	11	CADIZ	11	0.030420	0.597358	-1.044867

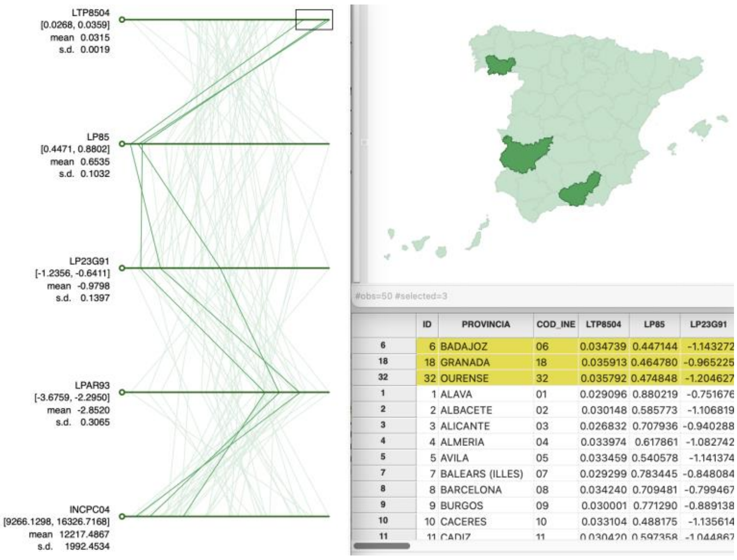
Figure 13. PCP of GDP growth rates and some explanatory variables in Spain, 1985-2004.

Hence, the observations are represented in the form of multiple segments that link their position on each axis according to the values of the variables they adopt. Each variable is rescaled so that the minimum value is on the far left and the maximum on the far right.

A primary aim of using PCP is to identify clusters and outliers within a multi-attribute space. Clusters appear as groups of lines that track similar trajectories. Outliers in a PCP are those lines that significantly deviate in pattern from the majority, resembling outlying points in a multidimensional scatter. The main use of this graph is to identify clusters of values or outliers in certain observations that may also be spatial in nature. Data standardization makes that data points on each axis are given

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2025 doi:10.20944/preprints202503.2044.v1

in standard deviational units and become directly comparable between different variables. It is also useful to detect outliers (Figure 14).



#obs=50 #selected=3

	ID	PROVINCIA	COD_INE	LTP8504	LP85	LP23G91
6	6	BADAJOZ	06	0.034739	0.447144	-1.143272
18	18	GRANADA	18	0.035913	0.464780	-0.965225
32	32	OURENSE	32	0.035792	0.474848	-1.204627
1	1	ALAVA	01	0.029096	0.880219	-0.751676
2	2	ALBACETE	02	0.030148	0.585773	-1.106819
3	3	ALICANTE	03	0.026832	0.707936	-0.940288
4	4	ALMERIA	04	0.033974	0.617861	-1.082742
5	5	AVILA	05	0.033459	0.540578	-1.141374
7	7	BALEARS (ILLES)	07	0.029299	0.783445	-0.848084
8	8	BARCELONA	08	0.034240	0.709481	-0.799467
9	9	BURGOS	09	0.030001	0.771290	-0.889138
10	10	CACERES	10	0.033104	0.488175	-1.135614
11	11	CADIZ	11	0.030420	0.597358	-1.044867

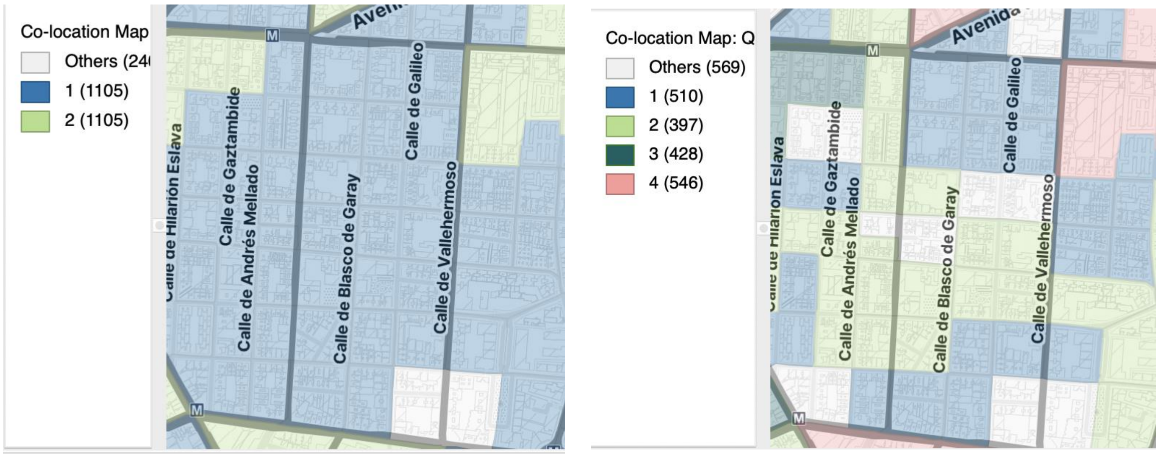
Figure 14. PCP of standardized GDP growth rates and some explanatory variables in Spain.

3.2. For Discrete Data: Co-Location Map

A co-location map combines the information from two or more categorical variables into a unique values map that shows those locations where the categories match. In essence, the process boils down to finding those locations where the codes for different categorical variables match (Anselin 2023).

When the variables under consideration take on binary values, a co-location map can be constructed by hand as a unique values map of the product of the respective indicator variables. Only those locations that take on a value of 1 for all variables will be coded as one in the resulting unique values map. However, a co-location map is different in that it also provides information on matches of locations with 0 for all variables, as well as on the locations of mismatches. In sum, rather than just two categories (match and no match), there are three: match of 1, match of 0, and mismatch.

When the variables pertain to multiple categories, the logic of the co-location map is the same, but it must be applied with caution. It is based on the equality of the categorical codes for each variable (Figure 15). It is up to the user to ensure that the categories across variables are meaningful, since the co-location is based on the variables having the same code. For example, this is useful when comparing the extent to which the quartiles across different variables occur at the same locations.



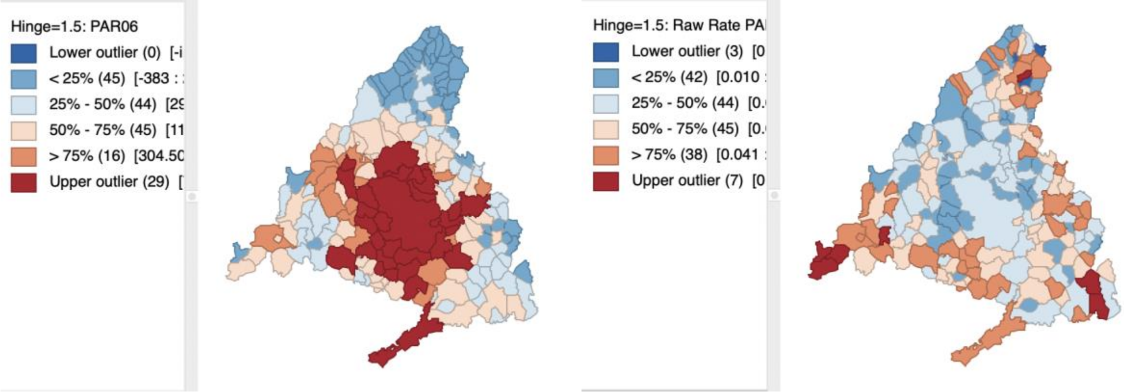
(a) Binary values

(b) Multiple values

Figure 15. Co-location map of census section length and area in Madrid, year 2024.

3.3. For Rates or Proportions

Rates or proportions are formed by dividing a numerator by a denominator. Typically, the **numerator represents a specific event**, like the occurrence of a disease, and the **denominator represents the population at risk**. This type of data is widely used in fields like public health and criminology. However, the underlying principles are also **applicable to any ratio or proportion**, such as the unemployment rate, which is calculated as the number of unemployed individuals in a region relative to the labor force, or other per capita measures, like gross regional product per capita. These variables are called by Anselin (2023) ***"spatially intensive"*** because they are not directly related to the size of the observational unit, being the only ones suitable for representation on choropleth maps. On the other side, the so-called ***"spatially extensive"*** variables, like all population-derived variables (total population, number of housing units, count of crimes, etc.), have larger values in larger observational units, and *vice versa*. These variables ought to be **standardized** in a manner that showcases intrinsic variation, as opposed to variations stemming from the size of the observational unit. This can easily be achieved by dividing by a certain **measure of size** (Figure 16).



(a) Spatially extensive variable

(b) Spatially intensive variable

Figure 16. Unemployment and Unemployment rate, year 2006.

The ESDA of variable of rates makes sense in itself, but also as part of the identification phase of an econometric model with a dependent variable of ratio. This is the case of the **spatial logit and probit models**, which can be considered a special case of a proportion (e.g., the proportion of success is either 0% or 100%), the tobit model, the beta regression model or fractional response models. The tobit models, which are traditionally censored data, when censoring is at 0 or 1. The **spatial beta regression** is designed for modeling rates taking values in the open interval (0, 1), which can handle data that exhibit skewness or are bounded at both ends (0 and 1), unlike linear regression models. The **spatial fractional response models** are used when the response variable can be a fraction or proportion that does not necessarily fall into a binary category but includes 0 and 1 as possible outcomes.

3.3.1. Raw Rate Map

A raw rate or crude rate map is an unadjusted rate or proportion where $\boldsymbol{O_i}$ denotes the numerator's value for unit $\boldsymbol{i}$ and $\boldsymbol{P_i}$ indicates the matching denominator, as in Eq. 2, as follows:

$$r_i = \frac{O_i}{P_i} \tag{2}$$

Hence, r_i is a spatially intensive variable of O_i. This variable is normalized by P_i, which must be a relevant scale factor like the total population in unit i or another aggregate like total area. The classic example is to use the **population at risk**, such as the population exposed to the risk of **a disease**. This implies that the rate is interpreted as an **estimate for the underlying risk**.

3.3.2. Relative Risk or Rate

Relative rate or relative risk is the ratio of the observed **rate at a specific location to a reference rate**. The concept of "risk" has many meanings but in this context, it represents **the probability of an event occurring**. There are two emblematic examples: the **standardized mortality rate** (SMR) in demography and public health, and the **location quotient** (LQ) in regional economics. Specifically, LQ is used for the proportion of employment in a specific sector. The approach involves comparing the observed rate in a small area with a **national or regional benchmark**. Relative risk or rate can be defined as follows:

$$RR_i = \frac{r_i}{\pi} \tag{3}$$

for π **reference risk level or aggregated** (national, regional) **benchmark**, which is, in turn, calculated by taking the ratio of the total of all numerators to the total of all denominators:

$$\tilde{\pi} = \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} P_i} \tag{4}$$

If the ratio $\tilde{\pi}$ would be uniformly applicable across all observations, it would represent the **expected value for each respective observation**. In fields like demography and public health, this expected value might correspond to the **anticipated number** of deaths or disease incidences. In regional economics, it could represent the **anticipated proportion** of employment -within a specific sector- matching the reference proportion. Generally, this concept can be articulated as:

$$E_i = \pi \times P_i \tag{5}$$

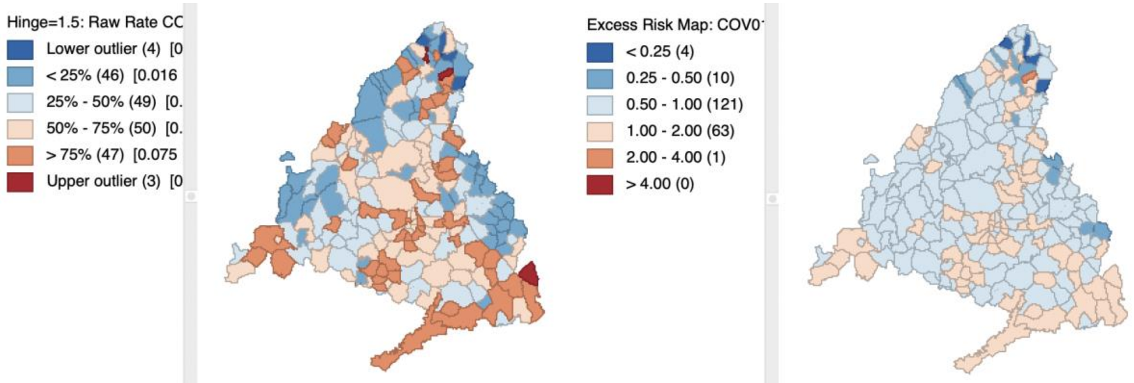
Specifically, it compares the observed number of events (O_i) against the expected number of events (E_i) that would occur if a reference risk level π (benchmark) were applied. Therefore, the relative risk is not only the ratio of the **observed rate over the reference rate**, but, equivalently, the **observed number of events over the expected number of events**:

$$RR_i = \frac{O_i/P_i}{E_i/P_i} = \frac{O_i}{E_i} \tag{6}$$

If a unit's rate aligns with the (regional) reference rate, its relative risk equals one. Rates above one indicates an **excess**, while rates below one indicates a **deficit**. The interpretation varies by context. For instance, in disease analysis, a relative risk greater than one suggests higher-than-expected disease prevalence in the area. Conversely, in regional economics, a location quotient above one indicates that **employment in a particular sector exceeds local demand**, suggesting it is an **export-oriented sector**.

3.3.3. Excess Risk Map

An excess risk map is a unique type of rate map that classifies observations according to whether they are **below or above the benchmark relative risk rate of one** (Figure 17). It features a divergent legend, with blue shades representing values less than one and brown shades for values more than one. The classification intervals include 1-2, 2-4, and greater than 4 for higher values (double and four times the reference rate), and 0.5-1, 0.25-0.5, and less than 0.25 for lower values (half and one fourth the reference rate).



(a) Raw rate map (b) Excess risk map

Figure 17. COVID-19 incidence raw rate and risk map, Madrid Region, January 2021.

3.3.4. Empirical Bayes Smoothed Rate Map

In numerous empirical studies, the primary focus consists of understanding the spatial distribution of the **underlying risk or likelihood of an event**. Given that this risk varies geographically, it is depicted through a conceptual '**risk surface**'—an unobserved entity that requires estimation.

The **crude rate** effectively estimates unknown risks because it is **unbiased**, hitting the target on average.[9] However, its **precision, or variance**, is not constant (heteroskedastic), fluctuating with the size of the **population at risk**, primarily due to the nature of how it is calculated—by dividing one quantity by another. The variance in these types of variables is affected because in **smaller denominators**, the ratio is **more sensitive to changes** in the numerator: even a slight change in the numerator, such as the occurrence of **a new single case**, can lead to significant fluctuations in the ratio, thereby **increasing the variance**.

Conversely, in **larger populations**, changes in the numerator have a **less pronounced effect on the ratio**, resulting in **lower variance**. The larger the denominator, the more it dilutes the impact of changes in the numerator. Heteroskedasticity can lead to inefficiencies in estimation and difficulties in inferential procedures, such as hypothesis testing and the construction of confidence intervals, also complicating the interpretation of maps, particularly in identifying outliers.

Moreover, creating meaningful maps is challenging in spatial units with **no observed events**. To improve precision across these variables, several methods have been proposed that **trade off some bias in order to increase overall precision**, by "*borrowing strength*". Here, some methods to refine risk estimates through **rate smoothing** will be discussed (Appendix A presents more formally some important concepts related to variance instability and Bayesian statistics).

[9] If the numerator of the ratio variable represents the number of successes or occurrences and the denominator represents the total opportunities for the event to happen, then their division gives a direct estimate of the **likelihood or probability of the event per unit of the population**. According to the **Law of Large Numbers**, the average of the results obtained from a large number of trials or sample size (e.g. districts) should be close to the expected value and will tend to become closer as more trials are performed: $\underset{n\to\infty}{plim}\, E(\hat{r}) = \pi$. On its side, the **Central Limit Theorem** supports the use of ratio variables as **unbiased estimators** by stating that when an **adequate sample size** is used, the distribution of the sample ratios will approximate a **normal distribution** centered around the true probability: $r{\sim}N(\pi, \sigma_r^2)$.

The Empirical Bayes (EB) method utilizes the Poisson-Gamma model,[10] in which the observed event count (O) is treated as a result of a Poisson distribution with a stochastic intensity (mean or expected value), denoted as $\boldsymbol{\pi} \times \boldsymbol{P}$ (being $\boldsymbol{\pi}$ stochastic and $\boldsymbol{P}$ deterministic). The Bayesian component involves using a prior Gamma distribution for the risk parameter $\boldsymbol{\pi}$. Consequently, this approach yields a specific form for the posterior distribution, which is also Gamma, characterized by specific mean and variance values:

$$E(\pi) = \mu = \frac{O + \alpha}{P + \lambda} \tag{7}$$

$$Var(\pi) = \frac{O + \alpha}{(P + \lambda)^2} \tag{8}$$

where α and λ are the **shape and rate** parameters of the prior Gamma distribution.[11]

In the EB method, the parameters **α and λ are derived from the data**. The **EB smoothed rate** is calculated as a **weighted average of the raw rate** (denoted as r) of each small area, and the **prior estimate** (denoted as λ), which is the **reference rate** (the overall state or regionwide) average or a similar benchmark[12], with weights based on the population at risk in those areas. The fundamental concept of smoothing entails *borrowing strength* from a **larger geographical area** (λ) to refine **the initial estimate** derived from raw data (the crude rate, r_i). Since r_i is the unbiased estimator of the underlying risk (π_i), this method introduces some bias as a trade off in order to increase overall precision:

$$\pi_i^{EB} = w_i r_i + (1 - w_i)\lambda \tag{8}$$

In practical terms, this means that areas with smaller populations at risk will see more significant adjustments to their rates, while larger areas will experience minimal adjustments (Anselin, Lozano-Gracia, and Koschinky 2006).

In this case, the weights are as follows:

$$w_i = \frac{\sigma^2}{\sigma^2 + \mu/P_i} \tag{9}$$

In the EB method, the mean μ and variance σ^2 of the prior, which define the rate (λ) and shape parameters (α) of the Gamma distribution, are derived from the data. For μ this estimate is simply the **reference rate** or aggregated benchmark: $\sum_{i=1}^{n} O_i / \sum_{i=1}^{n} P_i$. The estimate of the variance is as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n} P_i (r_i - \mu)^2}{\sum_{i=1}^{n} P_i} - \frac{\mu}{\sum_{i=1}^{n} P_i / n} \tag{10}$$

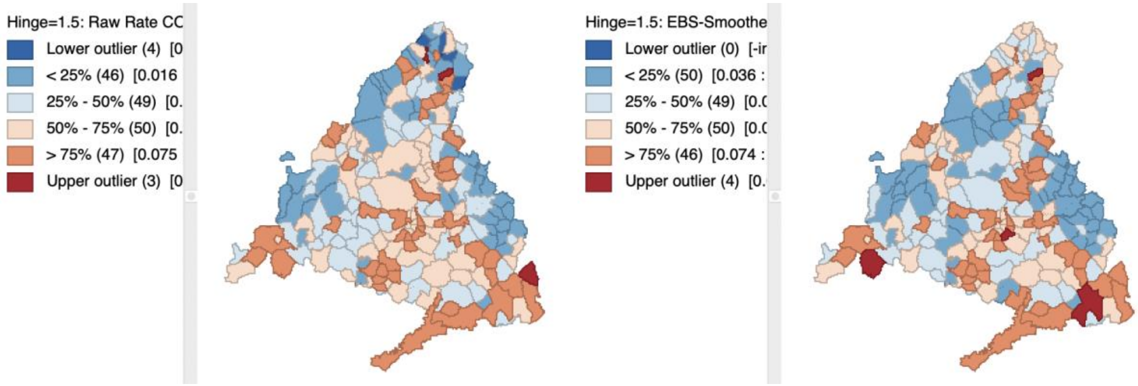
Although straightforward to compute, the estimate for the variance may result in negative values. When this occurs, the typical solution is to assign a value of zero to σ^2. Consequently, the weight w_i is also set to zero, effectively aligning the smoothed rate estimate (π_i^{EB}) with the reference rate (λ).

[10] The Poisson-Gamma or Gamma-Poisson distribution is a statistical distribution for overdispersed count data. It is a versatile **two-parameter family of continuous probability distributions**. The exponential distribution and chi-squared distribution are special cases of the gamma distribution. There are two equivalent parameterizations in common use: 1) With a shape parameter α and a scale parameter θ and 2) with a shape parameter α and a rate parameter $\lambda = 1/\theta$. Bayesian statisticians prefer the (α, λ) parameterization, while the parameterization with α and θ appears to be more common in econometrics and other applied fields (Wikipedia 2025a).

[11] A Gamma distribution, characterized by shape parameter α and scale parameter θ, has a mean of the product of its shape and scale parameters: $\mu = \alpha\theta = \alpha/\lambda$ and the variance is $\sigma^2 = \alpha\theta^2 = \alpha/\lambda^2$.

[12] From here, λ is the equivalent of π.

The impact of Empirical Bayes smoothing is that the initial raw rate is adjusted **towards the overall rate based on the size of the population**: regions with larger populations experience minimal adjustment, while regions with smaller populations may see significant alterations in their smoothed rates (Figure 18).



(a) Raw rate map

(b) EB smoothed rate map

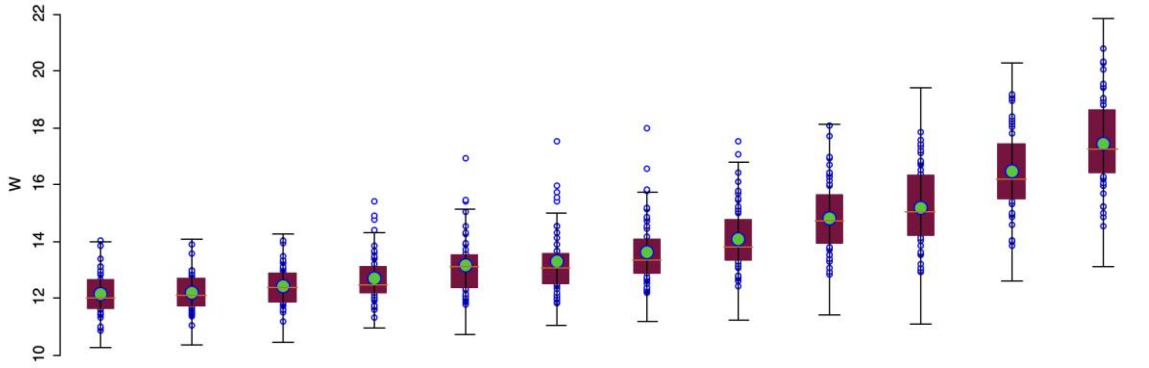
Figure 18. Comparative raw rate vs EB smoothed rate maps. Unemployment rate, year 2006.

3.4. For Space-Time Data

There are ESDA methods which enables the display of a variable's progression over time using maps and graphs, akin to comparative statics (Anselin 2023). However, a notable drawback is its orientation towards cross-sectional analysis. Specifically, each time period is treated as a distinct cross-sectional variable, offering considerable versatility in variable grouping but lacking inherent temporal awareness and a structured panel data framework. The ESDA of space-time data makes sense in itself, but also as part of the identification phase of panel data econometric models like spatial panels and difference-in-difference, among others.

3.4.1. Space-Time Box Plot

To evaluate how the distribution of a spatial variable x_i changes over time t, $x_{it}, x_{it-1}, x_{it-2}, \dots$, plotting box plots side-by-side can be very effective. This side-by-side arrangement is the standard approach when using the box plot feature for a variable that is grouped by time. Combining descriptive statistics with visual representations of the distribution offers a powerful method for analyzing broad trends over time (Figure 19).

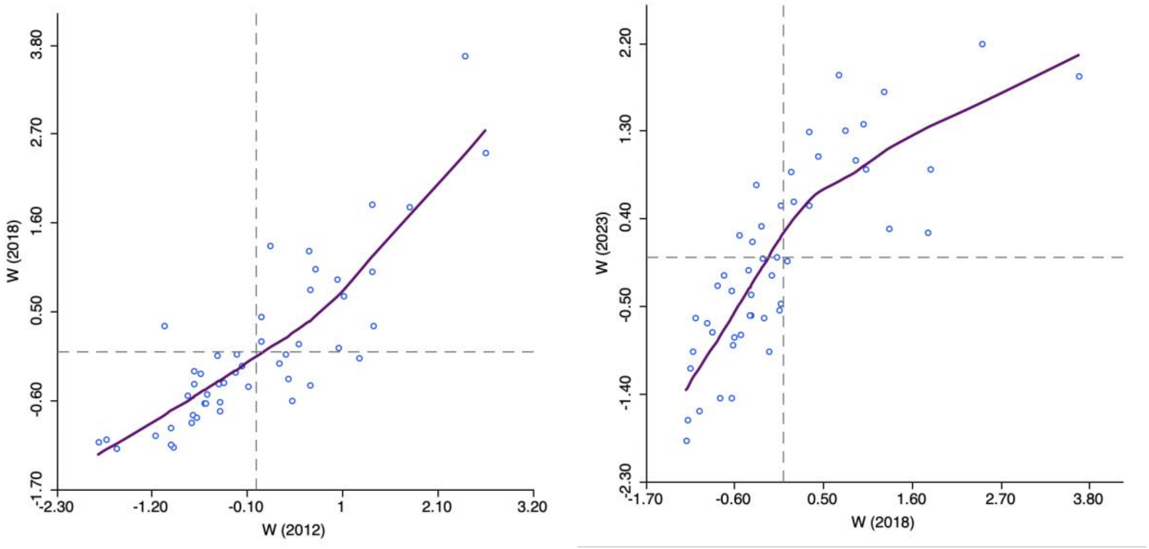


	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
min	10.8600	11.0200	11.1600	11.3300	11.7900	11.8100	12.2100	12.4100	12.8600	12.9100	13.8400	14.5500
max	14.0300	13.9000	14.0500	15.4100	16.9400	17.5500	18.0300	17.5600	18.1100	17.8900	19.2300	20.8300
Q1	11.6500	11.7400	11.8800	12.2000	12.3900	12.5200	12.8800	13.3300	13.9300	14.2300	15.5000	16.4200
median	11.9900	12.0600	12.3500	12.4300	13.1100	13.0500	13.3400	13.7900	14.7100	15.0100	16.1800	17.2500
Q3	12.5900	12.6700	12.8300	13.0500	13.5000	13.5100	14.0300	14.7200	15.6100	16.3200	17.4300	18.6100
IQR	0.9400	0.9300	0.9500	0.8500	1.1100	0.9900	1.1500	1.3900	1.6800	2.0900	1.9300	2.1900
mean	12.1490	12.1818	12.4057	12.7161	13.1512	13.2920	13.6404	14.0833	14.8402	15.2008	16.4967	17.4549
s.d.	0.7072	0.6081	0.6428	0.8804	1.0479	1.1786	1.1975	1.1979	1.2778	1.3817	1.4915	1.5398

Figure 19. Boxplot of the retail wage per hour in the USA states during 2012-2023.

3.4.2. Time-Wise Autoregressive Scatter Plot

When dealing with variables that change over time, you can generate a scatter plot comparing the values of the same variable at two different times. The slope of the linear fit in this plot represents the serial autoregressive coefficient, indicating the relationship between the variable's value in one period and its value in the preceding period. In more technical terms, this slope is the coefficient in a linear regression of $\boldsymbol{y_t}$ on $\boldsymbol{y_{t-1}}$ (Figure 20).

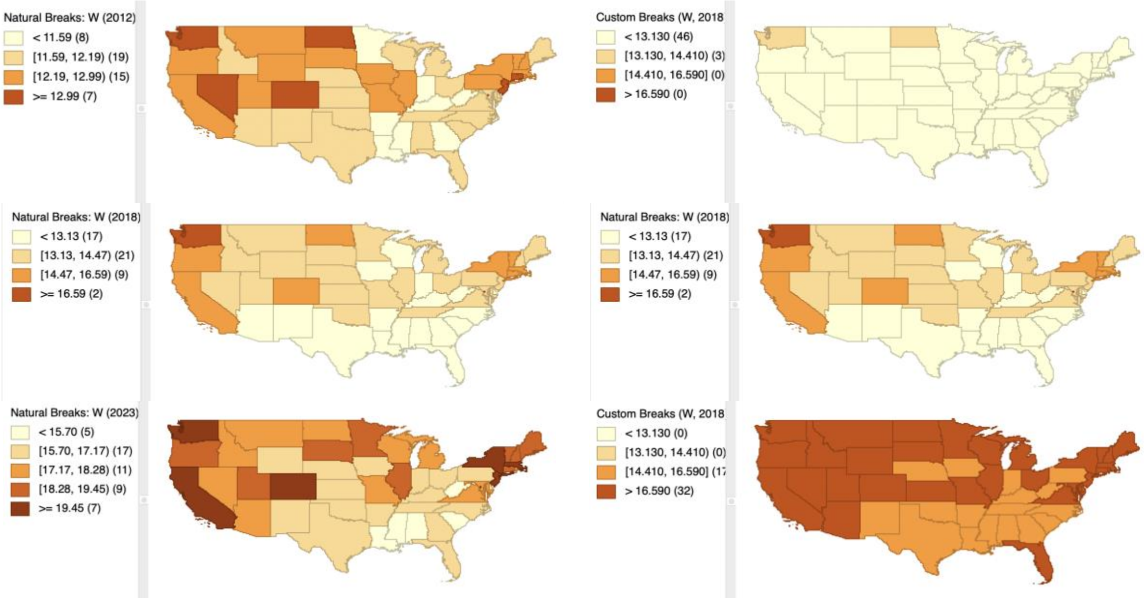


(a) $\boldsymbol{W_{2018}}$ on $\boldsymbol{w_{t-2012}}$ (b) $\boldsymbol{W_{2023}}$ on $\boldsymbol{W_{2018}}$

Figure 20. Scatter plot with time-lagged retail wage per hour in the USA states.

3.4.3. Space-Time Choropleth Map

A thematic map shows the spatial distribution of a single variable. However, since any map classification is by design relative to each point in time to which the variable refers, this may lead to an unsatisfactory outcome. In order to better represent the absolute changes considered in different moments of time, a set of custom categories must be created taking, for example, a time period as a reference. In Figure 21, the 2018 period categories have been taken as a reference to build the categories for all thematic maps. This allows us to see not only the spatial evolution of the variable, but a somewhat "deflated" temporal evolution.



(a) Relative changes (b) Absolute changes, categories: 2018

Figure 21. Natural breaks map of the retail wage per hour, USA states, 2012, 2016, 2018, 2023.

3.4.4. Treatment Effect Analysis: Averages Chart

The true strength of the Averages Chart stems from its ability to compare how the distribution of a variable differs between two subsets of data over time. This is especially useful in treatment effect analysis, where one set of observations undergoes a policy or treatment from a start to an end point. A target variable, or outcome, is measured for both time periods—before and after the implementation of the policy. The goal is to determine if the treatment significantly impacted the target variable by comparing its temporal change against that of a control group.

In the example used, the target variable is $\boldsymbol{W}$, retail wages per hour of less skilled workers in the United States of America (USA). It considers a policy experiment, where from 2000 to 2016 approximately, Amazon applied a targeted set of policies aimed at diversifying its business and gaining market share. In this way, the company experienced its peak expansion in the country with the launch of products such as Amazon Prime, Amazon Fresh, Amazon Music and Amazon Kindle, etc., specifically in certain states where it has a strong presence (more than 15,000 employees in 2021), which will be the *treated* group. The treated states are depicted in a unique value map for AMZ, with their selection highlighted in Figure 22.

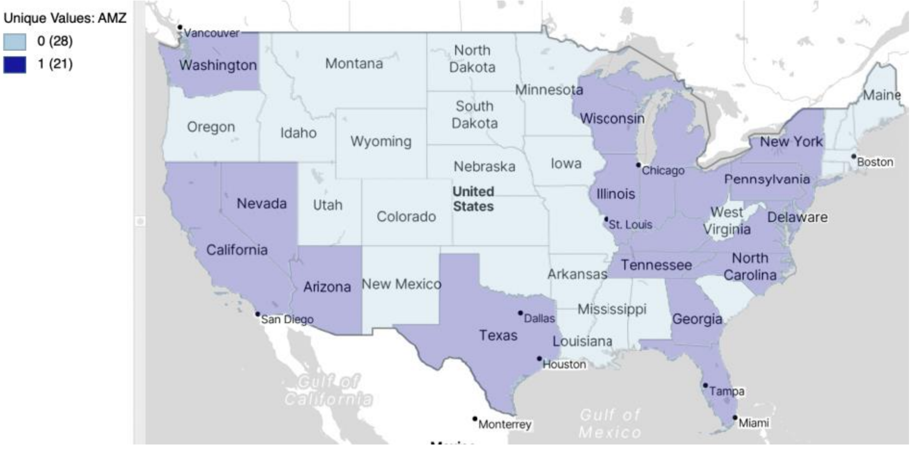


Figure 22. Unique map of the USA states with more than 15,000 employees in Amazon (2021).

The Averages Chart in GeoDa allows applying two methods: the Difference-in-Means test applied statically to the variable at different points in time and a more complete treatment effect analysis, implemented through the Difference-in-Difference test.

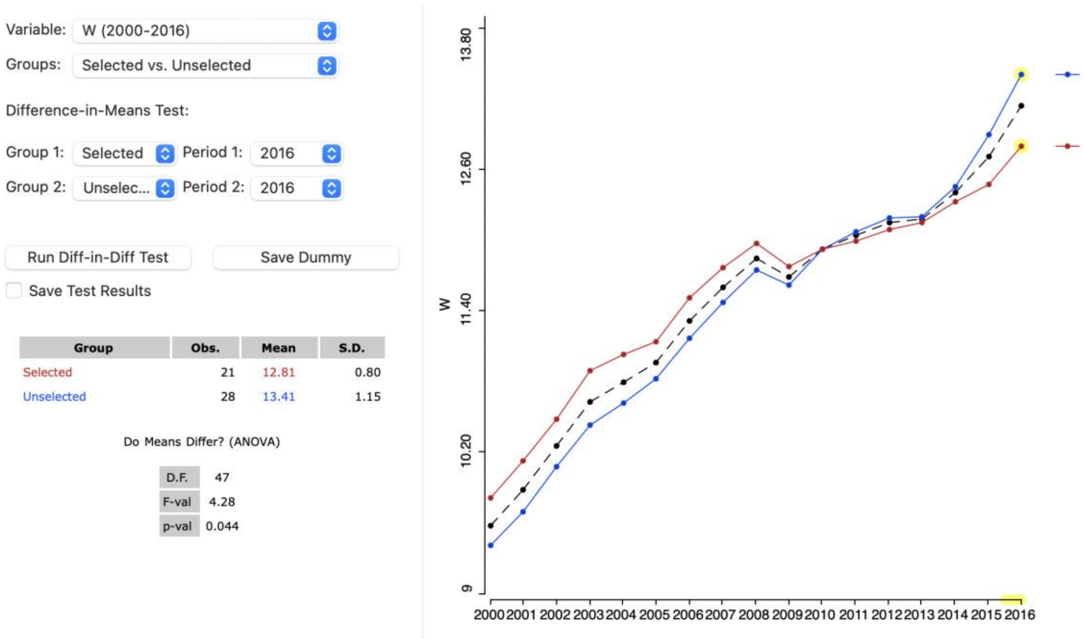
1) Difference-in-Means test

The primary function of this chart is to depict and quantify the difference in the mean of a variable between selected observations and their complement, the unselected observations. In GeoDa, this analysis is not conducted using a conventional t-test of difference in means[13].

[13] A t-test is a statistical method utilized to compare the means of two groups. It is employed in hypothesis testing to assess whether a specific treatment influences the target population, or to determine if there is a significant difference between two groups. $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s^2(1/n_1 - 1/n_2)}$, where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups being compared, $s^2 = s_x/\sqrt{n}$ is the pooled standard error of the two groups, (for s_x the pooled standard deviation and n the total sample), and n_1, n_2 denote the number of observations in each group. A higher t-value indicates that the

However, GeoDa use an *F*-statistic[14], which is derived from a regression that incorporates an indicator variable for the selection (assigned a value of 1 for selected observations and 0 for others). The *F*-statistic, which assesses the significance of the joint slopes in this regression, essentially serves the same purpose as a *t*-test on the coefficient of the indicator variable, given that there is only one slope involved.

To carry out a simple static difference in means test, we must specify the study variable (in our example, retail wage per hour, $\boldsymbol{W}$), the treated group (e.g., selecting the states on the unique values map) and the period for which the average test analysis must be performed (2016, in this case). Figure 23 shows the average chart of retail wages per hour for the group of selected (*treated*) states with more than 15,000 employees of Amazon in 2021, the group of the rest of states (*untreated*) and the overall group, for the year 2016.



Variable: W (2000-2016)
Groups: Selected vs. Unselected
Difference-in-Means Test:
Group 1: Selected Period 1: 2016
Group 2: Unselec... Period 2: 2016

□ Save Test Results

Group	Obs.	Mean	S.D.
Selected	21	12.81	0.80
Unselected	28	13.41	1.15

Do Means Differ? (ANOVA)

D.F.	47
F-val	4.28
p-val	0.044

Figure 23. Average chart of retail wages/hour for *treated* and *untreated* USA States, year 2016.

The *F* value of 4.28 allows to reject the null hypothesis of equality of means between groups with approximately 96% confidence. Therefore, it can be concluded that, in 2016, the average retail wage per hour in the selected or *treated* USA states is significantly lower than the average retail wage per hour in the *untreated* states while in the baseline year (2000), it was higher.

2) Difference-in-Difference test

The essence of treatment effect analysis lies in comparing how the target variable evolves in the treated group before and after policy implementation against a hypothetical scenario, termed the

difference between the group means exceeds the pooled standard error, suggesting more difference between the groups. (Scribbr 2025).

[14] The *F*-statistic fundamentally evaluates whether **the inclusion of an explanatory variable** yields a statistically significant **improvement in model fit relative to a model containing only the intercept**, or the overall mean. The computation of this statistic involves the regression sum of squared residuals ($RSS = \sum_{i=1}^{n} e_i^2$) and the total sum of squared deviations of the dependent variable from its mean ($RSY = \sum_{i=1}^{n}(y - \bar{y})^2$). The resulting *F*-statistic in the context of the simple regression, with a constant and a binary (dummy) variable, is given by: $F = (RSS - RSY/k - 1)/(RSY/n - k)$, where $k = 2$, which implies that the F-statistic is evaluated with degrees of freedom equal to 1 and $n - 2$. See Anselin and Rey (2014), pp. 98-99.

counterfactual, which speculates on the variable's behavior when the treatment is not been applied. This approach extends beyond a simple pre- and post-treatment mean comparison by also considering the evolution of the target variable within a control group, forming the foundation for a difference-in-differences (DiD) analysis (Angrist and Pischke 2015) The challenge here is that the counterfactual is not directly observable and must be inferred from the control group's outcomes.

A pivotal assumption in this analysis is that, excluding the intervention, the target variable would follow similar trends over time in both the treatment and control groups. This implies that any disparity observed between the groups in the first period would persist into the second period in the absence of the treatment.

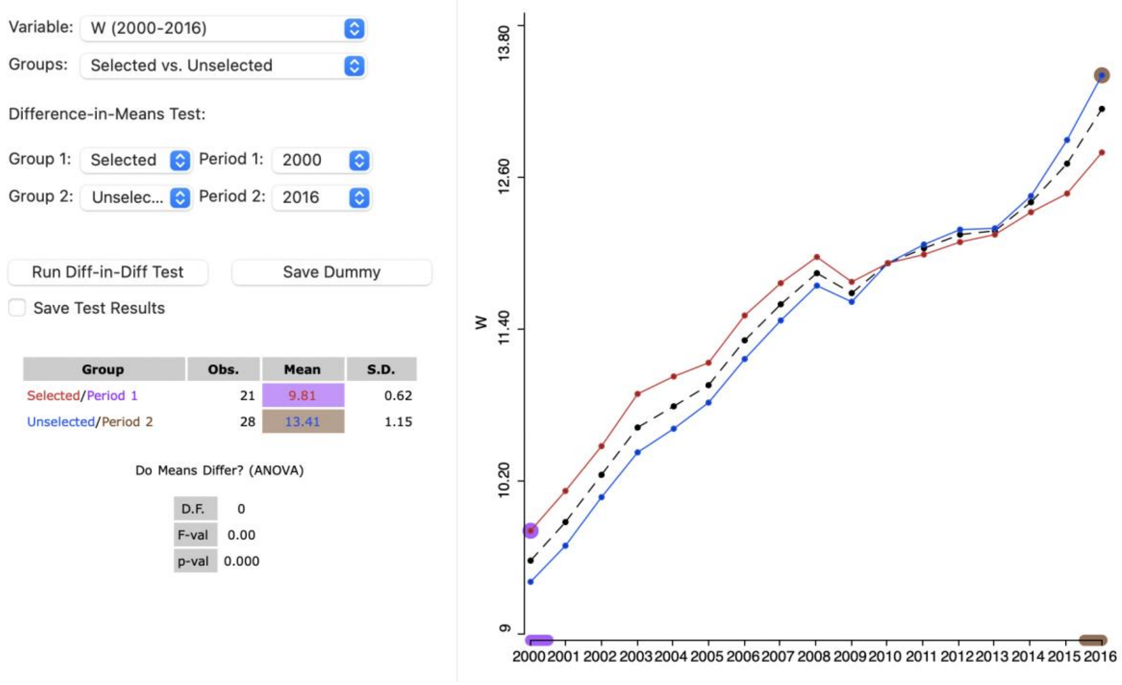
Therefore, the counterfactual is effectively an extrapolation of the treated group's trend. Without any treatment impact, the value of the target variable in period 2 would ideally equal the initial difference between the treated and control groups in period 1, plus the change observed in the control group from period 1 to period 2—this constitutes the trend extrapolation. The difference between this extrapolated counterfactual value and the actual observed value of the target variable in period 2 serves as the estimated treatment effect.

Formally, this methodology can be quantified through a linear regression model that stacks the target variable data across both periods and incorporates several dummy variables: a space dummy ($\boldsymbol{S}$, set to 1 for the treated group), a time dummy ($\boldsymbol{T}$, set to 1 for period 2), and an interaction dummy that captures the effect of the treatment in the second period ($\boldsymbol{S} \times \boldsymbol{T}$):

$$y_t = \beta_1 + \beta_2 S + \beta_3 T + \beta_4 (S \times T) + u_i \qquad (11)$$

The coefficient where $\boldsymbol{\beta}$ represents the estimated coefficients and $\boldsymbol{u}$ denotes the random error term. The coefficient $\boldsymbol{\beta_4}$ measures the treatment effect. The coefficient $\boldsymbol{\beta_1}$ indicates the average value for the control group in the baseline year. The coefficient $\boldsymbol{\beta_2}$ reflects the pure spatial effect, representing the difference in means between the treated and untreated in the baseline year. Lastly, $\boldsymbol{\beta_3}$ captures the time trend within the control group.

The implementation of the difference-in-differences analysis in the interface of GeoDa software utilizes the familiar Averages Chart interface, albeit the time frames set to period 1 in 2000 and period 2 in 2016 (Figure 24). Although the means of the two groups are displayed and the corresponding graph appears in the right-hand panel, the outcomes of the difference in means test are recorded as zero, because the analysis is carried out by selecting the "Run Diff-in-Diff Test" button (Figure 25).



Variable: W (2000-2016)

Groups: Selected vs. Unselected

Difference-in-Means Test:

Group 1: Selected Period 1: 2000

Group 2: Unselec... Period 2: 2016

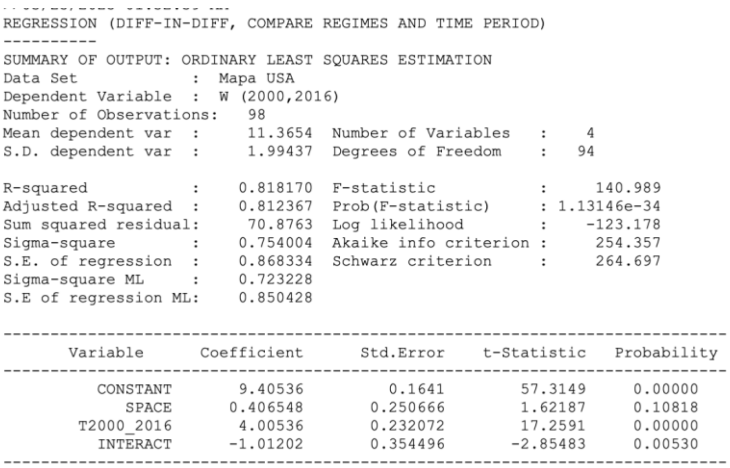
□ Save Test Results

Group	Obs.	Mean	S.D.
Selected/Period 1	21	9.81	0.62
Unselected/Period 2	28	13.41	1.15

Do Means Differ? (ANOVA)

D.F.	0
F-val	0.00
p-val	0.000

Figure 24. Difference in Difference setup, retail wage per hour in USA states, 2000 to 2016.



```
REGRESSION (DIFF-IN-DIFF, COMPARE REGIMES AND TIME PERIOD)
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data Set             :  Mapa USA
Dependent Variable   :  W (2000,2016)
Number of Observations:   98
Mean dependent var   :     11.3654  Number of Variables    :    4
S.D. dependent var   :     1.99437  Degrees of Freedom     :   94

R-squared            :    0.818170  F-statistic            :     140.989
Adjusted R-squared   :    0.812367  Prob(F-statistic)      : 1.13146e-34
Sum squared residual:      70.8763  Log likelihood         :    -123.178
Sigma-square         :    0.754004  Akaike info criterion :     254.357
S.E. of regression   :    0.868334  Schwarz criterion      :     264.697
Sigma-square ML      :    0.723228
S.E of regression ML:     0.850428

-----------------------------------------------------------------------------
       Variable       Coefficient       Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT        9.40536        0.1641        57.3149       0.00000
             SPACE       0.406548       0.250666       1.62187       0.10818
        T2000_2016        4.00536       0.232072       17.2591       0.00000
          INTERACT       -1.01202       0.354496      -2.85483       0.00530
-----------------------------------------------------------------------------
```

Figure 25. Difference in Difference results, retail wage per hour in USA states, 2000 to 2016.

In the regression output, the overall fit should be good, as it is the case in the example., with an $\boldsymbol{R^2} = \mathbf{0.82}$, and so should the coefficients, especially the corresponding to the treatment effect.

In this case, all the coefficients are highly significant at more than 99% except the estimator of the "pure space" variable, which is significant at 90% approximately. The treatment effect, which is quantified by the INTERACT coefficient, is –1.01 and highly statistically significant, indicating that Amazon's expansionist policies likely had a notable negative impact on low-skilled retail wage per hour.

Additional details are provided by the other coefficients: The CONSTANT, valued at 9.40, corresponds to the average for the *untreated* group in 2000. The SPACE coefficient, at 0.41, reflects the mean difference between the *treated* and *untreated* groups in 2000. The trend coefficient (T2000_2020) stands at 4.01. This is consistent with the changes in the means for the *untreated* group in period 1 (9.40) and in period 2 (13.41), with the difference, 4.01, aligning exactly with the time trend coefficient, T2000_2016.

4. Conclusions

In the conclusion of the manuscript, it is emphasized that Exploratory Spatial Data Analysis (ESDA) extends beyond traditional statistical methodologies to unveil hidden spatial patterns and relationships that may otherwise remain obscured. This exploration is crucial for a comprehensive understanding of spatial data, integrating geographical information systems (GIS) to enhance the complexity and detail of analyses. The paper advocates strongly for the integration of spatial context in statistical analyses, urging that both physical and socio-cultural dimensions of space be considered to understand better the dynamics that influence patterns and phenomena affecting human behaviors and environmental conditions.

ESDA methodologies, tailored to handle continuous, discrete, and spatial-temporal data, are discussed with an emphasis on their pivotal role in developing sophisticated econometric models. Techniques such as spatial visualization and dynamic linking are presented as effective tools for identifying spatial dependencies and heterogeneities across various datasets. The manuscript illustrates the versatility and critical insights provided by ESDA through its application in multiple disciplines, which substantially aid in informed decision-making across fields as diverse as urban planning and environmental science.

Moreover, the paper highlights the importance of exploratory analysis before the estimation of econometric models to ensure that variables are appropriately analyzed and understood. This approach helps in identifying outliers and understanding distribution patterns which are essential for accurate model specification. By fostering an understanding of spatial data's complexity and its analytical challenges, ESDA stands out as a powerful framework for exploring the intricate relationships inherent in spatial data. The conclusions drawn in this manuscript underline the ongoing relevance and necessity of developing robust ESDA techniques to keep pace with the increasing complexity of spatial datasets and the sophisticated models needed to analyze them.

Appendix A. Mapping Smoothed Rates (See Anselin 2023)

1) **Variance instability of rates**

Rate smoothing begins by understanding the number of events at a location i, O_i, as the outcomes of P_i draws from a distribution governed by a risk parameter π_i. This is analogous to drawing n balls from an urn containing red and white balls, with p representing the proportion of red balls. Given n draws, the interest lies in determining the likelihood of obtaining q red balls.

The appropriate mathematical model for this scenario is the **binomial distribution**, $B(n, p)$. The probability function of this distribution calculates the likelihood of k successes in n sequential draws from a population with a risk parameter p:

$$Prob[k] = \binom{n}{k} p^k (1-p)^{n-k} \tag{A1}$$

where $\binom{n}{k} = n!/k!\,(n-k)!$ is the binomial coefficient or the number of combinations of k observations out of n.

Translated into the context of O_i events out of a population of P_i, the corresponding probability is:

$$Prob[O_i] = \binom{P_i}{O_i} {\pi_i}^{O_i} (1-\pi_i)^{P_i - O_i} \tag{A2}$$

with π_i as the underlying risk parameter. The mean of this binomial distribution is $\pi_i P_i$, and the variance is $\pi_i(1-\pi_i)P_i$.

Returning to the raw rate $r_i = O_i/P_i$, it can be readily seen that its mean corresponds to the **underlying risk**:

$$E[r_i] = E\left[\frac{O_i}{P_i}\right] = \frac{E[O_i]}{P_i} = \frac{\pi_i P_i}{P_i} = \pi_i \tag{A3}$$

Note that only O_i is random; P_i is not.

Therefore, the **raw rate (r_i) is an unbiased estimator of the underlying risk (π_i)**. However, its variance exhibits certain undesirable characteristics, which can be derived algebraically as follows:

$$Var[r_i] = \frac{\pi_i(1-\pi_i)P_i}{P_i^2} = \frac{\pi_i(1-\pi_i)}{P_i} \tag{A4}$$

This outcome indicates that **the variance is dependent on the mean**, an unusual situation that introduces additional complexity. Crucially, it reveals that larger populations in an area (denoted as P_i in the denominator) result in lower variance for the estimator, enhancing its precision. Conversely, in areas with smaller populations (P_i), the risk estimate becomes less precise (exhibiting higher variance).

Given that population sizes typically vary across regions, the precision of each rate also varies. This instability in variance must be accounted for in map representations or corrected to prevent misleading depictions of the spatial distribution of the underlying risk. This need for adjustment underpins the subsequent discussion on rate smoothing.

2) **Borrowing strength**

Methods for smoothing rates, also known as **shrinkage estimators**, **enhance the accuracy of crude rates** by leveraging information from **additional observations**. This concept stems from the pioneering work of James and Stein, who introduced what is known as the **James-Stein paradox**, demonstrating that sometimes **biased estimators can offer greater precision in terms of mean squared error** (MSE) (James and Stein 1961).

The MSE of an estimator $\hat{\theta}$ of an unknown parameter θ is defined as the **sum of the variance and the squared bias** (Wikipedia 2025b):

$$MSE(\hat{\theta}) = E_{\theta}\left[(\hat{\theta}-\theta)^2\right] = Var_{\theta}(\hat{\theta}) - Bias(\hat{\theta},\theta)^2 \tag{A5}$$

For unbiased estimators, the bias component is zero making the MSE of the estimator equivalent to its variance.

The principle of borrowing strength involves a **deliberate trade-off**: accepting a slight increase in bias to achieve a significant reduction in the variance part of the MSE. Although this results in a biased estimator, it **tends to be more precise in terms of MSE**, effectively reducing the likelihood of deviating significantly from the true parameter value. This approach is grounded in **Bayesian statistical principles**, which will be briefly explained subsequently.

3) **Bayes Law**

The theoretical basis for rate smoothing lies within a Bayesian framework, where the distribution of a random variable is revised upon observing data. This process is guided by Bayes Law, which emerges from the **breakdown of the joint probability (or density) of two events, A and B,** into their respective conditional probabilities:

$$P[AB] = P(A|B) \times P(B) = P(B|A) \times P(A) \tag{A6}$$

where A and B are random events, and ∣ stands for the conditional probability of one event, given a value for the other. The second equality formally represents Bayes Law as:

$$P(A|B) = \frac{P((B|A) \times P(A)}{P(B)} \tag{A7}$$

In practical applications, **the denominator in this expression is often disregarded**, and the equality is commonly replaced with a proportionality sign:[15]

$$P(A|B) \propto P(B|A) \times P(A) \tag{A8}$$

In the process of estimation and inference, the variable A usually represents a parameter or a set of parameters, while B represents the data. The aim is to update the prior knowledge of parameter A, which is expressed as the prior distribution $P(A)$, after observing the data B, resulting in a posterior distribution $P(A|B)$ — what is now understood about the parameter following the data observation. **This connection between the prior and posterior distributions is made through the likelihood**, $P(B|A)$. Using conventional notation where θ symbolizes the parameters[16] and X the observations, the relationship can be articulated as (Wikipedia 2025c):

$$P(\theta|X) \propto P(X|\theta) \times P(\theta) \quad \text{(A9}$$

For each estimation issue, it is necessary to **define distributions for both the prior and the likelihood**, ensuring they combine to produce a valid posterior distribution. Conjugate priors are especially useful because they facilitate a closed-form solution for this combination. Common priors in the context of rate smoothing include the **Gamma and Gaussian** (normal) distributions. While a detailed mathematical discussion is outside our current scope, the basic idea of smoothing involves **adjusting the raw data**-derived estimate (i.e., the crude rate) **with prior information, such as a reference rate from a broader region**. This approach corrects unreliable estimates from small areas, where, for instance, a **zero-event occurrence does not necessarily indicate zero risk**, by integrating broader reference data.

[15] In a Bayesian framework, the **likelihood** is defined as the **probability of observing the data given** a specific value (or distribution) of the **parameters**. Conversely, classical statistics typically consider the probability of the parameters based on the observed data.

[16] Note that in this case, θ represents a set of parameters and not the scale parameter of the Gamma distribution.

doi:10.20944/preprints202503.2044.v1

References

Angrist, J, Pischke, J-S (2015) *Mastering Metrics, the Path from Cause to Effect*. Princeton, New Jersey: Princeton University Press.

Anselin, L. (2023) *An Introduction to Spatial Data Science with GeoDa*. Volume 1: Exploring Spatial Data. https://lanselin.github.io/introbook_vol1

Anselin, L, Rey, S.J. (2014) *Modern Spatial Econometrics in Practice, a Guide to Geoda, Geodaspace and Pysal*. Chicago, IL: GeoDa Press.

Anselin, L, Lozano-Gracia, N, Koschinky J (2006) *Rate Transformations and Smoothing*. Technical Report. Urbana, IL: Spatial Analysis Laboratory, Department of Geography, University of Illinois.

Bivand, R.S. (2010) Exploratory Spatial Data Analysis. In Fischer, M.M., Getis, A., Eds.; *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*; Springer: Berlin, Heidelberg; pp. 219–254 ISBN 978-3-642-03647-7.

Brewer, C.A. (1997) Spectral Schemes: Controversial Color Use on Maps. *Cartography and Geographic Information Systems* 49: 280–94.

Chasco, C., Vallone, A. (2023). *Introduction to Cross-Section Spatial Econometric Models with Applications in R*. Preprints. https://doi.org/10.20944/preprints202309.0413.v1

Inselberg, A. (1985) The Plane with Parallel Coordinates. *Visual Computer* 1: 69–91.

James, W, Stein, C (1961) *Estimation with Quadratic Loss*. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1: 361–79.

Jenks, G. F. (1977) *Optimal Data Classification for Choropleth Maps*. Occasional. Paper no. 2. Lawrence, KS: Department of Geography, University of Kansas.

Lefebvre, H. (1992) *The Production of Space*; Wiley-Blackwell: Oxford (UK) & Cambridge (USA), 1992; ISBN 978-0-631-18177-4.

Loader, C. (2004) Smoothing: Local Regression Techniques. In Gentle, J.E., Härdle, W., Mori, Y. (eds.) *Handbook of Computational Statistics: Concepts and Methods*, Berlin: Springer-Verlag, pp. 539–63.

Scribbr (2025) *An Introduction to t Tests | Definitions, Formula and Examples*. Accessed in March 2025. https://tinyurl.com/mud65hf9

Tobler, W.R. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234–240, https://doi.org/10.2307/143141

Tukey, J. (1977) *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

Wikipedia (2025a) *Gamma distribution*. Accessed in March 2025. https://tinyurl.com/48yy28bc

Wikipedia (2025b) *Bar chart*. Accessed in March 2025. https://tinyurl.com/2dtws4zn

Wikipedia (2025c) *Bayesian inference*. Accessed in March 2025. https://tinyurl.com/mrxrvyxr

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.