Introduction to Cross-Section Spatial Econometric Models with Applications in R

Coro Chasco [*] and Andrés Vallone

Article

Introduction to Cross-Section Spatial Econometric Models with Applications in R

Coro Chasco [1,2,*] and Andrés Vallone [3]

[1] Department of Applied Economics, Universidad Autónoma de Madrid, C/Francisco Tomás y Valiente, 5, 28049 Madrid, Spain; coro.chasco@uam.es
[2] ECEMIN Research Group, Nebrija University, C/Sta. Cruz de Marcenado, 27, 28015 Madrid, Spain
[3] Escuela de Ciencias Empresariales-Instituto de Políticas Públicas, Universidad Católica del Norte, Larrondo 1281, Coquimbo, Chile; avallone@ucn.cl
* Correspondence: coro.chasco@uam.es

Abstract: This paper introduces the spatial component in cross-section econometric estimations and specifically, the spatial dependence effect inherent in some of the variables involved in the modelling process. First, the spatial structure of the data from thematic maps is observed and Moran's spatial autocorrelation indicators are presented. Subsequently, the spatial weights matrix is built under different specifications. Finally, several modelling specification strategies are shown and the interpretation of the estimated coefficients. The theoretical concepts are illustrated with examples and their corresponding R software codes. This code and databases are available in a freely accessible repository in the BE2SHARE-EUDAT platform so that they can be easily reproduced.

Keywords: Spatial Econometrics; Spatial Weight Matrix; Spatial Autocorrelation; Cross-Section Models; R Software

1. Introduction

As first affirmed by the French philosopher and sociologist Henri Lefebvre in his work "*The Production of Space*", published in 1974 and translated into English in 1991: **"*space matters*"** ([1]). 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. Space can play an important role in determining the processes to be modelled. Hence it is relevant to find a way to incorporate this phenomenon into the estimation of spatial processes.

In fact, social phenomena are often not independent of the geographical space in which they occur. For example, in cities one can find certain neighbourhoods with high-income households surrounded by neighbourhoods with high-income households, and vice versa: spatial concentrations of neighbourhoods 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*" ([2], p 236).

Spatial data tend to be positively autocorrelated over geographic space so that the degree of autocorrelation usually decreases with physical distance. Therefore, observations of spatial variables are usually not independent of each other. In the context of linear regression models, this situation implies a violation of the basic assumption of independence of observations required by the Ordinary Least Squares (OLS) estimation method. In other words, in this context of spatial dependence, OLS estimators are no longer optimal and, thus, the significance statistics (t and F) could lead to erroneous conclusions. For this reason, it is necessary to explicitly consider geographical space in spatial econometric regression models. In this paper, this issue will be addressed. First, some Exploratory Spatial Data Analysis (ESDA) methods are presented to identify spatial structures and spatial autocorrelation effects in the variables, and then the main spatial regression models and their corresponding estimation and testing methods will be shown.

2. Exploratory Spatial Data Analysis (ESDA)

Exploratory spatial data analysis (ESDA) was developed from *a*-spatial exploratory data analysis (EDA). It is widely used in spatial statistics, spatial econometrics and geostatistics (see Bivand [3]). ESDA is a set of techniques and methods used to analyse 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.

2.1. 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 variable using symbols and colours to show differences in values in different units or regions. When the regions are coloured 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. For example, quantile maps represent the overall spatial trend of a variable by dividing and grouping the data into categories with the same number of observations (e.g., quartiles, quintiles, deciles, etc.). In this section of the paper we will employ the dataset of in Guerry's study [4], also used in [5].[1] These data can also be found along with the other databases used in this paper in Chasco and Vallone [6].

Figure 1 shows the **histogram** and the **quintile map** of the variable "*clergy*" representing the rate of Catholic priests per active service population in the provinces of France.

Table 1. R Code: ESDA of the rate of Catholic priests per active service population (Figure 1).

```r
load("guerry.rdata")
filled_color<-cut(guerry$clergy,breaks=5,include.lowest=TRUE)
breaks<-5
histogram_clergy<-ggplot(guerry,aes(x=clergy,fill=filled_color)) +
  geom_histogram(binwidth=(max(guerry$clergy)-min(guerry$clergy))/breaks,
                boundary=min(guerry$clergy)) +
  scale_fill_brewer(palette="PuRd", name="Priests") +
  theme_bw() + theme(legend.position='bottom') +
  xlab("Priests per active service population") + ylab("")
filled_color2<-cut(guerry$clergy,breaks=quantile(guerry$clergy,probs=seq(0,1,0.2
0)),
                                        include.lowest=TRUE)
map_clergy <- ggplot(guerry,aes(fill=filled_color2))+geom_sf() +
  scale_fill_brewer(palette="PuRd", name="Priests")+annotation_scale(location="t
l")+
  annotation_north_arrow(location = "bl", which_north="true") +
  theme_map() + theme(legend.position = "right")
grid.arrange(histogram_clergy,map_clergy,ncol=2)
```

---

[1] A full description of the database is available at https://geodacenter.github.io/data-and-lab/Guerry

The histogram creates intervals of the variable with **equal length** (in this case, 5 intervals of size 16.8% with approximately 17 observations each), starting from the minimum value of the variable (2%). It shows a practically uniform distribution of this variable. The quintile map groups the data into 5 categories with the **same number of observations** (17 in all cases), almost identically to the histogram. However, the observation of the map allows detecting that spatial units with similar values are **clustered in space**, uncovering the existence of positive spatial autocorrelation.

Table 1 presents the R code needed to generate the results of this Figure. The following five libraries[2] are required to run this code: "ggplot2", "ggspatial", "ape", "ggthemes", *and* "gridExtra". The main functions involved in this R code are "*load*", "*cut*", "*ggplot*", "*geom_histogram*", "*annotation_scale*", "*annotation_north_arrow*" and "*grid.arrange*". **load {base}** reload saved datasets (in this case, *"guerry.rdata")*. **cut {base}** divides the range of a numeric variable into intervals (factor variable) coding its values according to which interval they fall. **ggplot** is a function of **"ggplot2"**, which is a popular data visualization package in R programming. It stands for "Grammar of Graphics plot" and provides a consistent and structured way to create visualizations. **geom_histogram {ggplot2}** visualises the distribution of a single continuous variable by dividing the X axis into bins, counting the number of observations in each bin, and displaying the counts with bars. **annotation_scale {ggspatial}** allows creating the scale bar, which is a graphical means of depicting distance on a map. **annotation_north_arrow {ggspatial}** allows creating the north arrow on a map, which is a graphical representation that points to the geographical north (the option "true" points specifically to the North Pole). **grid.arrange {gridExtra}** allows arranging multiple graphical objects ("*grobs*") on a page.
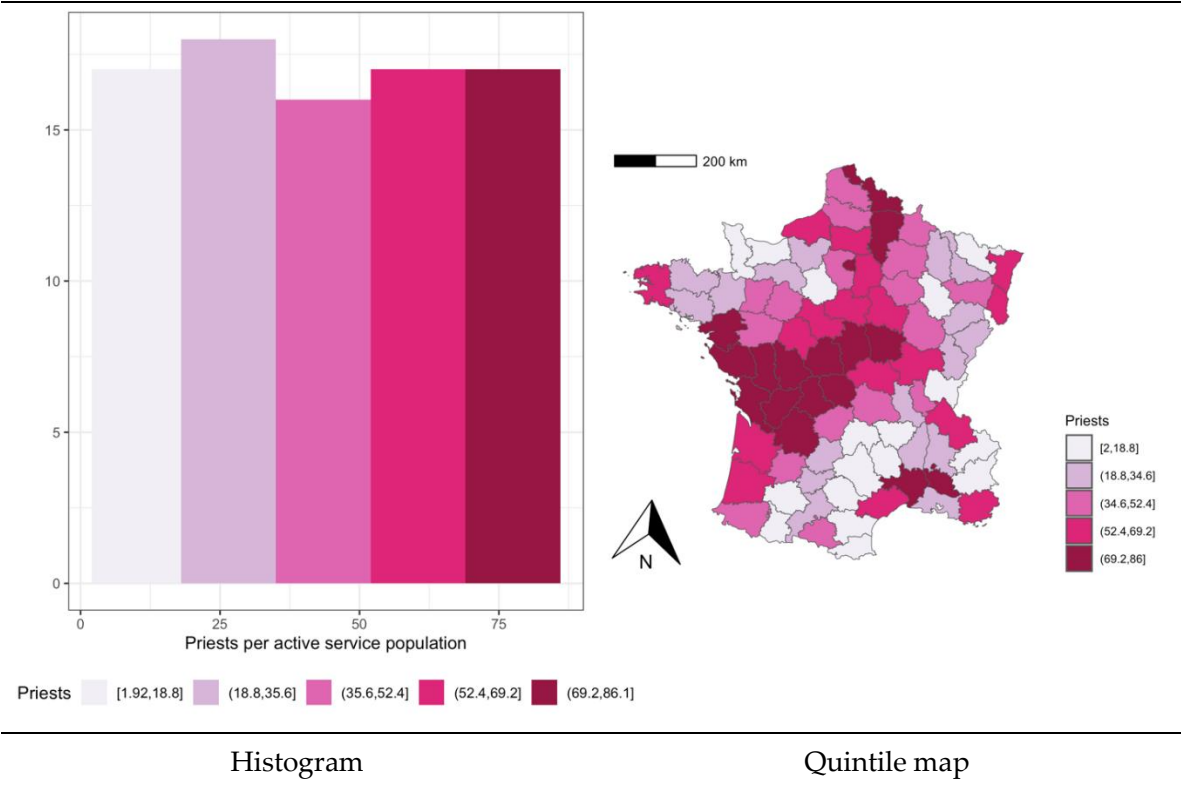

| Histogram | Quintile map |

Figure 1. ESDA of the rate of Catholic priests per active service population in France's provinces.

However, quintile maps are not useful when a determined variable has a **non-normal** distribution, as is the case for the variable "*suicids*" presenting the population per suicide in the provinces of France. Thus, when a variable is highly skewed or asymmetric, there is usually a high proportion of very similar values in the centre or at one extreme of the distribution, with one or more
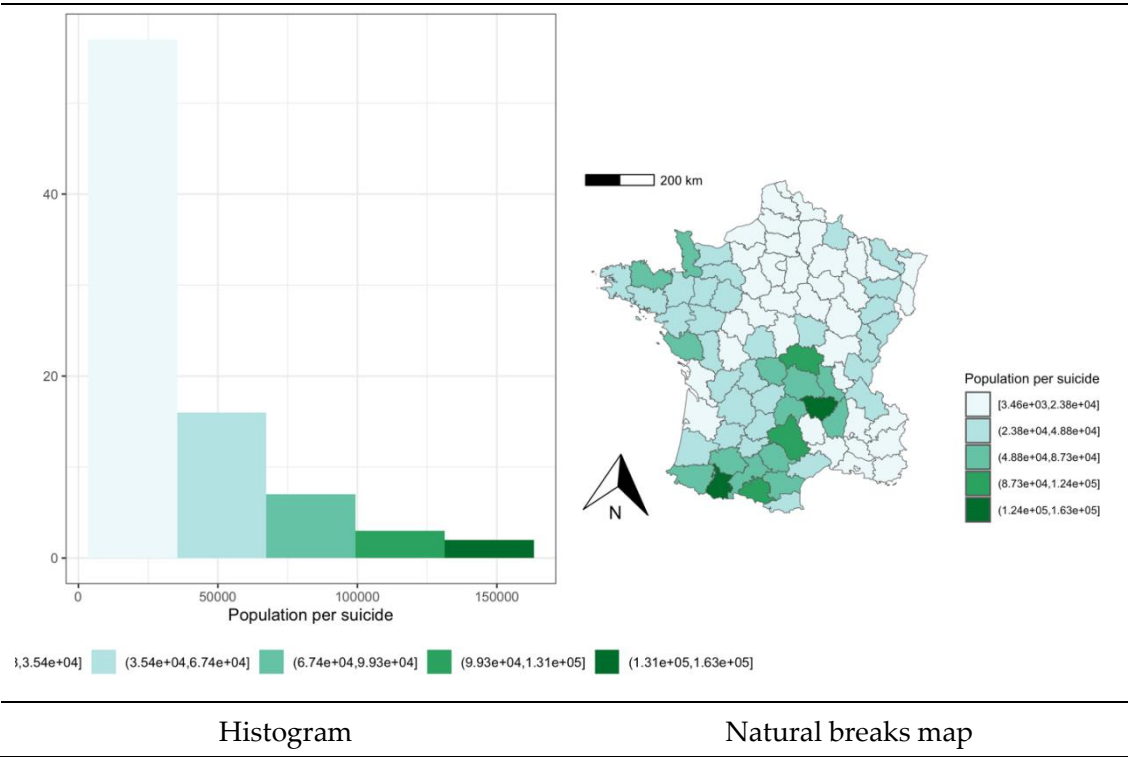
---

[2] From now on, only packages other than "base", which contains the R basic functions, will be shown.

values far away from the rest at the other extreme, as shown in the histogram in Figure 2. This prevents quintiles from being defined correctly by grouping very similar values into different intervals or very different values into a single interval. In these cases, it is better to use the choropleth map of **natural breaks** which is the one chosen to represent the variable "suicids" in Figure 2.

The R code is shown in Table 2.

Table 2. R Code: ESDA of the rate of suicides (Figure 2).

```
load("guerry.rdata")
filled_color<-cut(guerry$suicids,breaks=5,include.lowest=TRUE)
breaks<-5
histogram_suicids <- ggplot(guerry,aes(x=suicids,fill=filled_color))+
  geom_histogram(binwidth=(max(guerry$suicids)-min(guerry$suicids))/breaks,
                 boundary=min(guerry$suicids))+
  scale_fill_brewer(palette="BuGn", name="Suicids")+
  theme_bw() + theme(legend.position='bottom') +
  xlab("Population per suicide") + ylab("")
filled_color2<-cut(guerry$suicids,breaks=getJenksBreaks(guerry$suicids,k=6),
                 include.lowest=TRUE)
map_suicids <- ggplot(guerry,aes(fill=filled_color2))+
  geom_sf()+
  scale_fill_brewer(palette="BuGn", name="Population per suicide")+
  annotation_scale(location = "tl") +
  annotation_north_arrow(location = "bl", which_north="true") +
  theme_map() + theme(legend.position = "right")
grid.arrange(histogram_suicids,map_suicids,ncol=2)
```


| Histogram | Natural breaks map |

Figure 2. ESDA of the population per suicide in the French provinces.

The libraries and functions required for this code are the same as above. In addition, the function and corresponding library **getJenksBreaks {BAMMtools}** is added, which calculates the optimum breakpoints using Jenks natural breaks optimization. The values in a variable are binned into k+1 categories (5+1 in our case), according to the Jenks' classification method. This method is borrowed from the field of cartography, and seeks to minimize the variance within categories, while maximizing the variance between categories.

If a highly asymmetric distribution such as the suicide variable in the French provinces were represented by a quantile map, we could draw the wrong conclusions. In the quintile map in Figure 3, for example, the suicide variable obtains the highest values in the southern half and extreme west and east of France, leaving only a few northern provinces freer of this problem. However, a more appropriate representation of this variable with a map of natural breaks concentrates this phenomenon in certain provinces in the centre-south of the country.
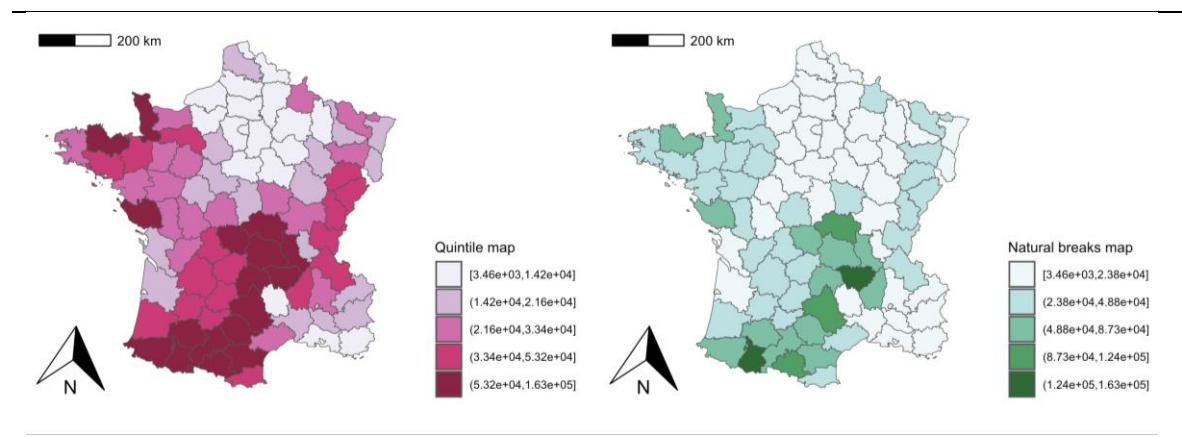

Figure 3. Comparison between the quintile and natural breaks maps of the population per suicide in the French provinces.

For this reason, it is important to carry out a preliminary exploratory analysis of the variables to analyse, even to know what type of map to use, because, as Monmonier [7] rightly explains, consciously or unconsciously, it is very easy "*to lie with maps*".

2.2. Spatial Autocorrelation Effect: The Spatial Weights Matrix

One of the most studied spatial effects so far is that of **spatial autocorrelation**, which is the one we will focus on in this section.[3] Spatial dependence[4] is said to exist in a variable when the value it takes in one observation *i* is influenced by the value it takes in another observation *j*, and *vice versa* (see [8]). The spatial autocorrelation effect can be positive, when spatial units with similar values tend to be together (high values close to high and low values close to low) or negative, when spatial units tend to be surrounded by neighbouring units with opposite values (high surrounded by low, and the other way around). The absence of spatial autocorrelation is known as spatial randomness. In Figure 4, various spatial patterns are shown for spatial autocorrelation and spatial randomness.

The code and R packages needed to produce this Figure can be found in Appendix A.1. The following five libraries are required to run this code: "ggplot2" *and* "gridExtra". The main functions involved in this R code are "*geom_rect*" and "*grid.arrange*". **geom_rect {ggplot2}** is a geometric object

---

[3] The other spatial effect is spatial heterogeneity, extensively presented at Anselin [8], chapter 9.

[4] Strictly speaking, the concepts of spatial dependence and spatial autocorrelation are not synonymous. Indeed, spatial autocorrelation is a form of spatial dependence defined statistically in a weaker form, through only the first moments of the joint distribution of a spatial variable. However, as most authors use both terms interchangeably this will be also done in this paper.

("*geom*") that draws a rectangle using the locations of the four corners ("*xmin*", "*xmax*", "*ymin*" and "*ymax*"). **grid.arrange {gridExtra}** is defined as above.
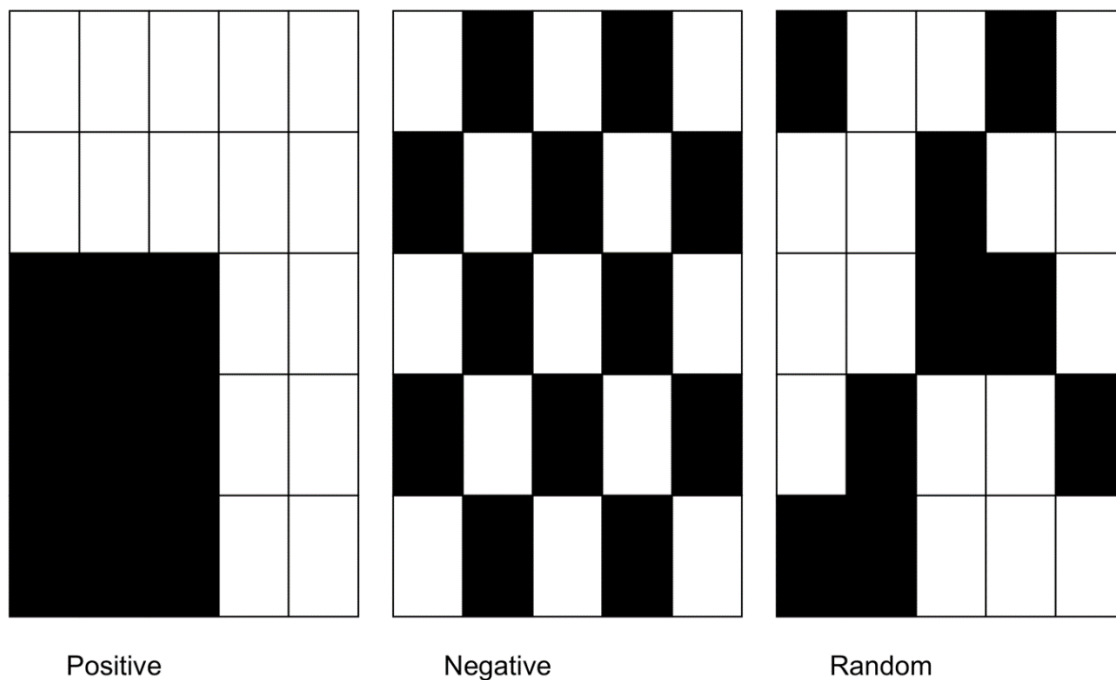

Figure 4. Spatial autocorrelation patterns.

The concept of autocorrelation alone refers to the modelling of a variable with itself. Thus, when referring to serial autocorrelation in econometric models, we state to the fact that the past values of a variable affect the present values of the same variable. $y_t = f(y_{t-1}, \dots, y_1)$. In this case, it is a one-way interaction of the variable: from the past to the present. However, the influence of geographical space is not unidirectional but multidirectional, in the environment or neighbourhood of a unit of analysis. It is therefore crucial to define the concept of spatial neighbourhood, in order to carry out any spatial analysis.

The spatial neighbourhood matrix, $\mathbf{W}_{n \times n}$, also called **spatial weights matrix**, shows the relationship among n spatial units under analysis. This matrix defines the spatial neighbourhood condition and thus the interaction between the spatial units. In its simplest form, the spatial weights matrix is a symmetric and binary matrix, where the element $w_{ij} = 1$ if spatial units i and j are neighbours and zero if they are not. By convention, it is established that $w_{ii} = 0$, i.e., a spatial unit cannot be a neighbour of itself. There are different criteria for defining this matrix depending on the process to be modelled and the characteristics of the data. But basically, we could identify the following two groups of spatial weighting matrices: spatial contiguity matrices and spatial weights matrices based on distances.

2.2.1. Spatial Contiguity Matrices

When working with a map in which the geographical units are polygons, one of the criteria presented in Figure 5 can be used to define the **W** matrix.

The **linear** criterion is the simplest. It considers as neighbours to a given spatial unit i all those units located to the north and south of this unit, provided they share a border in common with i. This criterion could also be applied if the neighbouring and contiguous units to i were located to the east and west sides.

The rest of the contiguity criteria follow the moves on a checkerboard to define spatial unit i's neighbourhood. The construction of a neighbourhood matrix under the **rook** criterion implies

considering as neighbours of spatial unit *i* those spatial units located in the four cardinal points (north, south, east, and west), provided that they share a common border with this unit. The use of the **bishop** criterion considers as neighbours of unit *i* those units located in the secondary or lateral cardinal points (north-east, north-west, south-east or south-west) of spatial unit *i*, provided that they have at least one point in common.
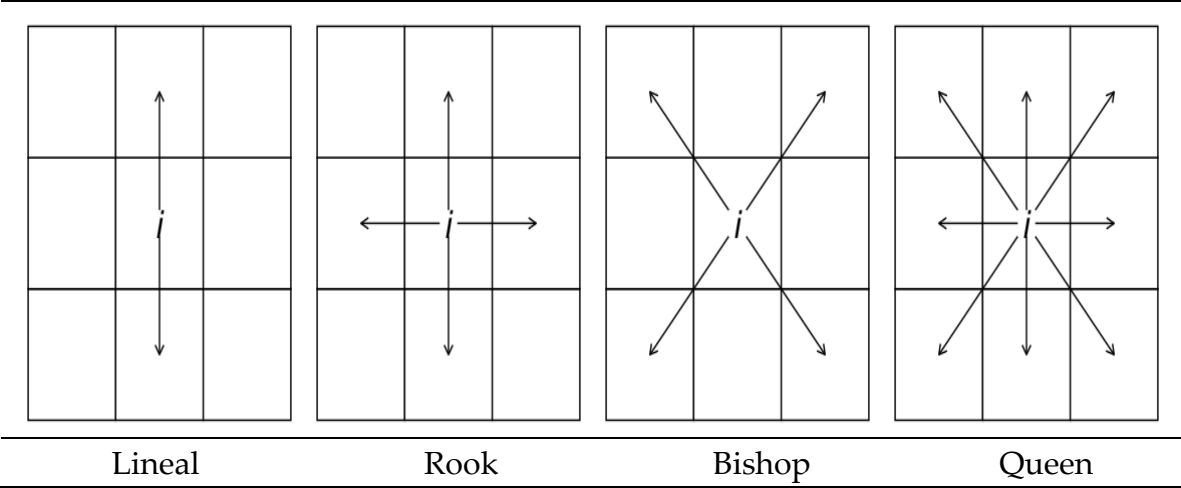

| Lineal | Rook | Bishop | Queen |

Figure 5. Spatial neighbourhood criteria in polygon data (Source: Martori et al. [9]).

The **queen** criterion is the most complete of all by considering as neighbourhood of spatial unit *i* those spatial units located in all main and secondary cardinal directions, provided that they have a point or a border in common with spatial unit *i* (Martori et al. [9]). Since the neighbourhood criterion in all these **W** matrices is based on the units or regions having a common border (edge or point), they are also called spatial contiguity matrices.

In certain contexts, such as spatial regression models, spatial weight matrices are **row standardised**, so that their elements are transformed as follows:

$$w_{ij}^* = \frac{w_{ij}}{\sum_{J=1}^n w_{ij}} \tag{1}$$

In simple words, each element of a row of the matrix **W*** is divided by the sum of the values of that row. This ensures that each element of the standardised spatial weights matrix lies between the values 0 and 1, and that the sum of the values in each row is always equals to 1.

The R code required for the construction of spatial contiguity matrices is presented in Table 3.

Table 3. R Code: construction of a queen contiguity matrix and its spatial network (Figure 6).

```r
load("guerry.rdata")
#Constructs the queen contiguity matrix
wqueen<-poly2nb(guerry, queen=TRUE)
#Constructs the row-standardized queen contiguity matrix
wqueens<-nb2listw(wqueen, style="W", zero.policy = TRUE)
# Plots first the map and then the neighbourhood network
plot(guerry$geometry)
plot(queens,st_centroid(guerry$geometry), col="blue", add = TRUE)
```

The "spdep" library is required to run this code (see Bivand and Wong [10]). It is a collection of functions to create spatial weights matrix objects from polygon 'contiguities', point patterns by distance and tessellations. It also loads the "spData" package (a group of spatial datasets for demonstrating, benchmarking, and teaching), which is underpinned by some legacy packages

("maptools", "rgdal" and "rgeos"). They will be retired during 2023, joint to all the "sp" classes, methods, and functions, and "sf", "terra" or other packages will be needed instead.

The main functions involved in this R code, besides "*load*", are "*poly2nb*", "*nb2listw*" and "*plot*". **poly2nb {spdep}** builds a neighbours list based on regions with contiguous boundaries, that are sharing one or more boundary points. **nb2listw {spdep}** adds a weights list with values given by a coding scheme style, the most commonly used being **B** (basic 0-1 binary coding) and **W** (row standardised).

plot is more than a function, it is placeholder for a family of related functions which allows to create a plot passing two vectors (of the same length), a dataframe, matrix, map, or even other objects, depending on its class or the input type. This function makes it possible to represent the spatial neighbourhood network of the W matrix corresponding to the provinces of France (Figure 6).
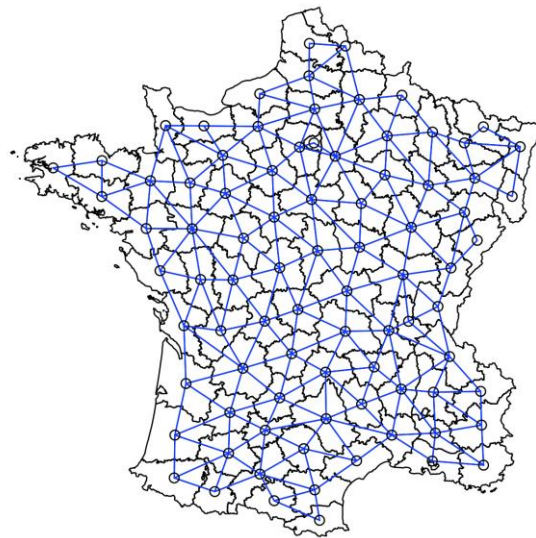

Figure 6. Neighbourhood network of a queen spatial contiguity matrix.

The construction of a rook-style spatial contiguity matrix requires the same packages and functions, but the *"queen"* parameter of the "*poly2nb*" function must be changed to "*FALSE* ", as in Table 4. When working with administrative regions, differences in the spatial network of the queen- and rook-based contiguity matrices are almost non-existent, as practically all regions have more than one point in common.

Table 4. R Code: construction of a rook spatial contiguity matrix and its spatial network.

```
load("guerry.rdata")
#Constructs the rook contiguity matrix
wrook<-poly2nb(guerry, queen = FALSE)
#Constructs the row-standardized rook contiguity matrices
wrooks<-nb2listw(wrook, style="W", zero.policy = TRUE)
# Plots first the map and then the neighbourhood relationship
plot(guerry$geometry)
plot(wrooks, st_centroid(guerry$geometry), col = "blue", add = TRUE)
```

2.2.2. Spatial Weights Matrices Based on Distances

There are situations where spatial contiguity matrices are not the most appropriate. For example, when spatial data are isolated points or locations, with no possibility of having common boundaries or in the case of polygonal data with "islands", i.e., units with no neighbours because they do not

have a common boundary with other units in the system. The zero-neighbour situation should be avoided in spatial weight matrices as it affects the calculations of the spatial autocorrelation statistics in which they participate.

In these cases, it is more convenient to define neighbourhood matrices based on distance-based criteria. These W matrices can be specified in two ways: local and global.

a) **Local** which, for each spatial unit, consider as a discrete or local neighbourhood domain, such as a given number of k nearest neighbours or a distance radius. As for the **k nearest neighbour W matrix**, its formal definition is the following:

$$\begin{cases} w_{ij} = 1 & ; & d_{ij} \leq d_i(k) \\ w_{ij} = 0 & ; & d_{ij} > d_i(k) \end{cases} ; \forall i \tag{2}$$

where d_{ij} is the distance between i and j and $d_i(k)$ is the distance at which each spatial unit i reaches a predetermined number of k nearest neighbours. Hence, in this type of configuration, all spatial units have exactly the same number k of spatial neighbours.

To illustrate the definition of the **W** matrices based on distances, a point map containing the centroids of the main Chilean cities, used in Vallone and Chasco [11] will be used. The R code required for the construction of k-nearest neighbour matrices is presented in Table 5. Two libraries are required to run this code: "spdep", which is defined as above, and "sf" (see Pebesma et al. [12]). The main functions involved in this R code, in addition to those already presented, are "*st_coordinates*", "*knearneigh*", "*knn2nb*", "*nb2listw*", "*plot*" and "*st_centroid*". **st_coordinates {sf}** retrieve the coordinates of spatial objects of class sf, sfc or sfg, in a matrix form of n rows and two columns for X and Y. **knearneigh {spdep}** returns a knn object, which is a matrix with the indices of points belonging to the set of the k nearest neighbours of each other; if longlat=*TRUE*, great-circle distances are used.[5] **knn2nb {spdep}** converts a knn object returned by the **knearneigh** function into a neighbours list of class nb with a list of integer vectors containing the index number (ID) of the neighbour regions. **st_centroid {sf}** is a geometric operation on a simple feature geometry (sf, sfc and sfg spatial objects), returning its centroid, which will be an object of the same class.

Table 5. R Code: construction of a 5-nearest neighbours W matrix (Figure 7A).

```
load("chilean_cities.rdata")
# The coordinates of the points are extracted.
coord <- st_coordinates(cities)
# Calculates the 5 nearest neighbours spatial weight matrix
w5knn <- knearneigh(coord, k=5, longlat= T) |> knn2nb()
w5knns<-nb2listw(w5knn, style="W", zero.policy = TRUE)
# Plots first the map and then the neighbourhood relationship
plot(cities$geometry)
plot(w5knn,st_centroid(cities$geometry),col="blue",add=TRUE)
```

Figure 7A presents the neighbourhood network corresponding to the **W** matrix of 5 nearest neighbours of the cities in Chile. As can be seen in the south of Chile, the neighbourhood network seems somewhat unrealistic as it connects cities that are actually more than 1,400 kilometres away by road from each other, as is the case of Puerto Aysén (Region XI) and Punta Arenas (Region XII).

Another local specification of the spatial weight matrices is based on the determination of a neighbourhood distance radius around each geographical unit. For example, the **minimum distance W matrix** considers as radius the minimum distance necessary for all regions or locations to have at least one neighbour. It is specified as follows:

---

[5] The great-circle distance is the shortest distance between two points on the surface of a sphere.

$$\begin{cases} w_{ij} = 1 & ; \quad d_{ij} < d_{min} \\ w_{ij} = 0 & ; \quad d_{ij} > d_{min} \end{cases} \tag{2}$$

for d_{min} the minimum distance necessary for all geographical units to have at least one neighbour.

The R code required for the construction of minimum distance neighbour matrices is presented in Table 6.

Table 6. R Code: construction of the minimum distance W matrix (Figure 7B).

```r
load("chilean_cities.rdata")
# The coordinates of the points are extracted.
coord <- st_coordinates(cities)
# Calculates the neighbourhood of 1 nearest neighbours
knn1 <- knearneigh(coord) |> knn2nb()
# Calculates the minimum distance for everybody's connection
min.dist <- max(unlist(nbdists(knn1,coord)))
# Generates the W matrix for which w_ij < d_min
k1 <- dnearneigh(coord, 0, min.dist)
wdmin <- nb2listw(k1)
wdmins<-nb2listw(k1, style="W", zero.policy = TRUE)
# Plots first the map and then the neighbourhood relationship
plot(cities$geometry)
plot(wdmin,st_centroid(cities$geometry),col="red", add=TRUE)
```


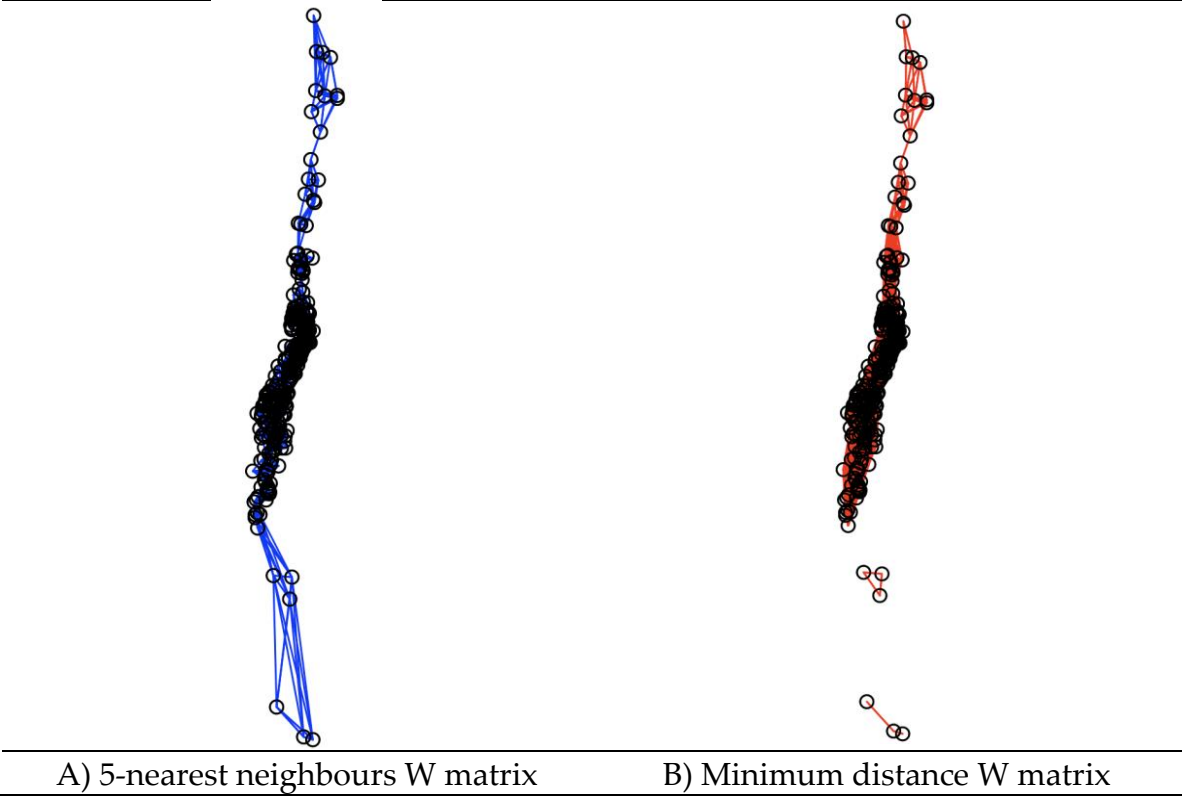| A) 5-nearest neighbours W matrix | B) Minimum distance W matrix |

Figure 7. Neighbourhood network of distance-based W matrices for the Chilean cities.

The libraries and functions needed to calculate these matrices are the same as in the previous cases, except for the functions "*max*" and "*unlist*" which are part of the basic R functions. and

"*nbdists*". **max {base}** returns the maximum of all the values present in its argument. **unlist {base}** simplifies a list structure to produce a vector which contains all the atomic components (logical, integer, real, complex, character and/or raw). **nbdists {spdep}** returns the Euclidean distances along the spatial neighbour links of an object type nb.

Figure 7B presents the neighbourhood network corresponding to the **W** matrix of minimum distance of the cities in Chile, which according to the integer atomic argument "min.dist" is equal to 215 km. Therefore, the cities of Puerto Aysén and Punta Arenas no longer appear to be linked in the neighbourhood network. Both cities are connected to other neighbouring cities in their respective regions, each forming its own urban network.

a) **Global** specifications of the spatial weight matrix consider as neighbours of each of the geographical units the rest of the units of the system, on a continuous basis, although with different weights or values depending on the distance that separates them. For example, the **inverse distance spatial weight matrix**, whose elements are defined as a gravity function is defined as follows:

$$w_{ij} = d_{ij}^{-\alpha} \tag{3}$$

for $\alpha \geq 0$, a friction parameter.

Although it is a matrix with non-zero elements (except for the main diagonal), from a certain distance between points the value of the w_{ij} elements will be practically zero. For this reason, a distance radius could usually be established to simplify the calculations (e.g., in the case of spatial systems with a large number of elements or regions), considering all the units beyond this radius as non-neighbours ($w_{ij} = 0$).

Table 7 presents the R code with the process of obtaining the inverse distance spatial weights matrix.

Table 7. R Code: construction of the inverse distance W matrices.

```r
load("chilean_cities.rdata")
# The coordinates of the points are extracted.
coord <- st_coordinates(cities)
# Calculates the neighbourhood of 5 nearest neighbours
knn5 <- knearneigh(coord, k=5, longlat= T) |> knn2nb()
# Calculates the inverse distance of knn5
dinv_knn5 <- lapply(knn5, function(x) 1/x)
# Calculates de row-standardized dinv_knn5 matrix
wdinv_knn5 <- nb2listw(knn5, style="W", zero.policy=TRUE)
# Obtains, for each unit, the nearest neighbour.
knn1 <- knearneigh(coord) |> knn2nb()
# Gets the minimum distance for all units having neighbours
min.dist <- max(unlist(nbdists(knn1,coord)))
# Calculates the neighbours in the minimum distance radius
k1 <- dnearneigh(coord, 0, min.dist)
# Computes the distances between neighbours in min.dist radius
dist_dmin <- nbdists(k1, st_coordinates(cities))
# Calculates the inverse distance of dist_dmin
dinv_dmin <- lapply(dist_dmin, function(x) 1/x)
# Calculates de row-standardized dinv_dmin matrix
wdinv_dmin <- nb2listw(k1, glist=dinv_dmin, style="W")
```

It establishes two types of cut-offs for each spatial unit: firstly, the group of 5 nearest neighbours and secondly, the radius of the minimum distance for which all spatial units have at least one neighbour. The libraries and functions needed to calculate these matrices are the same as in the previous cases, except for the function "*lapply*". **lapply {base}** returns a list of the same length as an "x" argument, each element of which is the result of applying a function to the corresponding element of "x".

Finally, the application of the spatial weights matrix to calculate a spatially lagged variable or **spatial lag** is presented. The spatial lag variable $\mathbf{W}y$ of the variable y is obtained by premultiplying the matrix $\mathbf{W}$ by that variable, as presented in the following example.

$$y = \begin{bmatrix} 120 \\ 300 \\ 240 \\ 60 \\ 180 \end{bmatrix} \ ; \ \mathbf{W}y = \begin{bmatrix} 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 120 \\ 300 \\ 240 \\ 60 \\ 180 \end{bmatrix} = \begin{bmatrix} 540/3 \\ 600/4 \\ 480/2 \\ 420/2 \\ 660/3 \end{bmatrix}$$

$$= \begin{bmatrix} 180 \\ 150 \\ 240 \\ 210 \\ 220 \end{bmatrix} \tag{4}$$

According to the matrix $\mathbf{W}$, the first spatial unit has three neighbours: units 2, 4 and 5. Therefore, each element of the first row of the standardised matrix, $\mathbf{W}^*$, is 1/3. The first element of the spatial lag variable is equal to 180, which is the average value of the variable y's neighbours of the first spatial unit: $\frac{(y_2 + y_4 + y_5)}{3} = \frac{300 + 60 + 180}{3} = \frac{540}{3} = 180 = \mathbf{W}^* y_1$. Therefore, each element i of the spatial lag variable, $\mathbf{W}^* y_i$, is a weighted average of the values of the variable y in the units neighbouring the spatial unit i.

2.3. Spatial Autocorrelation Measures

2.3.1. Moran's Scatterplot

A good tool for understanding spatial autocorrelation statistics is the Moran's scatterplot (see Anselin et al. [13]), which relates a variable to what is happening in its environment through its spatial lag in a scatter plot. Figure 8 shows this diagram: on the horizontal axis are the values of a variable and, on the vertical axis, the values of its respective spatial lag. Therefore, the origin of coordinates corresponds to the mean value of both variables.

Figure 8. Moran's scatterplot.
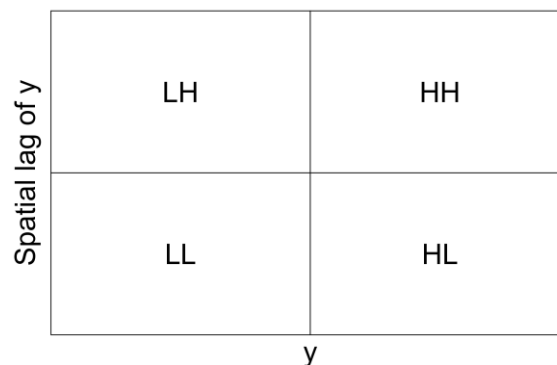
This diagram allows the identification of four quadrants: the first quadrant, called "High-High" (HH) contains the spatial units with values of both variables higher than the mean value of the system. The spatial units located in the third quadrant, "Low-Low" (LL) are the values of the variables below the average. Values located in the second and fourth quadrants, "Low-High" (LH) and "High-

Low" (HL), respectively, are those where the main variable is below/above the mean, with the values of their corresponding spatial lag being above/below the mean, respectively.

The R code with the process of obtaining a Moran's scatterplot representation is presented in Table 8. The only library needed to represent this diagram is "ggplot2". The functions employed, not previously presented, are "*annotate*" and "*theme_void*". **annotate {ggplot2}** adds geometric objects ("geoms") to a plot when their properties are introduced as vectors, which is useful for adding text labels (as it is the case here). **theme_void {ggplot2}** a completely empty theme of the non-data display.

Table 8. R Code: Moran's scatterplot (Figure 8).

```
ggplot() +
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 0, ymax = 1),
fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 0, ymax = 1),
fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 1, ymax = 2),
fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 1, ymax = 2),
fill="white",color="black")+
  annotate("text",x=.5,y=.5,label="LL", size=12) +
  annotate("text",x=1.5,y=1.5,label="HH", size=12) +
  annotate("text",x=.5,y=1.5,label="LH", size=12) +
  annotate("text",x=1.5,y=.5,label="HL", size=12) +
  annotate("text",x=1,y=-.1,label="y", size=12) +
  annotate("text",x=-.1,y=1,label="Spatial lag of y", size=12, angle=90) +
  theme_void()
```

If the values of the scatterplot are mostly located in the "HH" and "LL" quadrants, it means that spatial units with high/low values of the variable (above/below the average) are surrounded by spatial units that, on average, also have high/low values of the variable, respectively. This is an indication of the existence of positive spatial autocorrelation. If the concentration is in the "LH" and "HL" quadrants, it would be indicative of negative spatial autocorrelation: low surrounded by high and high surrounded by low, respectively. Using the same distribution rate of Catholic priests used in section 2.1 and using a spatial weights matrix defined with the queen criterion, Figure 9 presents the corresponding Moran scatterplot.

The R code with the process of obtaining this Moran's. scatterplot is presented in Table 9. The only library needed to represent this diagram is "spdep". A new function is employed, "*moran.plot*". **moran.plot {spdep}** plots a spatial variable data against its spatially lagged values.

Table 9. R Code: Moran's scatterplot (Figure 9).

```
load("guerry.rdata")
#Constructs the row-standardized queen contiguity matrix
wqueens<-poly2nb(guerry, queen=TRUE) |> nb2listw()
# Moran's scatter plot
moran.plot(guerry$clergy, wqueens, xlab="PRIESTS RATE",
           ylab="SPATIAL LAG OF PRIESTS RATE")
```

Figure 9 shows evidence of positive autocorrelation, i.e., French provinces with high/low rates of Catholic priests tend to be surrounded by provinces that, on average, also have high/low rates of Catholic priests, respectively. The Moran's scatterplot is only a graphical tool that needs to be complemented by more precise statistical measures, such as the Moran's I index of spatial autocorrelation.
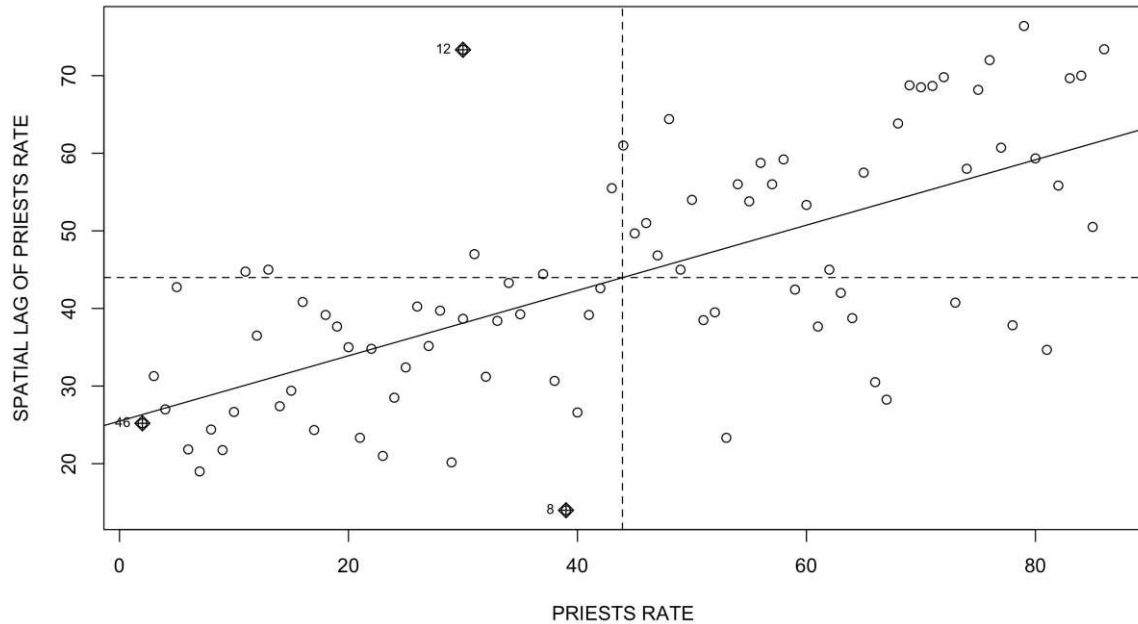

Figure 9. Moran's scatterplot of the rate of Catholic priests in the French provinces.

2.3.2. Global Moran's I Statistic

The Moran's I statistic of spatial autocorrelation is calculated as:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{5}$$

where n is equal to the number of spatial units in a system, $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ is the sum of all the elements, w_{ij}, of the spatial weights matrix, $z_i = y_i - \bar{y}$ is a variable y_i expressed in deviations from the mean value $(\bar{y})$, and the summation over j is such that only i's neighbouring values ($j \in J_i$; $i \neq j$) are included.

As can be seen, the second factor of Equation (6) consists of a quotient of the covariance of the variable y with its corresponding spatial lag, $\mathbf{W}y$, in the numerator and, in the denominator, the variance of the variable y or the product of the standard deviations of both variables (since both values are equal). As for the first factor of Equation (6), it is   equal to unity when the spatial weight matrix is row-standardised, since, in these cases, the sum of the weights is always equal to the sample size n. Hence, the Moran's I index can be interpreted as a linear correlation coefficient between a variable and its spatial lag, whose values lie within the range $[-1,1]$, so that the sign of the indicator coincides with the type of spatial autocorrelation: positive values ($I > 0$) are indicative of positive spatial autocorrelation and negative values ($I < 0$) are indicative of the existence of negative autocorrelation. Unlike what happens with the classical linear correlation coefficient, the complete absence of spatial autocorrelation does not correspond to the value zero exactly, except when n tends to infinity, but to the expression $-1/(n-1)$.

Moran's I value is usually represented in the Moran scatterplot with the **linear regression line**. For example, in Figure 9, the regression line of the variable y on $\mathbf{W}y$ is plotted, which in this case has a positive slope. The higher the degree of spatial autocorrelation in the variable y, the steeper the slope of the regression line. In addition, it is possible to calculate the level of statistical significance of the Moran's I value, through a **randomisation process**, by considering the null hypothesis of no

spatial autocorrelation and having as alternative the hypothesis of spatial autocorrelation, both positive and negative ([8]).

The R code with computation of the Moran's *I* indicator of the variable rate of Catholic priests plotted in Figure 9 is presented in Table 10. The only library needed to represent this diagram is "spdep". A new function is employed, "*moran.test*". **moran.test {spdep}** calculates the Moran's *I* test for spatial autocorrelation of a variable using a spatial weights matrix in weights list form.

Table 10. R Code and results: Moran's *I* value and statistical significance level.

```
load("guerry.rdata")
#Constructs the row-standardized queen contiguity matrix
wqueens<-poly2nb(guerry, queen=TRUE) |> nb2listw()
# Moran's I and p-value
moran.test((guerry$clergy), wqueens, randomisation=TRUE,
           alternative="two.sided")
##
##  Moran I test under randomisation
## data:  (guerry$clergy)
## weights: wqueens
## Moran I statistic standard deviate = 6.1632, p-value = 7.13e-10
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.421118648      -0.011904762       0.004936422
```

The p-value of $7.12 \cdot 10^{-10}$ does not allow to accept the null hypothesis of spatial randomness, so given that the value of $I = 0.421$ is of positive sign, we can conclude that the variable rate of Catholic priests has positive spatial autocorrelation in its provincial distribution in France.

As indicated in [5], there are variables whose distribution does not allow the null hypothesis of no spatial autocorrelation to be rejected. In these cases, it must be checked whether the absence of spatial autocorrelation occur **globally**, across the entire surface of the spatial system, or whether, on the contrary, there are certain spatial clusters of similar or dissimilar values **locally**, in some areas of the system. In other words, the spatial autocorrelation effect can occur globally and locally.

2.3.3. Local Moran's I Statistics

The effect of local spatial autocorrelation can be measured from LISA (Local Indicators of Spatial Autocorrelation) statistics and maps. LISA statistics are calculated as a decomposition of global spatial autocorrelation statistics, such as Moran's *I* test. In this way, it is possible to know the individual contribution of each spatial unit, and its environment, to the formation of the value of the global statistic. Moreover, it is possible to obtain, for each individual value and its neighbourhood, the level of statistical significance following a randomisation approach, as in the previous case (see Chasco [14]).

The formal expression of the local Moran's *I* statistics is the following:

$$I_i = \frac{z_i}{\sum_{j=1}^{n} z_j^2} \sum_{j=1}^{n} w_{ij} z_j \tag{6}$$

where the values of the variable z_i are expressed in deviations from the mean, an j only comprehends i's neighbours. For ease of interpretation, the weights w_{ij} may be row-standardized, and by convention, $w_{ij} = 0$.

As demonstrated in Anselin [15], a test for significant local spatial autocorrelation may be calculated by taking a conditional randomization approach. In this paper, it is also derived that the global I indicator can be expressed as an average of the local I_i statistics, corrected by a proportionality factor. Hence, it is possible to identify extreme values of I_i by, e.g., identifying outliers in a box plot, which will indicate the importance of the spatial unit i in determining the global I. They are not necessarily statistically significant, but are only similar to the identification of outliers, leverage, and influence points in the Moran's scatterplot (see Figure 9). In the map in Figure 10 we can find the hot spots (high values) painted in red and the cold spots (low values) painted in blue.

The R code with the process of obtaining the local Moran's statistics and map is presented in Table 11. The libraries needed to represent this diagram are "spdep", "ggplot2", "ggspatial", "ggthemes" and "grDevices", which contains functions for graphics devices and support for colours and fonts. The functions employed, not previously presented, are "*localmoran*", "*as.character*", "*attributes*", "*ifelse*", "*as.factor*", "*c*", "*rgb*" and "*scale_fill_manual*".

Table 11. R Code: Local Moran's *I* statistics and map (Figure 10).

```r
# Constructs the row-standardized queen contiguity matrix
wqueens<-poly2nb(guerry, queen=TRUE) |> nb2listw()
# Computation of the local Moran's statistics
local_m_Is <- localmoran(guerry$clergy, wqueens)
# Extract the position of the obs. in the Moran scatterplot
quad <- as.character(attributes(local_m_Is)$quadr[,1])
# Detects the obs. with significant values
sig <- local_m_Is[,5] <= .1
cluster <- ifelse(sig == FALSE, "non-significant",
                  as.character(quad)) |> as.factor()
# Configures the colors and other attributes of the map
colors <- c("Low-Low"="blue", "Low-High"=rgb(0,0,1,alpha=0.4),
            "High-Low"=rgb(1,0,0,alpha=0.4),
            "High-High"="red", "non-significant"="grey")
ggplot(guerry)+geom_sf(aes(fill=cluster))+
  scale_fill_manual(values = colors, name="Local Moran's Is")+
  annotation_scale(location = "tl") +
  annotation_north_arrow(location = "bl", which_north="true") +
  theme_map() +
  theme(legend.position = "bottom")
```

localmoran {spdep} computes the local Moran's I statistics for each spatial unit, based on a specified spatial weights object, joint to their corresponding returned "z-value" (significance level) that may be used as a diagnostic tool. **as.character {base}** creates a character vector of the specified length. **attributes {base}** returns an object's attribute list, understanding by attribute a set and not a vector, i.e., the order of the elements does not matter, but they must have unique names. **ifelse {base}** applies to all the elements of a vector, in a single call, two functions: **if {base}**, which is used to cause an operation to be executed only if a certain condition is met, and **else {base}**, which is used to state what to do in case a condition is not met. **as.factor {base}** encode a vector as a factor, being the factor levels ordered when the argument "ordered" is TRUE. **c() {base}** returns a vector (a one-dimensional array). **rgb {grDevices}** creates colours corresponding to the given intensities (between 0 and max) of the red, green, and blue primaries, the colour specification being referred to the standard sRGB colour

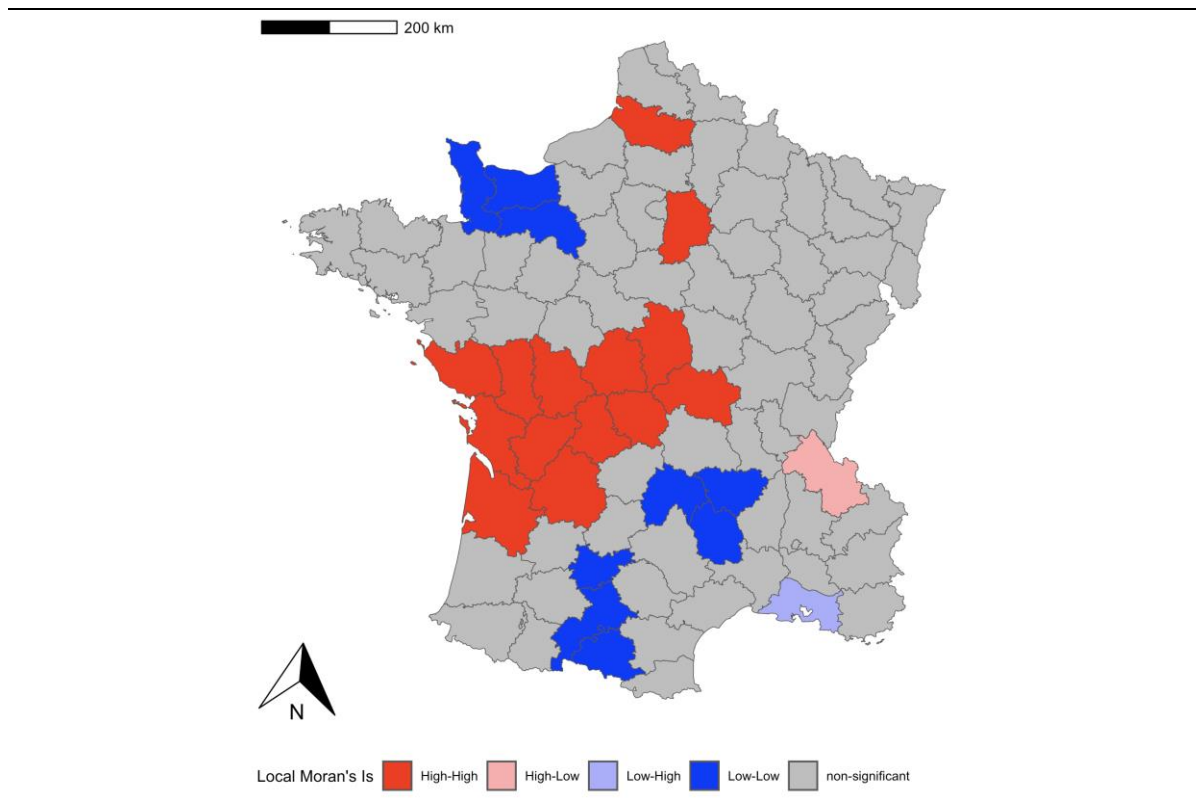space (IEC standard 61966). **scale_fill_manual {ggplot2}** define both colour and fill aesthetic mappings.

Figure 10. Moran's scatterplot of the rate of Catholic priests in the French provinces.

3. Cross-Section Spatial Econometric Models

3.1. Modelling Strategies for Spatial Autocorrelation Models

Spatial cross-sectional models are a particular case of cross-sectional econometric models and, as is the case with them, they must be identified before proceeding to their estimation and testing. Hence, it is important to follow a specific identification or modelling strategy for spatial models, which allows the researcher to know the correct population parameters from the observation of a data sample.

Traditionally, spatial econometrics has solved this problem by assuming that the model specification is *a priori* known, either from an existing economic theory, or from the results obtained by the application of an ESDA on the variables of the model, or by applying certain strategies consisting of the comparison of several competing models. Within the latter option, we can highlight two widely used modelling strategies: the one that goes from the specific (basic model without spatial autocorrelation effects) to a general model (with spatially lagged explanatory variables), STG, and the one that starts from a general model (the spatial Durbin model) to end up in a simpler spatial autocorrelation model or the basic regression model itself without spatial effects, GTS. But, from these two previous approaches, it is possible to propose a third hybrid strategy, which considers the good properties of the previous ones.

3.1.1. Anselin's Specific-To-General Strategy (STG)

The STG strategy, also known as "classical", was proposed by Anselin ([8]). The starting point of this strategy is a **basic linear regression model** without spatial effects:

$$y = \alpha \iota_n + \mathbf{X}\beta + \varepsilon \tag{7}$$

where y is the vector of the dependent variable, of order $(n \times 1)$; X is the matrix of explanatory variables, of order $(n \times k)$; ι_n is a vector of ones, of order $(n \times 1)$; α, β are the set of $(k + 1)$ parameters to be estimated; and ε is the random disturbance variable, of order $(n \times 1)$, which is distributed as $\varepsilon \sim N(0, \sigma_\varepsilon I_n)$, where I_n is the identity matrix of order n. This model is estimated by the **Ordinary Least Squares** (OLS) method. Table 12 presents the R code needed to generate the results of the OLS estimation of the basic linear regression model.

Table 12. R Code: OLS estimation of the basic linear regression model.

```
load("gdpmap.rdata")
# OLS estimation of the regression model
gdp_ols <- lm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap)
summary(gdp_ols)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0192207 -0.0043510  0.0003206  0.0040435  0.0216733
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.926e-01  6.216e-03  63.172  < 2e-16 ***
## LGH85       -3.984e-02  4.246e-03  -9.383 6.19e-16 ***
## BANK         6.565e-05  1.431e-05   4.588 1.13e-05 ***
## UNI01        5.107e-04  9.013e-05   5.667 1.06e-07 ***
## PAT00        4.979e-02  1.989e-02   2.504   0.0137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006755 on 117 degrees of freedom
## Multiple R-squared:  0.5849, Adjusted R-squared:  0.5708
## F-statistic: 41.22 on 4 and 117 DF, p-value: < 2.2e-16
```

The following two libraries are required to run this code: "sp" *and* "stats". The main functions, not previously presented, involved in this R code are "*lm*" and "*summary*". **lm {stats}** is used to fit linear models, including multivariate ones. **summary {base}** is used to produce result summaries of the results of various model fitting functions.

The code sequences show how to estimate and test the model with a dataset of municipalities (NUTS 5) of the urban areas of Spain used in Mella and Chasco [16]. With these data, a model of urban economic growth is formulated in which the average rate of change of GDP per capita, in logarithms, over the period 1985-2003 (LPGH), is explained as a function of GDP per capita in 1985 logarithms (LGH85), the rate of change in the number of banking institutions in the period 1985-2003 (BANK), the percentage of people with secondary and university education out of the population aged 16 and over in 2001 (UNI01) and the rate of the number of patents per inhabitant in 2000 (PAT00). As can be seen, all estimators are statistically significant at least at 99% confidence. The model has an explanatory level of 58.49%.

In order to test whether the variable of the OLS regression errors is spatially white noise,[6] the **Lagrange Multiplier** (LM) tests are calculated on this variable. This basic hypothesis is rejected as soon as any of these tests, which are distributed as a Chi-square with 1 degree of freedom (χ^2_1), is statistically significant. In particular, these tests are focused on a single alternative hypothesis. Thus, if the Lagrange multiplier test for the alternative hypothesis of spatially lagged dependent variable (LMLAG) is statistically significant instead of the Lagrange multiplier test for the alternative hypothesis of residual dependence (LMERR), the model that would best explain the data would be the **spatial lag model or spatial autoregressive model** of order 1 (SAR):

$$y = \rho \mathbf{W} y + \alpha \iota_n + \mathbf{X}\beta + \varepsilon \qquad (8)$$

being ρ a spatial autoregressive parameter to be estimated and $\mathbf{W}$ a row-standardised spatial weight matrix of order $(n \times n)$. But if it is the LMERR test that is statistically significant, instead of the LMLAG, the best model would be the **Spatial Error Model** (SEM):

$$\begin{cases} y = \alpha \iota_n + \mathbf{X}\beta + u \\ u = \lambda \mathbf{W} u + \varepsilon \end{cases} \qquad (9)$$

where λ is a spatial autoregressive parameter to be estimated. In this case, it should be noted that $y = \alpha \iota_n + \mathbf{X}\beta + (\lambda \mathbf{W} u + \varepsilon)$ and $u = y - \alpha \iota_n - \mathbf{X}\beta$. Therefore, the SEM would be equivalent to a spatial lag model that includes, on the right-hand side of the equation, in addition to the terms of the basic model, the spatially lagged dependent variable and the k spatially lagged explanatory variables as follows:

$$y = \lambda \mathbf{W} y + \alpha \iota_n + \mathbf{X}\beta - \lambda \mathbf{W X}\beta + \varepsilon; \text{ for } \theta = -\lambda \beta \qquad (10)$$

The restriction $\gamma = -\lambda \beta$ is called the **common factor (COMFAC) hypothesis** and when it holds, this model is called, by Anselin [8], as **Spatial Durbin Model** (SDM).

Regarding the LM tests previously presented, it is possible that both are statistically significant because, although they are tests oriented towards an alternative spatial hypothesis or model, they are also sensitive to the existence of the other type of spatial autocorrelation. In these cases, to decide on the most appropriate spatial model, spatial lag (SAR) or SEM, Anselin, Bera, Florax and Yoon [18] propose a solution by formulating robust versions of the LMLAG and LMERR tests, the new LMLE (a test to spatial lag dependence which is robust to ignored spatial error dependence) and LMEL (a test to spatial error robust to ignored spatial lag), respectively. If the values of LMLE > LMEL, the spatial lag (SAR) model should be selected, while if LMEL > LMLE, the SEM should be chosen.[7]

Table 13 presents the R code needed to generate the results of the computation of the LM test on the OLS estimation residuals. The following three libraries are required to run this code: "sp", "stats" *and* "spdep". The main functions, not previously presented, involved in this R code are "*coordinates*" and "*lm.LMtest*". **coordinates {sp}** retrieves spatial coordinates from a spatial object of class sp. **lm.LMtest {spdep}** reports the estimates of the LMERR test for error dependence (which is called 'LMerr' by the function), the LMLAG test for a missing spatially lagged dependent variable ('LMlag'), their corresponding robust variants LMEL ('RLMerr') and LMLE ('RLMlag'), respectively, and a

---

[6] In the context of spatial regression models, it is not recommended to use the Moran's *I* to test for the presence of spatial autocorrelation in the residuals, because the rejection of the null hypothesis of no spatial autocorrelation does not provide additional information about the model under the alternative hypothesis. Moreover, as demonstrated by Anselin and Rey ([17]) this test was shown to be very sensitive to misspecification errors, such as non-normality or heteroscedasticity.

[7] These robust tests are widely used, although according to Mur and Angulo [19] they do not seem to be so. necessary. Indeed, as they demonstrated with a Monte Carlo experiment, if the values LMLAG > LMERR, it will always be given that LMLE > LMEL, and *vice versa*. In other words, according to these authors, when the LMLAG and LMERR tests are statistically significant, it would be sufficient to compare their values, without the need to calculate their corresponding robust tests.

SARMA test, which is test for a mixed residual spatial autoregressive process (SAR) and a spatial moving average (SMA).

The LMLAG and LMERR tests are both highly significant, so the null hypothesis of no spatial autocorrelation must be rejected with more than 99% confidence. Additionally, of the robust tests, it is only possible to reject the null hypothesis for LMEL (not for LMLE). Therefore, according to the classic modelling strategy, the most appropriate specification for this model would be the SEM.

Table 13. R Code: LM tests on the OLS estimation residuals.

```r
load("gdpmap.rdata")
#Constructs the row-standardized minimum distance matrix
coord <- coordinates(gdpmap)
knn1 <- knearneigh(coord) |> knn2nb()
min.dist <- max(unlist(nbdists(knn1,coord)))
k1 <- dnearneigh(coord, 0, min.dist)
dmin <- nb2listw(k1)
dmins <-nb2listw(k1,style="W") # W: Row-standardized
# OLS estimation of the regression model
gdp_ols <- lm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap)
## LM tests
lm.LMtests(gdp_ols, dmins, test="all")
##
##  Lagrange multiplier diagnostics for spatial dependence
## data:
## model: lm(formula = LPGH ~ LGH85 + BANK + UNI01 + PAT00, data = gdpmap)
## weights: dmins
## LMerr = 21.641, df = 1, p-value = 3.288e-06
## LMlag = 6.944, df = 1, p-value = 0.00841
## RLMerr = 14.715, df = 1, p-value = 0.0001251
## RLMlag = 0.017854, df = 1, p-value = 0.8937
## SARMA = 21.659, df = 2, p-value = 1.981e-05
```

Since the spatial lag model (SAR) includes as an explanatory variable the spatially lagged endogenous variable ($\boldsymbol{W}y$) referred to the same moment in time as the dependent variable (y), a situation of simultaneity or contemporaneous dependence arises. Therefore, its estimation by OLS produces bias, inefficiency, and inconsistency in the estimators. As for the SEM, due to the heteroskedastic form of the variance and covariance matrix of the random disturbance u, its estimation by the OLS method results in inefficient, although consistent, estimators. Due to the problems of OLS in these spatial models, estimation by the **Maximum Likelihood** (ML) method in the case of normality in the OLS error variable is recommended as more appropriate (see Anselin [8], chap. 6).

Table 14 presents the R code needed to generate the results of the ML estimations of the spatial lag (SAR) model and the SEM. The following four libraries are required to run this code: "sp", "stats", "spdep" and "tseries".

The main functions, not previously presented, involved in this code are "*jarque.bera.test*", "*lagsarlm*" and "*errorsarlm*". **jarque.bera.test {tseries}** tests the null of normality for a variable using the Jarque-Bera test statistic). **lagsarlm {spatialreg}** provides ML estimation of spatial lag (SAR) models and SDM. **errorsarlm {spatialreg}** provides a ML estimation of SEM.

Table 14. R Code: ML estimation of the spatial lag (SAR) model and the SEM.

```r
load("gdpmap.rdata")
# Constructs the row-standardized minimum distance matrix 'dmins' (like in Table
 13)
# JB normality test on the OLS residuals
gdp_ols <- lm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap)
jarque.bera.test(gdp_ols$res)
##
## X-squared = 5.0298, df = 2, p-value = 0.08087
##
# ML estimation of the SAR model
gdp_sar <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins)
# ML estimation of the SEM
gdp_err <- errorsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins,
                      tol.solve=1e-16)
# Output of the SEM ML estimation
## Call: errorsarlm(formula = LPGH ~ LGH85 + BANK + UNI01 + PAT00, data =
gdpmap,
##      listw = dmins, tol.solve = 1e-16)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -1.7697e-02 -3.6030e-03  3.1832e-05  3.7508e-03  2.1121e-02
## Type: error
##
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  3.9282e-01  6.7922e-03 57.8342 < 2.2e-16
## LGH85       -3.9828e-02  4.5709e-03 -8.7134 < 2.2e-16
## BANK         6.7328e-05  1.3820e-05  4.8718 1.106e-06
## UNI01        5.2750e-04  8.7403e-05  6.0352 1.587e-09
## PAT00        2.9814e-02  1.8091e-02  1.6480   0.09936
##
## Lambda: 0.57331, LR test value: 12.96, p-value: 0.00031815
## Asymptotic standard error: 0.12688
##      z-value: 4.5187, p-value: 6.222e-06
## Wald statistic: 20.419, p-value: 6.222e-06
## Log likelihood: 445.6128 for error model
## ML residual variance (sigma squared): 3.813e-05, (sigma: 0.006175)
## Number of observations: 122
## Number of parameters estimated: 7
## AIC: NA (not available for weighted model), (AIC for lm: -866.27)
```

In Table 13, the LM tests on the OLS residuals showed the existence of spatial aucorrelation in the residuals and recommended the estimation of a SEM to correct for this problem. In Table 14, the Jarque-Bera test cannot reject the null hypothesis of normality of the OLS error terms with more than 95% confidence, so it is possible to estimate the SEM by the ML method. For this reason, only the results of this estimation in their entirety are presented in Table 14.

According to the STG strategy, the SEM is the model which best fits the data generation process. The value of the spatial autoregressive parameter $\lambda = 0.57331$ has no interpretation, unlike the spatial autoregressive ρ parameter, which is estimated in the spatial lag model (SAR). In fact, it is called by Anselin and Rey as a "nuisance" parameter and, therefore, no inference is performed for it ([17]). All estimators of the model are statistically significant although, with respect to the estimation of the basic model (Table 12), the estimator of the patent variable is only significant for a confidence level above 90%. In the case of the ML estimates, the R^2, as a measure of goodness of fit, is not presented, but rather the ML and information criteria (Akaike and AIC).

The modelling strategy proposed by Anselin and Rey ([17]) is presented in Figure 11.
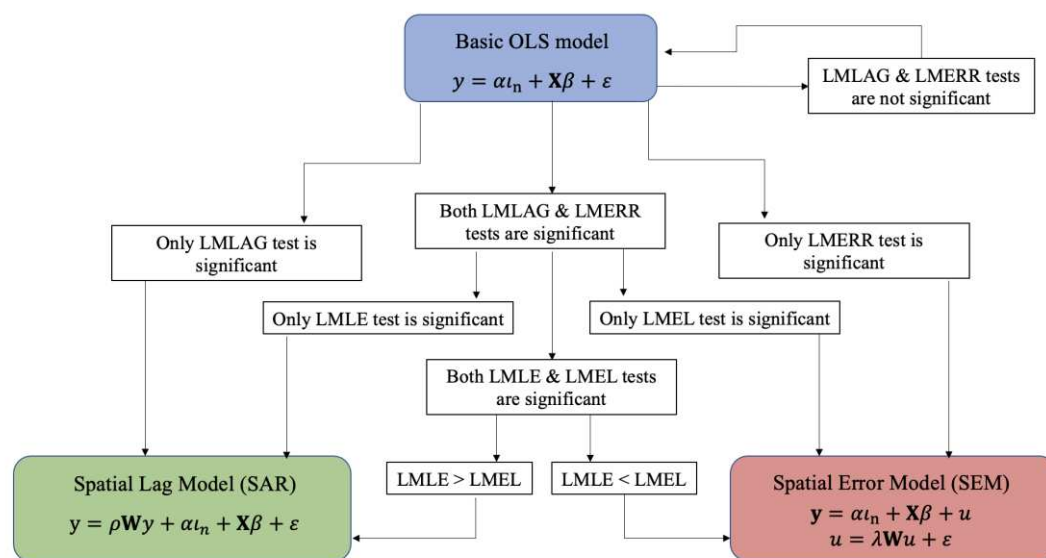

Figure 11. Classic STG strategy for identifying spatial autocorrelation models.

When the OLS estimation errors follow a non-normal distribution, the spatial lag model (SAR) should be estimated by the **Spatial Two-Stage Least Squares** method (S2SLS) which, in the version of Kelejian and Robinson [20], consists of using the spatially lagged exogenous variables of several orders of contiguity as instruments of the spatially lagged endogenous variable, resulting in a consistent, though not very efficient, autoregressive estimator ($\hat{\rho}$).

As for the SEM, Arraiz, Drukker, Kelejian and Prucha [21] present an estimation by the **General Method of Moments** (GMM), building on the initial proposal made by Kelejian and Prucha [22]. Although, given that the estimators are unbiased and consistent, it is also considered acceptable to estimate the SEM by OLS by performing robust inference of the variance-covariance matrix of the estimators by the KP-HET method proposed by Kelejian and Prucha [23], which takes into account the joint existence of heteroscedasticity and spatial autocorrelation in the regression errors.

Table 15 presents the R code needed to generate the results of the STSLS and GMM estimations of the spatial lag (SAR) model and the SEM, respectively. The following three libraries are required to run this code: "sp", "spdep" *and* "spatialreg".

The main functions, not previously presented, involved in this R code are ""*stsls*" and "*GMerrorsar*". **stsls {spatialreg}** fits a spatial lag model (SAR) by STSLS. **GMerrorsar {spatialreg}** fits a SEM by the Kelejian and Prucha's GMM.

Table 15. R Code: Other estimations of the spatial lag (SAR) model and the SEM.

```
load("gdpmap.rdata")
# Constructs the row-standardized minimum distance matrix 'dmins' (like in Table
 13)
# STSLS estimation of the SAR model
gdp_tsls <- stsls(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins)
# GMM estimation of the SEM
gdp_GMerr <- GMerrorsar(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins,
                        returnHcov=FALSE, method="nlminb", verbose=TRUE)
```

3.1.2. LeSage's General-To-Specific Strategy (GTS)

The second strategy, GTS, is to start from the most general spatial autocorrelation model possible. According to Manski [24], this general model is the one that includes the three possible types of spatial interaction, endogenous ($\mathbf{W}y$), exogenous ($\mathbf{W}X$) and unobserved ($\mathbf{W}u$) effects:

$$\begin{cases} y = \rho\mathbf{W}y + \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + u \\ u = \lambda\mathbf{W}u + \varepsilon \end{cases} \tag{11}$$

where θ is a vector of k spatial autoregressive parameters. The problem with this Manski's model is that, as the author demonstrates, it is impossible to identify its parameters. Therefore, it is necessary to reduce the three types of spatial interaction to two, which gives rise to three possible general sub-models of spatial autocorrelation. First, if we exclude the spatial endogenous effect ($\rho = 0$), we obtain the so-called, by LeSage and Pace [25], **Spatial Durbin Error Model** (SDEM):

$$\begin{cases} y = \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + u \\ u = \lambda\mathbf{W}u + \varepsilon \end{cases} \tag{12}$$

Secondly, if what is excluded is the spatial exogenous effect (($\theta = 0$), the **SARAR model** or Kelejian-Prucha's model is obtained, as it was proposed by these authors ([22])[8]:

$$\begin{cases} y = \rho\mathbf{W}y + \alpha\iota_n + \mathbf{X}\beta + u \\ u = \lambda\mathbf{W}u + \varepsilon \end{cases} \tag{13}$$

Finally, if the spatial unobserved effect is excluded, we obtain the unconstrained **Spatial Durbin Model** (SDM), where the constraint presented in Equation (11) do not hold; that is, when $\theta \neq -\lambda\beta$. Hence:

$$y = \rho\mathbf{W}y + \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + \varepsilon \tag{14}$$

As with the SAR model, the presence of the spatial lag of the dependent variable on the right-hand side of the equation does not result in OLS estimators with good properties, so this model must be estimated by ML.

This is the general model proposed by LeSage and Pace [25] as a starting point for the GTS modelling strategy.[9] The SDM model fulfils the identifiability condition by including two of the three possible types of spatial interaction, endogenous and exogenous effects, and thus includes all spatially lagged explanatory variables. In this way, a possible bias in the estimators caused by the omission of any relevant spatial variable is avoided.

Additionally, LeSage and Pace demonstrated that if this model also had residual spatial autocorrelation problems, the omission of the spatially lagged error variable would lead to

---

[8] For a more complete summary of the specification, estimation and testing of this and other models proposed by these authors, see Chasco [26].

[9] Although in section 2.2 of their book, LeSage and Pace present the SDM as the Anselin's restricted model ([8]), from chapter 2 forward, they call the unrestricted SDM as simply 'SDM'. Therefore, from here onwards, when we say SDM, we will refer to the unconstrained SDM.

inefficiency, but not to bias in the estimators. The SDM model likewise has the property of nesting several models, from the basic model without spatial effects to the SAR and SEM spatial autocorrelation models (when the COMFAC hypothesis is satisfied), as well as the so-called, by LeSage and Pace [25], **Spatial Lag of X (SLX) model**, firstly called "mixed regressive-spatial cross-regressive model" by Florax and Folmer [27].

$$y = \alpha \iota_n + \mathbf{X}\beta + \mathbf{WX}\theta + \varepsilon \qquad (15)$$

This model can be estimated by the OLS method since, if the explanatory variables are exogenous, their corresponding spatial lags will also be exogenous. In addition, the SLX model has two more good properties: on the one hand, it is more flexible to estimate or parameterise the **W** matrix and, on the other hand, it has better properties to capture spatial spillover effects when no clear theoretical model is available to support the inclusion of the endogenous spatial interaction effect (**W**y), as shown by Halleck Vega and Elhorst [28].

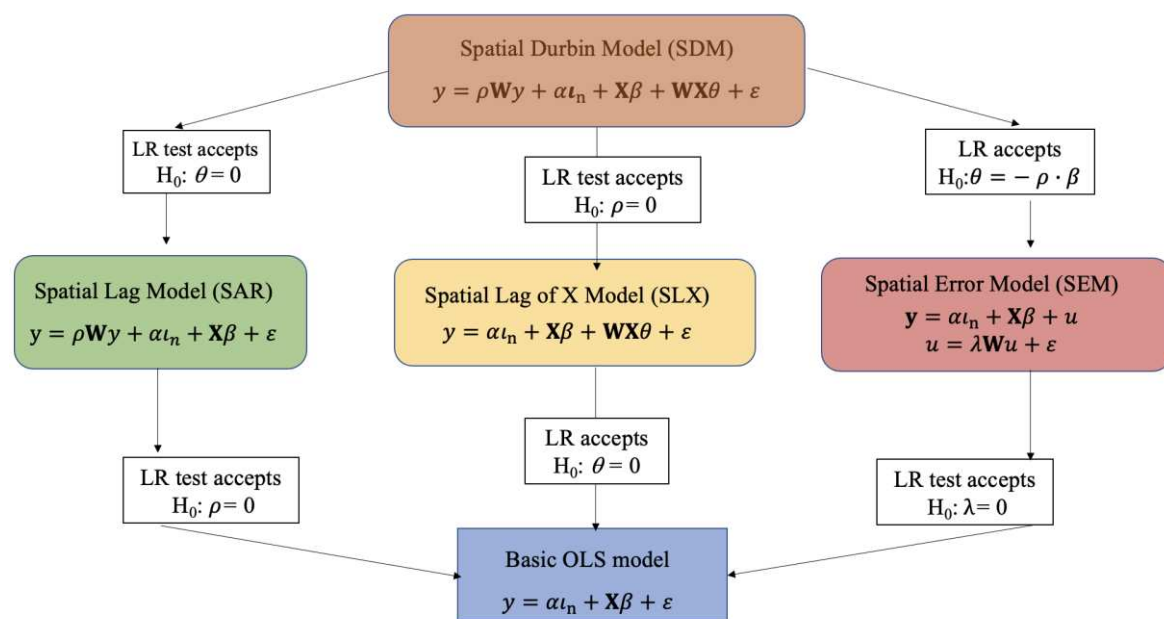The modelling strategy proposed by LeSage and Pace ([25]) is presented in Figure 12.

Figure 12. LeSage and Pace's GTS strategy for identifying spatial autocorrelation models.

Table 16 presents the R code needed to generate the results of the ML estimation of the SDM model. The following three libraries are required to run this code: "sp", "spdep" *and* "spatialreg". In this case, there are no new functions.

In the output, the estimation results highlight the contrast between the high statistical significance of the model's explanatory variables and the low significance of their corresponding spatial lags. The only exception is the variable of patents per capita and its corresponding spatially lagged variable, for which both estimated coefficients are highly significant, especially in the case of the latter.

However, the lack of statistical significance of the spatial autoregressive coefficient $\hat{\rho}$ (p-value: 0.40193) is striking, raising doubts about the suitability of this identification as the most appropriate for the model. Finally, the function also calculates an LM test on the residuals of this regression, which is not significant, demonstrating the non-existence of spatial autocorrelation in the residuals (see Anselin and Rey [17] for further information about this LM test).

As can be seen in Figure 12, the decision on the most appropriate model for the data generating process, according to this GTS strategy, requires the calculation of several Likelihood Ratio (LR) tests, which will be discussed in more detail in the next section, where we will present a hybrid strategy, which combines the two strategies seen so far: STG and GTS.

Table 16. R Code: ML estimation of the SDM.

```
load("gdpmap.rdata")
# Constructs the row-standardized minimum distance matrix (like in Table 13)
# ML estimation of the SDM
gdp_sdm <- lagsarlm(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins,
                    type="mixed")
summary(gdp_sdm)
##
## Residuals:
##         Min          1Q      Median          3Q         Max
## -0.01641172 -0.00300317 -0.00034465  0.00309880  0.02128742
## Coefficients: (asymptotic standard errors)
##                 Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  3.4167e-01  8.1083e-02  4.2138 2.511e-05
## LGH85       -4.2636e-02  4.6043e-03 -9.2602 < 2.2e-16
## BANK         6.8636e-05  1.3429e-05  5.1112 3.202e-07
## UNI01        5.3610e-04  8.5042e-05  6.3040 2.901e-10
## PAT00        4.6891e-02  1.8237e-02  2.5712 0.0101356
## lag.LGH85   -9.2796e-03  1.3781e-02 -0.6733 0.5007253
## lag.BANK    -3.2110e-06  4.3915e-05 -0.0731 0.9417113
## lag.UNI01   -3.7962e-05  2.4924e-04 -0.1523 0.8789439
## lag.PAT00    3.1468e-01  9.4438e-02  3.3322 0.0008617
##
## Rho: 0.17621, LR test value: 0.70253, p-value: 0.40193
## Asymptotic standard error: 0.19735
##     z-value: 0.8929, p-value: 0.37191
## Wald statistic: 0.79726, p-value: 0.37191
## Log likelihood: 452.4774 for mixed model
## ML residual variance (sigma squared): 3.5084e-05, (sigma: 0.0059232)
## Number of observations: 122
## Number of parameters estimated: 11
## AIC: -882.95, (AIC for lm: -884.25)
## LM test for residual autocorrelation
## test value: 0.90182, p-value: 0.3423
```

3.1.3. Elhorst's Hybrid Strategy

Based on the two previous approaches, Elhorst [29] proposes a hybrid strategy, which takes into account the good properties of both proposals. For this reason, this strategy will be the one we select as the most suitable for identifying spatial autocorrelation models. As presented in Figure 13, the Elhorst's hybrid strategy starts, like the STG strategy, with the OLS estimation of a basic model without spatial effects. The error variable of this regression is analysed with the LMLAG and LMERR tests, to check whether they are white noise. At this point, it may happen that one of the tests is statistically significant or that none of them is. Firstly, if any of the LM tests is significant, it is recommended to select the SDM model, as proposed by the GTS strategy.

The ML estimation of this model allows the likelihood ratio (LR), whose distribution follows a Chi-square with k degrees of freedom (χ_k^2), to be used to test the null hypotheses $H_0(\theta = 0)$ and $H_0(\theta = -\rho \cdot \beta)$. If the second (COMFAC) hypothesis cannot be rejected, the SDM should be simplified to a spatial lag (SAR) model, provided that the LMLAG > LMERR tests. If the first hypothesis cannot be rejected, the SEM should be selected, provided that LMERR > LMLAG tests. If there is no agreement between the results of the LR test and the LM tests, then the SDM would be the model that best describes the data.

Secondly, if after the OLS estimation of the basic model none of the LM tests is statistically significant, then the basic model would have to be re-estimated as an SLX model, including all spatially lagged exogenous variables or a subset of them, in order to test the null hypothesis $H_0(\theta = 0)$. If this hypothesis cannot be rejected, the basic model should be chosen as the one that best describes the data, i.e., there would be no evidence of the need for spatial autocorrelation effects to explain the dependent variable. But if, on the contrary, the hypothesis $H_0(\theta = 0)$ can be rejected, the SDM model would have to be estimated to test, again, the null hypothesis $H_0(\rho = 0)$. If this hypothesis can be rejected, the selection would be the SDM; on the contrary, it should be settled that a model with spatially lagged independent variables (complete or parsimonious SLX) only suffices.
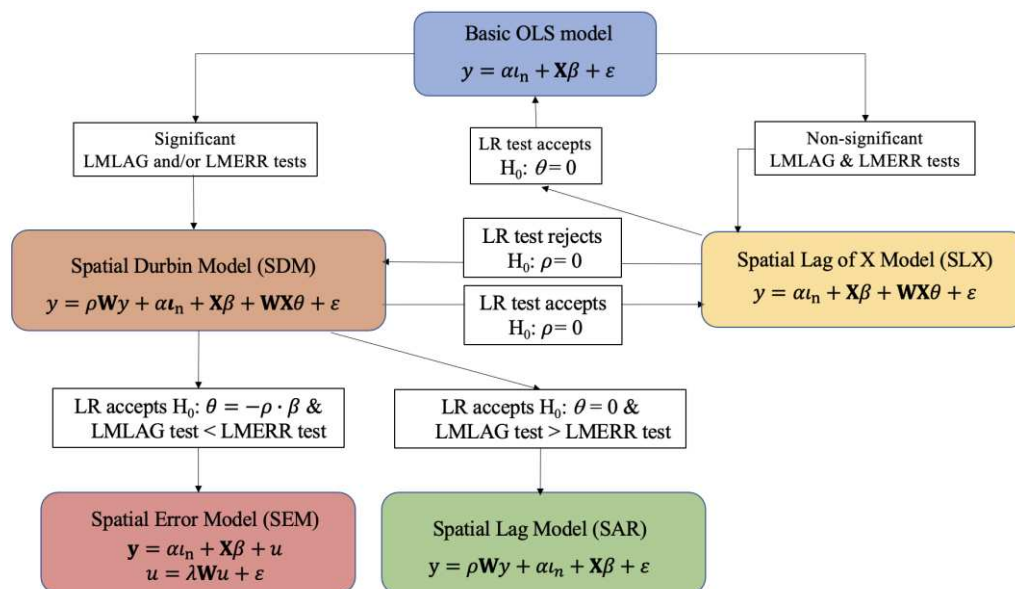

Figure 13. Elhorst's hybrid strategy for identifying spatial autocorrelation models.

Additionally, Halleck Vega and Elhorst [28] introduced the SLX model into the model selection process as a new SDM's nested model (like in Figure 12), recommending its choice when the null hypothesis $H_0(\rho = 0)$ cannot be rejected. This addition may lead to differences in the final model selected, as will be seen below.

Table 17 presents the R code needed to compute the LM and LR tests necessary to determine the best model specification according to the Elhorst's hybrid strategy. The following four libraries are required to run this code: "sp", "stats", "spdep" *and* "spatialreg". The main functions, not previously presented, involved in this R code are "*lmSLX*" and "*LR.Sarlm*". **lmSLX {spatialreg}** fits a SLX model, i.e., an OLS model augmented with the spatially lagged regressor variables. **LR.Sarlm {spatialreg}** is a function which provides a likelihood ratio test.

As seen in the results of the classical strategy (Table 13), the LM tests on the OLS residuals are both statistically significant, with LMLAG < LMERR. Therefore, the SEM was identified as the most appropriate specification for the urban growth model. However, according to Elhorst's hybrid strategy, the significance of any or all of the LM tests involves estimating the SDM and then comparing it with the LR tests of their corresponding more restricted nested models. In his first version in 2010 [29], Elhorst proposes to compare the SDM with the spatial lag model (SAR) and the

SEM. As can be seen in Table 17, the null hypothesis of the fulfilment of the COMFAC hypothesis, $H_0(\theta = -\rho \cdot \beta)$, and the null hypothesis of $H_0(\theta = 0)$ must be rejected. In other words, the most appropriate specification for the model is the SDM. However, if SLX is incorporated into the comparison of rival models, which is what Elhorst does in his second formulation in 2015 [28], the null hypothesis $H_0(\rho = 0)$ cannot be rejected. Therefore, the more appropriate specification for the data-generating process would be the SLX model, rather than the SDM.[10] Specifically, a more parsimonious SLX model is selected that only includes, as explanatory variables, the spatially lagged variables that are statistically significant, as shown in the last rows of the Table 17.

Table 17. R Code: Elhorst's hybrid model specification strategy.

```
load("gdpmap.rdata")
# OLS estimation of the basic model  (like in Table 12)
# Constructs the row-standardized minimum distance matrix (like in Table 13)
# LM tests on the OLS model residuals (like in Table 13)
# ML estimation of the SDM (like in Table 16)
# ML estimation of the SAR model (like in Table 14)
# ML estimation of the SEM (like in Table 14)
# OLS estimation of the SLX model
gdp_slx <- lmSLX(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins)
# LR tests on SDM and its nested models
LR.Sarlm(gdp_sdm, gdp_sar) # SDM vs. SAR
## Likelihood ratio = 21.37, df = 4, p-value = 0.0002674
LR.Sarlm(gdp_sdm, gdp_err) # SDM vs. SEM (COMFAC)
## Likelihood ratio = 13.729, df = 4, p-value = 0.008211
LR.Sarlm(gdp_sdm, gdp_slx) # SDM vs. SLX
## Likelihood ratio = 0.70253, df = 1, p-value = 0.4019
# OLS estimation of a parsimonious SLX model
gdp_slx2 <- lmSLX(LPGH~LGH85+BANK+UNI01+PAT00, data=gdpmap, listw=dmins,
                Durbin=~LGH85+PAT00)
```

Therefore, the selection of the most appropriate final model is still an open question since, as we have seen, the outcome depends on the modelling strategy adopted. In the proposed example of urban growth in Spain, if the Anselin's classical STG strategy is followed, the selected model would be the **SEM**. According to the original proposal of Elhorst's hybrid strategy, the proposed model would be the **SDM**. Finally, according to Elhorst's second proposal, and also LeSage and Pace's GTS strategy, the model finally selected would be the **SLX**.

All these models must be estimated by the OLS and ML methods to be able to use the LR as a testing tool, although it should be noted that spatial autocorrelation models can also be estimated with **Bayesian methodology** using the Markov Chains Monte Carlo (MCMC) approach, as explained in LeSage and Pace [25], chapter 5.

To conclude this section on the identification of true data generation process of a dependent variable, it must be said that there is still a long way to go to create a method that considers not only the existence of the spatial autocorrelation effect, but also **spatial heterogeneity**, as shown by Debarsy and Le Gallo [30]. Spatial heterogeneity can manifest itself in various forms, such as diversity of coefficients or of the functional relationships themselves in various locations or groups of locations

---

[10] The application of these results to the GTS strategy presented in Figure 12 also leads to this result: that the SLX model is the most appropriate model to explain urban economic growth data in Spain.

(spatial regimes), spatial clustering, hierarchical structures, etc. But this is a topic that will not be dealt with in this paper.

3.2. Interpretation of the Estimators of Spatial Autocorrelation Models

Only in spatial autocorrelation models where the endogenous effect ($\mathbf{W}y$) is not present in the right-hand side of the model, can the estimated coefficients ($\hat{\beta}$) be interpreted directly, as in the basic model without spatial effects, because the reduced form coincides with the structural form. Of the specifications presented above, this would be the case for the basic OLS model, the SDEM and the SLX model. That is, the marginal effect of a change in the value of a continuous explanatory variable for a given location coincides with the value of the estimator corresponding to that variable:

$$\frac{\partial y_i}{\partial x_{i,k}} = \beta_k \tag{16}$$

In all other cases, the correct interpretation of the estimators involves first **moving from the structural form to the reduced form**. Thus, for example, in the spatial lag (SAR) model of Equation (9) the reduced form would be (under certain invertibility conditions):

$$y = (\mathbf{I} - \rho\mathbf{W})^{-1}(\alpha\iota_{\mathrm{n}} + \mathbf{X}\beta + \varepsilon) \tag{17}$$

The term $(\mathbf{I} - \rho\mathbf{W})^{-1}$ is called **the spatial multiplier** and, using the potential expansion, can also be expressed as follows:

$$(\mathbf{I} - \rho\mathbf{W})^{-1} = \mathbf{I} + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \rho^3\mathbf{W}^3 + \cdots \tag{18}$$

If this new expression is used in Equation (18), for the conditional mean, the value of y at a given location i is a function not only of the value of the explanatory variables at that location, but also the value of the explanatory variables at neighbouring locations (through the term $\rho\mathbf{W}\mathbf{X}\beta$), of the value of the explanatory variables at neighbours' neighbours (through the term $\rho^2\mathbf{W}^2\mathbf{X}\beta$), etc.

$$\mathrm{E}(y|\mathbf{X}) = \mathbf{X}\beta + \rho\mathbf{W}\mathbf{X}\beta + \rho^2\mathbf{W}^2\mathbf{X}\beta + \rho^3\mathbf{W}^3\mathbf{X}\beta + \cdots \tag{19}$$

Therefore, if we restrict our analysis to the relationship between the variable y as a whole and a given variable x_k of the matrix $\mathbf{X}$, the marginal effect of a change in all the values of the column vector x_k on all the values of y would be as follows:

$$\frac{\partial y}{\partial x_k'} = \frac{\partial(x_k\beta_k + \rho\mathbf{W}x_k\beta_k + \rho^2\mathbf{W}^2 x_k\beta_k + \rho^3\mathbf{W}^3 x_k\beta_k + \cdots)}{\partial x_k'} \tag{20}$$

Unlike the marginal effect obtained for the basic model (Equation 17), the result of this expression does not refer to a given location i, but to all of them as a whole, and it is not a scalar (β_k) either, but a matrix of order $n \times n$. If we now calculate the effect of discrete changes in the elements of x_k on the values of y:

$$\Delta y = (\Delta x_k)\beta_k + \rho\mathbf{W}(\Delta x_k)\beta_k + \rho^2\mathbf{W}^2(\Delta x_k)\beta_k + \rho^3\mathbf{W}^3(\Delta x_k)\beta_k + \cdots \tag{21}$$

As can be seen in the right-hand side of the equation, the total effect on the values of the dependent variable and of changes in the values of an explanatory variable, x_k, is the result of a direct effect (first summand) plus an indirect effect from the spatial multiplier (remaining summands). From this expression, two types of analysis can be distinguished, one particular and the other general.

First, Anselin and Rey [17] analyse the particular case of the total effect produced by a unit change of x_k, in all the locations at the same time, on each and every value of y, which would be equal to a scalar, $\beta_k/(1 - \rho)$. This is the result of a direct effect (β_k) plus an indirect effect driven by the spatial multiplier, ($\beta_k \cdot \rho/(1 - \rho)$). As explained by these authors, this expression can be used to simulate the total effects of certain changes in an explanatory variable on all the observations of the dependent variable.

Second, LeSage and Pace [25] present the more general case of the effects caused by the change in the value of "one" location, j, of the variable x_k on "another" location, i, of the variable y. Here, each effect is no longer a scalar, but a complete matrix, $S_k(\mathbf{W})_{ij}$, of order $n \times n$. That is, each explanatory variable x_k in the model will have its own full matrix of impacts on the dependent variable.

$$S_k(\mathbf{W})_{ij} = \frac{\partial y_i}{\partial x'_{jk}} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\beta_k \tag{22}$$

In this case, we can also distinguish between direct and indirect effects. The direct effect would be the effect caused by changes in the value of x_k, at location i on the value of y at that same location i. These effects constitute the values of the main diagonal of the matrix $S_k(\mathbf{W})_{ii}$. The indirect effect is constituted by the rest of the values of the matrix $S_k(\mathbf{W})_{ij}$, which would be the "feedback loops" in which the value of x_k, in a location j affects the value of y in location i, and *vice versa*, being possible longer runs in which the effect in one location could reach the last observation n of the spatial system, and then go back to the starting point. For example, the value of $S_k(\mathbf{W})_{11}$ would mean the direct effect of a change in the value of variable x_k at location 1 ($x_{1,k}$) on the value of variable y at location 1 (y_1), while the value of $S_k(\mathbf{W})_{13}$ would be the indirect effect of a change in the value of variable $x_{1,k}$ on the value of variable y at location 2 (y_2). In the rows, the matrix $S_k(\mathbf{W})_{ij}$ has the effects of x_k from each location i "to" every one of the locations j, while the columns represent the effect on each location j "from" every one of the locations i.

Since it is not possible to construct an **inferential process** for all impacts in the matrix $S_k(\mathbf{W})_{ij}$, LeSage and Pace propose to make the inference on the mean value of the direct and total effects, extracting the indirect effects by difference:

$$\overline{M}(k)_{direct} = \text{tr}[S_k(\mathbf{W})]/n$$
$$\overline{M}(k)_{total} = \iota'_n S_k(\mathbf{W})\iota_n/n \tag{23}$$
$$\overline{M}(k)_{indirect} = \overline{M}(k)_{total} - \overline{M}(k)_{direct}$$

In LeSage and Pace [25], chapter 2, the same process for the more complex SDM are laid out.

Therefore, the interpretation of the coefficients in spatial autocorrelation models is not always equivalent to the coefficient estimated by the regression, $\hat{\beta}$, and it is first necessary to express the model in its reduced for to identify the correct mathematical expression. Moreover, it is also typical of some of these models that, in addition to the direct effect caused by unit variations in the independent variables on the dependent variable, in each observation, there is an indirect or **spillover effect** from variations in neighbouring units. Therefore, each explanatory variable produces a total effect on the explanatory variable that is the sum of the two previous ones. As explained by Halleck Vega and Elhorst [28], the spillover effect cannot be observed in some spatial dependence models, such as the SEM (Table 18). The spillover effect is also absent in the basic OLS model because of the implicit assumption that outcomes for different observations are independent of each other. In a similar way to that used by the authors, we present these effects, for the models used in this paper:

Table 18. Direct and spillover effects corresponding to different model specifications.

Spatial model	Direct effect	Indirect/spillover effect
Basic OLS model (Equation 8) $y = \alpha\iota_n + \mathbf{X}\beta + \varepsilon$	$\hat{\beta}$	0
SEM (Equation 10) $y = \alpha\iota_n + \mathbf{X}\beta + \lambda\mathbf{W}u + \varepsilon$	$\hat{\beta}$	0
SLX model (Equation 16) $y = \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + \varepsilon$	$\hat{\beta}$	$\hat{\theta}$
Spatial lag (SAR) model (Equation 9) $y = \rho\mathbf{W}y + \alpha\iota_n + \mathbf{X}\beta + \varepsilon$	Mean diagonal elements: $(I - \hat{\rho}\mathbf{W})^{-1}\hat{\beta}$	Mean off-diagonal elements: $(I - \hat{\rho}\mathbf{W})^{-1}\hat{\beta}$
SDM (Equation 15) $y = \rho\mathbf{W}y + \alpha\iota_n + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\theta + \varepsilon$	Mean diagonal elements: $(I - \hat{\rho}\mathbf{W})^{-1}(\hat{\beta} + \mathbf{W}\hat{\theta})$	Mean off-diagonal elements: $(I - \hat{\rho}\mathbf{W})^{-1}(\hat{\beta} + \mathbf{W}\hat{\theta})$

Source: Elaborated from Halleck Vega and Elhorst [28], Table 1.

Table 19 presents the R code needed to calculate the direct, indirect, and total effect matrices, as well as the inference and the impact matrix of an explanatory variable for the SDM model, in the case of the model of urban economic growth in Spain. The following four libraries are required to run this code: : "sp", "spdep", "spatialreg" *and* "coda", *which provides functions for summarizing and plotting the output from Markov Chain Monte Carlo (MCMC) simulations.* The main functions, not previously presented, involved in this R code are "*as_dgRMatrix_listw*", "*trW*", "*set.seed*", "*impacts*", "*HPDinterval*", "*rep*", "*names*", "*matrix*" "*diag*" and "*solve*".

Table 19. R Code: Effects, inference and a variable impact matrix of a SDM.

```r
load("gdpmap.rdata")
# Constructs the row-standardized minimum distance matrix (like in Table 13)
# ML estimation of the SDM (like in Table 16)
# Calculation of effects for SDM model (LeSage & Pace)
set.seed(1234) # Simulations for the inferential process
gdpsdm_impacts <- impacts(gdp_sdm, tr=trMat, R=1000)
summary(gdpsdm_impacts, zstats=TRUE, short=TRUE)
HPDinterval(gdpsdm_impacts, choice="direct")
HPDinterval(gdpsdm_impacts, choice="indirect")
HPDinterval(gdpsdm_impacts, choice="total")
plot(gdpsdm_impacts, choice="direct")
plot(gdpsdm_impacts, choice="indirect")
plot(gdpsdm_impacts, choice="total")
plot(gdpsdm_impacts, trace=TRUE, density=FALSE, choice="direct")
plot(gdpsdm_impacts, trace=TRUE, density=FALSE, choice="indirect")
plot(gdpsdm_impacts, trace=TRUE, density=FALSE, choice="total")
# Calculation of the impact matrix for the variable LGH85
clear.pr <- rep(NA, dim(gdpmap)[1])
names(clear.pr) <- gdpmap$MUNICIPIO
svec <- rep(0,dim(gdpmap)[1])
eye <- matrix(0,nrow=dim(gdpmap)[1],ncol=dim(gdpmap)[1])
diag(eye) <- 1
for(i in 1:length(clear.pr)){
  cvec <- svec
  cvec[i] <- 1
  res <- solve(eye - gdp_sar[["rho"]]*Wsp) %*%
cvec*gdp_sar[["coefficients"]][["LGH85"]]
  clear.pr[i] <- res[i]
}
mult <- solve(eye - gdp_sar[["rho"]]*Wsp)
deriv.LGH85 <- solve(eye -
gdp_sar[["rho"]]*Wsp)*gdp_sar[["coefficients"]][["LGH85"]]
```

as_dgRMatrix_listw {spatialreg} converts a weights list to a sparse matrix class object. **trW {spatialreg}** is used to prepare a vector of traces of powers of a spatial weight matrix. **set.seed {base}** is the recommended way to specify seeds. **impacts {spatialreg}** calculates the impacts for spatial lag

(SAR) and SDM. **HPDinterval {coda}** Create Highest Posterior Density (HPD) intervals for the parameters in MCMC sample. **rep {base}** replicates elements of vectors and lists. **names {base}** gets or sets the names of an object. **matrix {base}** creates a matrix from the given set of values. **diag {base}** extract or replace the diagonal of a matrix or construct a diagonal matrix. **solve {base}** solves a system of equations.

The output of the calculation of the effects for the SDM model is presented in Table 20.

Table 20. Output of the effects and inference of the SDM of urban growth in Spain.

```
## Impact measures (mixed, trace):
##             Direct       Indirect        Total
## LGH85 -4.285846e-02 -2.016259e-02 -6.302105e-02
## BANK   6.875336e-05  1.066622e-05  7.941958e-05
## UNI01  5.368490e-04  6.784563e-05  6.046946e-04
## PAT00  5.116346e-02  3.877536e-01  4.389171e-01
##
## Simulated standard errors
##             Direct       Indirect        Total
## LGH85 4.430020e-03 1.144304e-02 0.0111533399
## BANK  1.295476e-05 5.361165e-05 0.0000541133
## UNI01 8.414860e-05 2.931461e-04 0.0002956689
## PAT00 1.824064e-02 9.927590e-02 0.1044755706
##
## Simulated z-values:
##          Direct   Indirect     Total
## LGH85 -9.722559 -1.8058825 -5.714514
## BANK   5.292804  0.1470840  1.412822
## UNI01  6.386995  0.2694986  2.084964
## PAT00  2.837808  3.9729278  4.270658
##
## Simulated p-values:
##        Direct    Indirect   Total
## LGH85 < 2.22e-16 0.070937   1.1002e-08
## BANK  1.2045e-07 0.883066   0.157708
## UNI01 1.6918e-10 0.787546   0.037073
## PAT00 0.0045424  7.0995e-05 1.9490e-05
```

Table 20 shows that the model estimates are -practically all- statistically significant and, except in the case of the LGH85 variable, they are positive. The negative sign of the coefficient of LGH85 demonstrates the existence of income convergence in the group of large Spanish cities. The total impact of a 10% growth in GDP per capita in a city in the initial period (1985) implied a fall in the average rate of change of GDP per capita in the period 1985-2003 of -0.63% in that city. This impact is the sum of the direct effect caused by the growth of GDP per capita in the city itself (-0.43), which is the direct effect, and the indirect effect coming from the growth of GDP per capita in the rest of the cities (-0.20). Additionally, the total effect of the growth of 1 patent per inhabitant led to a growth of GDP per capita in the period of 0.44%, of which 0.05% came from the growth of patents per capita in

the city itself (direct effect) and the remaining 0.39% was caused indirectly by the growth of patents in the rest of the cities.

4. Conclusions

This paper introduces the spatial component in econometric estimation and, in particular, the spatial dependence effect inherent in some of the variables involved in the modelling process, first, the spatial structure of the data is observed from maps, and the Moran's spatial autocorrelation indicator is presented. Then, the spatial weights matrix is constructed under different specifications. Finally, the taxonomy of spatial econometric models is shown, the hybrid specification strategy is presented, and the interpretation of the estimated coefficients ends.

Author Contributions: Conceptualization, A.V. and C.C; methodology, A.V. and C.C.; software, A.V. and C.C.; validation, A.V. and C.C.; formal analysis, A.V. and C.C.; investigation, A.V. and C.C.; resources, A.V. and C.C.; data curation, A.V. and C.C; writing—original draft preparation, A.V and C.C..; writing—review and editing, A.V. and C.C.; visualization, A.V. and C.C. Authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The data and scripts are available in the Spatial and Regional Economics research group (ECONRES) of the Universidad Autónoma de Madrid and the following repository: [6].

Conflicts of Interest: The authors declare no conflicts of interest.

Appendices

A.1. R Code: Spatial Autocorrelation Patterns (Figure 4)

```r
# Spatial autocorrelation patterns
library(ggplot2)
library(gridExtra)
positive <- ggplot() +
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 4, ymax = 5),fill="white",color="black")+
  annotate("text",x=1.5,y=-.4,label="Positive",size=6)+
  theme_void()
negative <- ggplot() +
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 0, ymax = 1),fill="white",color="black")+
```

```r
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 3, ymax = 4),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 3, ymax = 4),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 3, ymax = 4),fill="black",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 4, ymax = 5),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 4, ymax = 5),fill="black",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 4, ymax = 5),fill="white",color="black")+
  annotate("text",x=1.5,y=-.4,label="Negative",size=6)+
  theme_void()
random <- ggplot() +
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 0, ymax = 1),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 0, ymax = 1),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 1, ymax = 2),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 1, ymax = 2),fill="black",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 2, ymax = 3),fill="black",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 2, ymax = 3),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 3, ymax = 4),fill="black",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 3, ymax = 4),fill="white",color="black")+
  geom_rect(aes(xmin = 0, xmax = 1, ymin = 4, ymax = 5),fill="black",color="black")+
  geom_rect(aes(xmin = 1, xmax = 2, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 4, ymax = 5),fill="white",color="black")+
  geom_rect(aes(xmin = 3, xmax = 4, ymin = 4, ymax = 5),fill="black",color="black")+
  geom_rect(aes(xmin = 4, xmax = 5, ymin = 4, ymax = 5),fill="white",color="black")+
  annotate("text",x=1.5,y=-.4,label="Random",size=6)+
  theme_void()
grid.arrange(positive,negative,random,ncol=3)
```

References

1. Lefebvre, H. *The Production of Space*; Wiley-Blackwell: Oxford (UK) & Cambridge (USA), 1992; ISBN 978-0-631-18177-4.
2. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234–240, doi:10.2307/143141.

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

4. Friendly, M. A.-M. Guerry's "Moral Statistics of France": Challenges for Multivariable Spatial Analysis. *Stat. Sci.* **2007**, *22*, 368–399.

5. Anselin, L. A Local Indicator of Multivariate Spatial Association: Extending Geary's c. *Geogr. Anal.* **2019**, *51*, 133–150, doi:10.1111/gean.12164.

6. Chasco, C.; Vallone, A. Introduction to Cross-Section Spatial Econometric Models with Applications in R [Data and Codes] 2023, B2SHARE, EUDAT Repository, EU, 1 file, 840.4 kB. https://b2share.eudat.eu. doi:10.23728/b2share.4066171faa9b48329daf3c03de04305e

7. Monmonier, M. *How to Lie with Maps, Third Edition*; University of Chicago Press: Chicago, IL, 2018; ISBN 978-0-226-43592-3.

8. Anselin, L. *Spatial Econometrics: Methods and Models*; Springer Science & Business Media, 1988; ISBN 978-90-247-3735-2.

9. Martori, J.C.; Hoberg, K.; Madariaga, R. La Incorporación Del Espacio En Los Métodos Estadísticos: Autocorrelación Espacial y Segregación. In Proceedings of the Conference: Actas del X Coloquio Internacional de Geocrítica; Barcelona, Spain, May 30 2008.

10. Bivand, R.S.; Wong, D.W.S. Comparing Implementations of Global and Local Indicators of Spatial Association. *TEST* **2018**, *27*, 716–748, doi:10.1007/s11749-018-0599-x.

11. Vallone, A.; Chasco, C. Spatiotemporal Methods for Analysis of Urban System Dynamics: An Application to Chile. *Ann. Reg. Sci.* **2020**, *64*, 421–454, doi:10.1007/s00168-019-00960-9.

12. Pebesma, E.; Bivand, R.; Racine, E.; Sumner, M.; Cook, I.; Keitt, T.; Lovelace, R.; Wickham, H.; Ooms, J.; Müller, K.; et al. Sf: Simple Features for R 2023.

13. Anselin, L.; Bera, A.K.; Florax, R.; Yoon, M.J. Simple Diagnostic Tests for Spatial Dependence. *Reg. Sci. Urban Econ.* **1996**, *26*, 77–104, doi:10.1016/0166-0462(95)02111-6.

14. Chasco, C. *Econometría Espacial Aplicada a La Predicción-Extrapolación de Datos Microterritoriales.*; Consejería de Economía e Innovación Tecnológica, Comunidad de Madrid.; Madrid, 2003; ISBN 978-84-451-2442-0.

15. Anselin, L. Local Indicators of Spatial Association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115, doi:10.1111/j.1538-4632.1995.tb00338.x.

16. Mella, J.M.; Chasco, C. Urban Growth and Territorial Dynamics: A Spatial-Econometric Analysis of Spain. In *Spatial Dynamics, Networks and Modelling*; Edward Elgar Publishing, 2006 ISBN 978-1-78100-747-1.

17. Anselin, L.; Rey, S.J. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*; GeoDa Press LLC, 2014; ISBN 978-0-9863421-0-3.

18. Anselin, L.; Bera, A.K.; Florax, R.; Yoon, M.J. Simple Diagnostic Tests for Spatial Dependence. *Reg. Sci. Urban Econ.* **1996**, *26*, 77–104, doi:10.1016/0166-0462(95)02111-6.

19. Mur, J.; Angulo, A. Model Selection Strategies in a Spatial Setting: Some Additional Results. *Reg. Sci. Urban Econ.* **2009**, *39*, 200–213, doi:10.1016/j.regsciurbeco.2008.05.018.

20. Kelejian, H.H.; Robinson, D.P. A Suggested Method of Estimation for Spatial Interdependent Models with Autocorrelated Errors, and an Application to a County Expenditure Model. *Pap. Reg. Sci.* **1993**, *72*, 297–312, doi:10.1111/j.1435-5597.1993.tb01878.x.

21. Arraiz, I.; Drukker, D.M.; Kelejian, H.H.; Prucha, I.R. A Spatial Cliff-Ord-Type Model with Heteroskedastic Innovations: Small and Large Sample Results*. *J. Reg. Sci.* **2010**, *50*, 592–614, doi:10.1111/j.1467-9787.2009.00618.x.

22. Kelejian, H.H.; Prucha, I.R. A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *J. Real Estate Finance Econ.* **1998**, *17*, 99–121, doi:10.1023/A:1007707430416.

23. Kelejian, H.H.; Prucha, I.R. Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances. *J. Econom.* **2010**, *157*, 53–67, doi:10.1016/j.jeconom.2009.10.025.

24. Manski, C.F. Identification of Endogenous Social Effects: The Reflection Problem. *Rev. Econ. Stud.* **1993**, *60*, 531–542, doi:10.2307/2298123.

25. LeSage, J.; Pace, R.K. *Introduction to Spatial Econometrics*; Chapman and Hall/CRC: New York, 2009; ISBN 978-0-429-13808-9.

26. Chasco, C. Geodaspace: A Resource For Teaching Spatial Regression Models. *Rect Rev. Electrónica Comun. Trab. ASEPUMA* **2013**, *4*, 119–144.

27. Florax, R.; Folmer, H. Specification and Estimation of Spatial Linear Regression Models: Monte Carlo Evaluation of Pre-Test Estimators. *Reg. Sci. Urban Econ.* **1992**, *22*, 405–432, doi:10.1016/0166-0462(92)90037-2.

28. Halleck Vega, S.; Elhorst, J.P. The SLX Model. *J. Reg. Sci.* **2015**, *55*, 339–363, doi:10.1111/jors.12188.

29. Elhorst, J.P. Applied Spatial Econometrics: Raising the Bar. *Spat. Econ. Anal.* **2010**, *5*, 9–28, doi:10.1080/17421770903541772.

30. Debarsy, N.; Gallo, J.L. Identification, Causality, and Spatial Econometrics. *Post-Print* 2022.