

# Análisis de datos espaciales

Gustavo A. García

[ggarci24@eafit.edu.co](mailto:ggarci24@eafit.edu.co)

Econometría Espacial

PhD/Maestría en Economía

Universidad EAFIT

Link slides en formato **html**

Link slides en formato **PDF**

## En este tema

- Principios básicos en el tratamiento de datos espaciales
- Efectos espaciales
- Heterogeneidad espacial
- Autocorrelación o dependencia espacial
- Matriz de pesos espaciales
- Análisis exploratorio de datos en el espacio
- Análisis confirmatorio de datos espaciales
- Ejercicio aplicado en R

# Lecturas

- Elhorst, J.P. (2010). "Applied Spatial Econometrics: Raising the Bar". *Spatial Economic Analysis*, 5(1):9–28
- Millo, G. y Piras, G. (2012). "splm: Spatial Panel Data Models in R". *Journal of Statistical Software*, 47(1):1–37
- Elhorst, J.P. (2014). *Spatial Econometrics from Cross-Sectional Data to Spatial Panels*, Springer
- LeSage, J. y Pace, R. (2014). "Interpreting spatial econometrics models". En Fischer, M. y Nijkamp, P. (Eds.), *Handbook of Regional Science*, Springer
- Halleck Vega, S. y Elhorst, J.P. (2015). "The SLX model", *Journal of Regional Science*, 55(3):339–363
- Golgher, A. y Voss, P. (2016). "How to interpret the coefficients of spatial models: spillovers, direct and indirect Effects". *Spatial Demography*, 4:175–205
- Belotti, F., Hughes, G. y Mortari, A. (2017). "Spatial panel-data models using Stata", *The Stata Journal*, 17(1):139–180.

# Principios básicos en el tratamiento de datos espaciales

Paelinck y Klaassen (1979) destacan cinco principios básicos en el campo de la econometría espacial y el tratamiento de datos de corte transversal en general:

- **Interdependencia**: todo modelo espacial ha de caracterizarse por su interdependencia, es decir, deben incorporarse relaciones mutuas entre las observaciones de las variables económicas, sociales, demográficas, etc.
- **Asimetría**: las relaciones espaciales son, en principio, asimétricas
- **Alotopía**: se ha de buscar a priori "la causa" de un fenómeno espacial en otro lugar
- **No linealidad**: la no linealidad de soluciones espaciales óptimas *ex-ante* conduce a modelos econométricos *ex-post* que requieren una atención particular en lo que respecta a su especificación, lo cual generalmente será no lineal
- **Inclusión de variables topológicas**: dado que la vida económica se desarrolla necesariamente en el espacio geográfico, un modelo espacial debe incorporar variables topológicas: coordenadas, distancias, superficies, densidades, etc.

De acuerdo a Paelinck y Klaassen (1979), no siempre será posible observar estos cinco principios de construcción de modelos espaciales y probablemente pueden haber otros además de los aquí especificado

# Efectos espaciales

Causas de la dependencia espacial:

- la delimitación arbitraria de las unidades espaciales de observación (ejemplo, zonas censales, límites municipales, departamentales...)
- problemas de agregación espacial
- la presencia de externalidades y efectos de desbordamiento

Causa de la heterogeneidad espacial:

- falta de estabilidad en el espacio del comportamiento u otras relaciones bajo estudio
- esto implica que la forma funcional y los parámetros varían con la localización y no son homogéneos en los datos
- puede ocurrir al estimar modelos econométricos con datos de sección cruzada de unidades espaciales diferentes, como regiones ricas y pobres

La heterogeneidad espacial se puede tratar con la econometría estándar que tenga en cuenta la inestabilidad estructural

# Heterogeneidad espacial

- **Definición:** se refiere a la variación en las relaciones en el espacio
- Aspectos de la heterogeneidad espacial:
  - la inestabilidad estructural: falta de estabilidad en el espacio del comportamiento de la variables bajo estudio. La forma funcional y los parámetros de una regresión pueden variar según la localización, por tanto, no son homogéneos en toda la muestra
  - la heterocedasticidad: proviene de la omisión de variables u otras formas de error de especificación que llevan a la aparición de errores de medida
- La heterogeneidad espacial puede tratarse por medio de las técnicas econométricas estándar, en concreto:
  - parámetros variantes, coeficientes aleatorios (Hildreth y Houck, 1968)
  - *Switching regressions* (Quant, 1958)
  - técnicas de filtraje adaptativo espacial (Foster y Gorr, 1983)
  - expansión espacial de parámetros (Casetti, 1972)
  - regresiones ponderadas geográficamente (Fotheringham et al., 1998)

# Autocorrelación o dependencia espacial

- **Definición:** aparece como consecuencia de la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar
- El valor que toma una variable en una región no viene explicado únicamente por condicionantes internos sino también por el valor de esa misma variable en otras regiones vecinas, incumpléndose por tanto el supuesto de independencia entre las observaciones muestrales
- La autocorrelación espacial puede ser positiva o negativa
- Positiva: la presencia de un fenómeno determinado en una región lleva a que se extienda ese mismo fenómeno hacia el resto de regiones que la rodean, favoreciendo así la concentración del mismo
- Negativa: cuando la presencia de un fenómeno en una región impida o dificulte su aparición en las regiones vecinas a ella, es decir, cuando unidades geográficas cercanas sean netamente más disímiles entre ellas que entre regiones alejadas en el espacio (tablero de ajedrez)
- Cuando la variable analizada se distribuye de forma aleatoria, no existirá autocorrelación espacial



# Autocorrelación o dependencia espacial

**Causas:** la existencia de errores de medida y fenómenos de interacción espacial

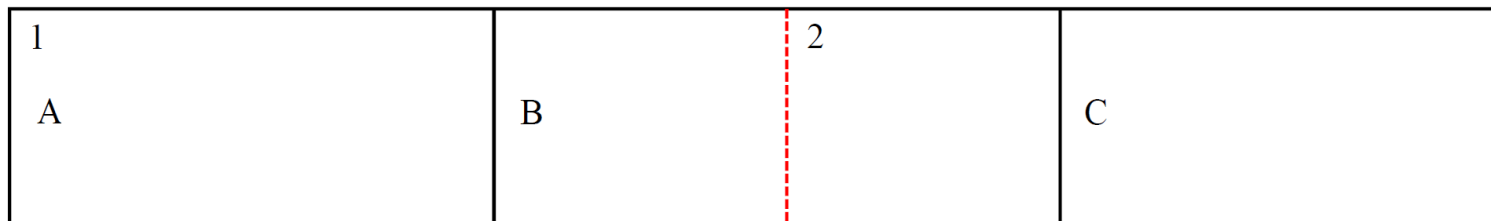
Errores de medida: pueden surgir, entre otros aspectos, como consecuencia de una escasa correspondencia entre la extensión espacial del fenómeno económico bajo estudio y las unidades espaciales de observación

Ejemplo:

La correcta delimitación espacial de una variables  $x$  corresponde a las áreas ABC

Las observaciones disponibles de  $x$  son agregadas a nivel espacial en dos niveles: 1 y 2

Consecuencias:  $x_1$  observada contendrá a  $x_A$  y parte de  $x_B$ , al tiempo que  $x_2$  contendrá a  $x_C$  y parte de  $x_B$



Resultado:  $x$  estará correlacionada espacialmente aunque de forma espuria

Interacción espacial: también entendido como efectos de desbordamiento y de jerarquías espaciales. La existencia de efectos desbordamiento de las infraestructuras de transporte o la difusión tecnológica entre economías son ejemplos claros de fenómenos que favorecen la aparición de interdependencias entre unidades espaciales

# Autocorrelación o dependencia espacial

- Es posible detectar cierta similitud entre los conceptos de autocorrelación espacial y temporal en la medida en que, en ambos casos, se produce un incumplimiento de la hipótesis de independencia entre las observaciones muestrales
- Diferencias entre estos dos tipos de autocorrelaciones:
  - la dependencia temporal es únicamente unidireccional  $\implies$  el pasado explica el presente
  - la dependencia espacial es multidireccional  $\implies$  una región puede no sólo estar afectada por otra región contigua a ella sino por otras muchas que la rodean, al igual que ella puede influir sobre aquéllas
- La multidireccionalidad de la dependencia espacial imposibilita la utilización del operador de retardos  $L$ ,  $L^p x_t = x_{t-p}$ , ya que recoge sólo únicamente una relación unidireccional

# Matriz de pesos espaciales (W)

La solución al problema de la multidireccionalidad en el contexto espacial para por la definición de la denominada **matriz de pesos espaciales, de retardos o de contactos  $\mathbf{W}$** :

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & 0 & \dots & w_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ w_{N1} & w_{N2} & \dots & 0 \end{bmatrix}$$

una matriz cuadrada no estocástica cuyos elementos  $w_{ij}$  reflejan la intensidad de la interdependencia existente entre cada par de regiones  $i$  y  $j$

No existe una definición de  $\mathbf{W}$  unánimemente aceptada, si bien se ha de cumplir que dichos pesos sean no negativos y finitos (Anselin, 1980)

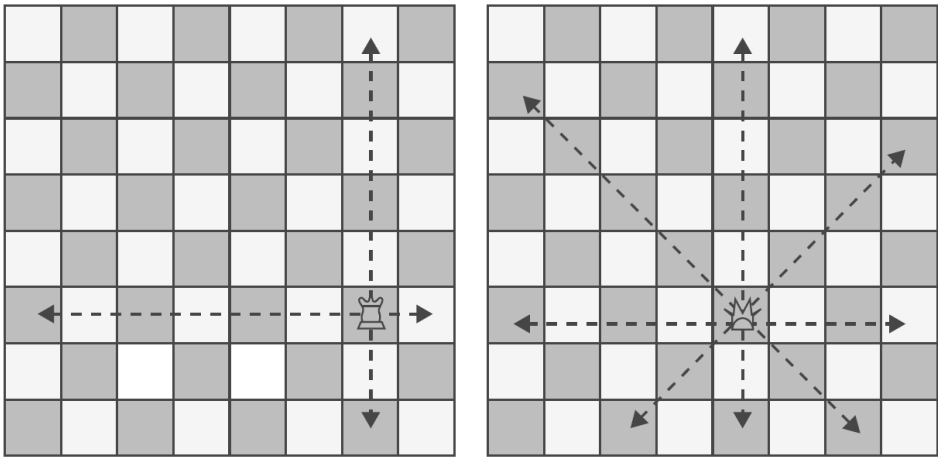
De forma habitual se recurre al concepto de congüidad física de primer orden, donde  $w_{ij}$  es igual a 1 si las regiones  $i$  y  $j$  son físicamente adyacentes o a 0 en caso contrario (se asume por definición que  $w_{ii} = 0$ )

# Matriz de pesos espaciales (W)

Existen diversos criterios para la identificación de las regiones vecinas:

Cuadro 1. Criterios de contigüidad física en una cuadrícula regular

Criterio de vecindad	Número total de vecinos	Definición
Criterio lineal	2	Serán vecinas de $i$ las regiones que comparten el lado izquierdo o derecho de $i$
Criterio torre o <i>rook</i>	4	Serán vecinas de $i$ las regiones que comparten algún lado con $i$
Criterio alfil o <i>bishop</i>	4	Serán vecinas de $i$ las regiones que comparten algún vértice con $i$
Criterio reina o <i>queen</i>	8	Serán vecinas de $i$ las regiones que comparten algún lado o vértice con $i$



# Matriz de pesos espaciales ( $W$ )

Limitaciones de la matriz  $W$ :

- es simétrica, no siendo posible incorporar influencias no recíprocas, violando el segundo de los cinco principios básicos de la econometría espacial
- considera la adyacencia física como único determinante de las interdependencias regionales, descuidando con ello, por ejemplo, posibles influencias mutuas entre regiones, que, aun estando alejadas, mantienen estrechas relaciones comerciales

# Matriz de pesos espaciales (W)

Definiciones de **W** basadas en la utilización de la distancia entre regiones:

- Cliff y Ord (1973, 1981):  $w_{ij} = d_{ij}^{-a} \beta_{ij}^b$   
 $d_{ij}$ : distancia entre  $i$  y  $j$   
 $\beta_{ij}$ : longitud relativa de la frontera común entre  $i$  y  $j$  con relación al perímetro de  $i$   
 $a$  y  $b$ : parámetros a estimar
- Dacey (1968):  $w_{ij} = \gamma_{ij} \beta_{ij} \alpha_i$   
 $\beta_{ij}$ : igual que antes  
 $\gamma_{ij}$ : es un factor de contigüidad binario  
 $\alpha_i$ : es el área de la región  $i$  en relación al área total del sistema
- Anselin (1980): matriz inversa de distancias al cuadrado, de manera que la intensidad de la interdependencia entre dos regiones disminuye con la distancia que separa sus respectivos centros
- Bodson y Peeters (1975): función logística que mide la influencia de varios canales de comunicación entre regiones como podrían ser las carreteras, el ferrocarril y otros medios de transporte

$$w_{ij} = \sum_{n=1}^N K_N \left\{ \frac{a}{1 + b e^{-c_j d_{ij}}} \right\}$$

$K_N$ : la importancia relativa del medio de comunicación  $n$

$d_{ij}$ : la distancia entre dos regiones  $i$  y  $j$   $N$ : el número de medios de comunicación

$a$ ,  $b$  y  $c_j$ : parámetros a estimar

# Matriz de pesos espaciales (W)

Definiciones de **W** basadas en la utilización de la distancia entre regiones:

- Case et al. (1993): **W** basada en distancias económicas,  $w_{ij} = \frac{1}{|x_i - x_j|}$   $x_i$  y  $x_j$ : observaciones de características socioeconómicas, tales como la renta per capita
- Vayá et al. (1998a, 1998b) y López-Bazo et al. (1999): **W** recoge el grado de intercambio comercial entre regiones analizadas

Otras consideraciones sobre **W**:

- La matriz de pesos debe ser exógena
- Estandarización de la matriz **W**
  - se divide cada elemento  $w_{ij}$  por la suma total de la fila a la que pertenece, de forma que la suma de cada fila será igual a la unidad

$$w_{ij}^* = \frac{w_{ij}}{\sum_j^n w_{ij}}$$

- la posibilidad de ponderar por igual la influencia total que recibe cada región de sus vecinas, con independencia del número total de vecinos de cada una de ellas, explicaría dicha transformación
- Anselin (1988) plantea que la estandarización de **W** no es siempre adecuada, especialmente cuando ésta se basa en un concepto de distancia dado que, en este caso, la matriz estandarizada carecería de significado

# Matriz de pesos espaciales (W)

Matriz de pesos como instrumento para recoger las interdependencias

Un ejemplo de la matriz **W**:



Matriz **W**

	1	2	3	4	5	6	7	8	9	10	11	12	
	North East	North West	Yorksh-Humber	East Midlands	West Midlands	East of England	London	South East	South West	Wales	Scotland	Northern Ireland	Row sum
1 North East	0	1	1	0	0	0	0	0	0	0	1	0	3
2 North West	1	0	1	1	1	0	0	0	0	1	1	0	6
3 Yorkshire and The Humber	1	1	0	1	0	0	0	0	0	0	0	0	3
4 East Midlands	0	1	1	0	1	1	0	1	0	0	0	0	5
5 West Midlands	0	1	0	1	0	0	0	1	1	1	0	0	5
6 East of England	0	0	0	1	0	0	1	1	0	0	0	0	3
7 London	0	0	0	0	0	1	0	1	0	0	0	0	2
8 South East	0	0	0	1	1	1	1	0	1	0	0	0	5
9 South West	0	0	0	0	1	0	0	1	0	1	0	0	3
10 Wales	0	1	0	0	1	0	0	0	1	0	0	0	3
11 Scotland	1	1	0	0	0	0	0	0	0	0	0	0	2
12 Northern Ireland	0	0	0	0	0	0	0	0	0	0	0	0	0

Matriz **W** estandarizada (**W\***)

	1	2	3	4	5	6	7	8	9	10	11	12	
	North East	North West	Yorksh-Humber	East Midlands	West Midlands	East of England	London	South East	South West	Wales	Scotland	Northern Ireland	Row sum
1 North East	0	0.3333	0.3333	0	0	0	0	0	0	0	0.3333	0	1
2 North West	0.1667	0	0.1667	0.1667	0.1667	0	0	0	0	0.1667	0.1667	0	1
3 Yorkshire and The Humber	0.3333	0.3333	0	0.3333	0	0	0	0	0	0	0	0	1
4 East Midlands	0	0.2	0.2	0	0.2	0.2	0	0.2	0	0	0	0	1
5 West Midlands	0	0.2	0	0.2	0	0	0	0.2	0.2	0.2	0	0	1
6 East of England	0	0	0	0.3333	0	0	0.3333	0.3333	0	0.2	0	0	1
7 London	0	0	0	0	0	0.5	0	0.5	0	0	0	0	1
8 South East	0	0	0	0.2	0.2	0.2	0.2	0	0.2	0	0	0	1
9 South West	0	0	0	0	0.3333	0	0	0.3333	0	0.3333	0	0	1
10 Wales	0	0.3333	0	0	0.3333	0	0	0	0.3333	0	0	0	1
11 Scotland	0.5	0.5	0	0	0	0	0	0	0	0	0	0	1
12 Northern Ireland	0	0	0	0	0	0	0	0	0	0	0	0	0



# Matriz de pesos espaciales (W)

## El operador de retardo espacial

El retardo espacial de una variable resulta del producto de la matriz  $W$  y la variable que se quiere retardar espacialmente:

$$L(y) = \mathbf{W}^* y = \sum_j^n w_{ij}^* y_j$$

$\mathbf{W}^* =$ 

0	0.3333	0.3333	0	0	0	0	0	0	0	0.3333
0.1667	0	0.1667	0.1667	0.1667	0	0	0	0	0.1667	0.1667
0.3333	0.3333	0	0.3333	0	0	0	0	0	0	0
0	0.2	0.2	0	0.2	0.2	0	0.2	0	0	0
0	0.2	0	0.2	0	0	0	0.2	0.2	0.2	0
0	0	0	0.3333	0	0	0.3333	0.3333	0	0	0
0	0	0	0	0	0.5	0	0.5	0	0	0
0	0	0	0.2	0.2	0.2	0.2	0	0.2	0	0
0	0	0	0	0.3333	0	0	0.3333	0	0.3333	0
0	0.3333	0	0	0.3333	0	0	0	0.3333	0	0
0.5	0.5	0	0	0	0	0	0	0	0	0

$y$	$\mathbf{W}^* y$
86.2	90.06667
88.6	87.93333
84.7	88.00000
89.2	93.50000
89.1	91.48000
96.8	112.40000
139.7	102.55000
108.3	100.92000
89.8	92.96667
81.5	89.16667
96.9	87.40000

# Análisis exploratorio de datos en el espacio

- El análisis exploratorio de datos espaciales (ESDA por sus siglas en inglés) se centra de forma explícita en los efectos espaciales:
  - identificar localizaciones atípicas (*outliers* espaciales)
  - descubrir esquemas de asociación espacial (*cluster* espacial)
  - sugerir diferentes regímenes espaciales u otras formas de inestabilidad espacial
- El centro de este concepto lo ocupa la noción de autocorrelación espacial, es decir, el fenómeno por el cual la similitud locacional (observaciones con proximidad espacial) se une con la similitud de valores (correlación de atributos)
- Dimensiones del ESDA:
  - distinción entre indicadores globales y locales de asociación espacial
  - distinción entre los estadísticos basados en la vecindad y la distancia

# Análisis exploratorio de datos en el espacio

## Indicadores globales de asociación espacial

- La dependencia espacial se resume en un sólo indicador
- Suelen utilizarse para conocer el rango de interacción espacial en los datos
- Estadísticos: I de Moran y C de Geary

## Indicadores locales de asociación espacial (LISA por sus siglas en inglés)

- Un LISA es un indicador que consigue dos objetivos:
  - que el valor del estadístico obtenido para cada observación suministre información acerca de la relevancia de una agrupación espacial de valores similares alrededor de la misma
  - que la suma del valor del estadístico para todas las observaciones sea proporcional a un indicador global de asociación espacial
- Los LISA resultan fáciles de interpretar mediante la visualización en un mapa

# Análisis exploratorio de datos en el espacio

Modelos de datos en los cuales la autocorrelación espacial puede ser analizada:

- datos geoestadísticos
  - datos puntuales como una muestra de una distribución continua subyacente
  - se asume que la interacción espacial es una función suave de la distancia entre pares de observaciones
- datos *lattice*
  - una colección fija de localizaciones espaciales discretas (puntos o polígonos)
  - la interacción espacial se entiende como una función a pasos según la cual una localización interactúa con un grupo dado de vecinos
  - esta perspectiva es la más comúnmente seguida en la estadística espacial y ciencias sociales

# Análisis confirmatorio de datos espaciales

- El análisis confirmatorio trata los datos espaciales desde una perspectiva de **modelización** y está constituido por los distintos métodos de **estimación**, **contrastes de especificación** y procedimientos de validación necesarios para implementar **modelos multivariantes** en los que las observaciones son de corte transversal y están georeferenciados
- Tradicionalmente, el modelo suele estimarse en un primer momento sin incorporar ningún tipo de efecto espacial, de forma que los resultados de la estimación del mismo (y especialmente los residuos) sean el punto de partida de los diagnósticos de dependencia espacial
- Idealmente estos diagnósticos apuntan hacia la dirección correcta en que debe introducirse dicha dependencia espacial en el modelo
- **Autocorrelación espacial residual**: cuando se deduce la existencia de autocorrelación residual, se reespecifica el término de error con el objetivo de incorporar dicha estructura de dependencia espacial en el mismo
- **Autocorrelación espacial sustantiva**: en este caso se procede a incorporar la variable dependiente retardada espacialmente como una variable explicativa más en el modelo

# Análisis confirmatorio de datos espaciales

- La estimación de tales modelos debe realizarse mediante métodos basados en el principio máximo verosímil o en el método genral de los momentos, entre otros
- Una vez hecha la estimación se utilizan diagnósticos y otros procedimientos de validación a fin de seleccionar el más adecuado
- Este conjunto de estadísticos y métodos crean el cuerpo de lo que se conoce como **econometría espacial**

# Ejercicio aplicado en R

En este ejercicio se hace una aplicación de la construcción de la matrix  $W$  utilizando los datos espaciales del United Kingdom (UK), analizando la contiguidad entre las 12 regiones que la compone. En los siguientes links se encuentran los datos, el shapefile y el código utilizado en R:

- Datos
- Shapefile
- Código en R