

**ECONOMETRÍA ESPACIAL APLICADA
A LA PREDICCIÓN-EXTRAPOLACIÓN DE DATOS
MICROTERRITORIALES.**

CORO CHASCO YRIGOYEN
Ganadora de la Quinta Edición de Premios
sobre Estudios de Economía Regional

CONSEJERÍA DE ECONOMÍA E
INNOVACIÓN TECNOLÓGICA
Comunidad de Madrid

**ECONOMETRÍA ESPACIAL APLICADA
A LA PREDICCIÓN-EXTRAPOLACIÓN DE DATOS
MICROTERRITORIALES.**

**ECONOMETRÍA ESPACIAL APLICADA
A LA PREDICCIÓN-EXTRAPOLACIÓN DE DATOS
MICROTERRITORIALES.**

**TESIS DOCTORAL
CORO CHASCO YRIGOYEN**

Comunidad de Madrid
CONSEJERÍA DE ECONOMÍA E INNOVACIÓN TECNOLÓGICA
Dirección General de Economía y Planificación



No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio ya sea eléctrico, mecánico, por fotocopia, por registro u otros métodos, ni su préstamo, alquiler o cualquier otra forma de cesión de uso del ejemplar, sin el permiso previo y por escrito de los titulares del Copyright.

Tirada: 400 ejemplares

Coste unitario: 36€

1ª Edición: Abril 2003

Copyright? 2003, by Comunidad de Madrid. Consejería de Economía e Innovación Tecnológica.

Autora: Coro Chasco Yrigoyen

Edita: Consejería de Economía e Innovación Tecnológica.

Príncipe de Vergara, 132. 28002 Madrid (España)

ISBN: 84-451-2442-0

Depósito Legal: M-19651-2003

Impreso en España – Printed in Spain

Imprime: Artes Gráficas Palermo.

Índice

Prólogo	XI
Consideraciones de la autora	XIII
CAPÍTULO I. PLANTEAMIENTO Y OBJETIVOS	1
1. Introducción y objetivos de la investigación	1
2. Estructura de la investigación	4
CAPÍTULO II. ECONOMETRÍA DE DATOS ESPACIALES	7
1. Econometría espacial	8
1.1. Concepto	8
1.2. Antecedentes	10
1.3. Aplicaciones	12
2. Datos espaciales	17
2.1. Naturaleza de los datos espaciales	17
2.2. Sistemas de Información Geográfica (GIS)	19
2.3. Fuentes estadísticas de información microterritorial	21
2.4. Problemas en el tratamiento de datos espaciales	23
CAPÍTULO III. ANÁLISIS EXPLORATORIO DE DATOS ESPACIALES ...	27
1. Introducción al análisis exploratorio de datos espaciales (AEDE)	28
1.1. Análisis exploratorio de datos	28
1.2. AEDE y paquetes informáticos especializados	30
1.3. Métodos gráficos del AEDE	36
1.3.1. Métodos de representación de distribuciones espaciales	37
1.3.2. Representación del fenómeno de dependencia espacial	39
1.3.3. Representación del fenómeno de heterogeneidad espacial	45
2. Dependencia o autocorrelación espacial	49
2.1. Concepto	49
2.2. Formulación matemática del fenómeno de dependencia espacial	53
2.2.1. Matriz de interacciones espaciales	53
2.2.2. Matriz de pesos espaciales	56
2.2.3. Operador retardo espacial	60
2.3. Perspectivas en el análisis de autocorrelación espacial: global y local	62
2.3.1. Contrastes de autocorrelación espacial global	63
2.3.2. Contrastes de autocorrelación espacial local	71

CAPÍTULO IV. ANÁLISIS CONFIRMATORIO: MODELOS DE REGRESIÓN ESPACIAL	79
1. Especificación de los efectos espaciales en modelos de regresión	80
1.1. Modelos de dependencia espacial o modelos espaciales dinámicos	81
1.1.1. Primera taxonomía de Anselin	82
1.1.2. Taxonomía de Florax y Folmer	86
1.1.3. Taxonomía de modelos de externalidades espaciales	94
1.2. Modelos de heterogeneidad espacial	104
1.2.1. Especificación de la heteroscedasticidad espacial	105
1.2.2. Modelos de inestabilidad paramétrica	108
2. Métodos de estimación y contraste de los efectos espaciales	117
2.1. Estimación y contrastes en el modelo básico de regresión lineal	118
2.1.1. Estimación por el método de mínimos cuadrados ordinarios ..	118
2.1.2. Contrastes de heteroscedasticidad	119
2.1.3. Contrastes de autocorrelación espacial	121
2.2. Estimación y contrastes en modelos de dependencia espacial	129
2.2.1. Mínimos cuadrados ordinarios y dependencia espacial	129
2.2.2. Modelo del retardo espacial	132
2.2.3. Modelo del error espacial	138
2.3. Estimación y contrastes en modelos de heterogeneidad espacial	141
2.3.1. Modelo del error heteroscedástico	141
2.3.2. Modelos de inestabilidad paramétrica continua	143
2.3.3. Modelo de espacial de regresiones cambiantes	144
 CAPÍTULO V. PREDICCIÓN-EXTRAPOLACIÓN ESPACIAL DE DATOS MICROTERRITORIALES	147
1. Principios básicos de la predicción-extrapolación espacial	148
1.1. Concepto de predicción-extrapolación espacial	148
1.2. El problema de la unidad espacial modificable (MAUP)	152
1.2.1. Problema de la escala	153
1.2.2. Problema de la agregación	157
1.3. Inferencias ecológicas (IE)	158
1.3.1. Concepto de IE	158
1.3.2. Modelos de IE	165
1.4. Efectos espaciales y el sesgo de la agregación	171
1.5. Problemas y vías de solución	173
2. Modelos de predicción-extrapolación de datos microterritoriales	178
3. Proceso de realización de un ejercicio de predicción-extrapolación de la renta familiar disponible municipal de la Comunidad de Madrid en 2000	185
3.1. Fundamentos teóricos	187
3.2. Selección y tratamiento de la información disponible	192
3.3. Análisis exploratorio espacial	198
3.3.1. Análisis exploratorio de datos municipales	198
3.3.2. Análisis exploratorio de datos provinciales	209
3.4. Análisis confirmatorio	220
3.5. Predicción-extrapolación de datos microterritoriales	227
3.6. Análisis y validación de los resultados	229
 CAPÍTULO VI. CONCLUSIONES Y REFLEXIÓN FINAL	237
1. Principales conclusiones y resumen de la investigación	237
2. Líneas futuras de investigación	248
 REFERENCIAS BIBLIOGRÁFICAS	251
 ANEXOS	275

Prólogo

Los estudios y análisis que lleva a cabo la Consejería de Economía e Innovación Tecnológica representan un eslabón más en la larga cadena de estudios y publicaciones que desde hace años lleva a cabo la Comunidad de Madrid para impulsar un mejor conocimiento de la economía, la sociedad y el territorio de nuestra región. Es indudable que las decisiones estratégicas que deben tomar las instituciones públicas y las empresas para enfrentar con éxito los retos del presente pueden verse apoyadas por la existencia de una amplia información y reflexión previas sobre la realidad a transformar. Por tanto, las obras que se publican pretenden tener utilidad tanto para estudiosos como para emprendedores y responsables públicos interesados en conocer, comprender y actuar en la región con el objetivo de impulsar su competitividad económica y el empleo, mejorando, al tiempo, el bienestar de sus habitantes. Se está fomentando, por tanto, la actividad investigadora en materia socioeconómica.

La administración por sí misma no puede, ni probablemente deba, desarrollar todos los estudios necesarios para conocer nuestra economía, en este sentido, es preciso sesgar la actividad investigadora de las trece universidades madrileñas. La Convocatoria de las sucesivas ediciones de Premios sobre Estudios de Economía Regional de la Comunidad de Madrid, cuya primera edición data de 1998, ha sido todo un éxito. Se ha potenciado con carácter de permanencia la realización de estudios e investigaciones desde uno de los principales motores del desarrollo de dicha realidad: el mundo Universitario, y más concretamente dentro de la Universidad, el sector de los jóvenes investigadores que en ella realizan sus estudios de doctorado.

Se está apoyando a la preparación para la inserción profesional de graduados de la Comunidad de Madrid, mediante la concesión de becas para períodos de prácticas de técnicos jóvenes graduados, con especialidad en Economía Regional, con objeto de que vean aumentadas sus posibilidades de capacitación profesional, a la vez que entren a conocer los objetivos y actividades propias de la Administración Regional, materias que, seguramente, les son de real utilidad para su inserción en el mercado laboral.

Durante el año 2002 se tramitó la Quinta Edición de Premios sobre Estudios de Economía Regional de la que resultó ganadora del primer premio D^a. María del Coro Chasco Irigoyen, con la tesis titulada “Econometría Espacial aplicada a la predicción-extrapolación de datos microterritoriales”. Obra que aborda un tema relacionado con la Comunidad de Madrid, en concreto, la predicción-extrapolación de la renta familiar disponible de los municipios madrileños, como un ejercicio no circunscrito únicamente a la obtención de unos resultados a partir de técnicas más o menos complejas, sino con el deseo también de aportar un mayor conocimiento de la realidad económica de la Comunidad de Madrid, evaluada con el máximo nivel de actualización posible (año 2000), desde el punto de vista del poder adquisitivo medio de los hogares en los 179 municipios de la región.

Resulta indudable la importancia de esta obra, en la medida en que constituye una interesante aportación que nos ayudará a comprender mejor nuestra realidad. Madrid se ha convertido en un centro industrial y empresarial de primer orden que, al concentrar la mayor oferta de formación e investigación de España, se asegura un puesto de vanguardia en las innovaciones tecnológicas y estructurales de este continente europeo.

LUIS BLÁZQUEZ TORRES
Consejero de Economía e
Innovación Tecnológica.

Consideraciones de la autora

Esta obra reproduce con algunas modificaciones mi tesis doctoral en ciencias económicas y empresariales, terminada de escribir en mayo de 2002. La principal diferencia entre dicha tesis y este libro es la incorporación de varias de las sugerencias aportadas por los miembros del tribunal durante la defensa de la tesis y el resumen del capítulo de anexos para adaptar el trabajo original a la presente edición.

Este estudio contribuye a una mayor profundización tanto en el campo científico de la econometría, como, en general, en las ciencias sociales, con una especial trascendencia para la economía de la Comunidad de Madrid. En lo que se refiere al ámbito científico, se propone la predicción-extrapolación de datos microterritoriales como parte de la econometría espacial, presentando una metodología fundamentada en instrumentos exploratorios y confirmatorios propios de esta ciencia. La predicción-extrapolación de datos geográficos podría definirse como el conjunto de métodos inferenciales destinados a obtener datos microterritoriales, no disponibles, a partir de información espacial agregada proporcionada por fuentes oficiales. Para ello, se realiza una revisión completa y clara de los principios básicos de la econometría espacial, una disciplina no bien conocida con la suficiente profundidad en España, con la excepción de algunos investigadores.

Esta obra constituye también una importante contribución al análisis económico en cuanto que acomete la estimación de un importante indicador de bienestar social, la renta familiar disponible, a un nivel espacial microterritorial municipal. Pero la trascendencia de este estudio se extiende, en general, a las ciencias sociales, dado que el proceso de predicción-extrapolación espacial, en cuanto obtención de datos espaciales no disponibles en una escala geográfica inferior, supera el ámbito del análisis económico y abarcaría también temas como la ciencia política, sociología, marketing u otros campos de aplicación, como la sanidad.

Por último, la importancia concreta para la Comunidad de Madrid se evidencia con la predicción-extrapolación de los datos de renta familiar disponible realizada en este estudio para los municipios de esta Comunidad. Se trata de un ejercicio no circunscrito únicamente a la obtención de unos resultados a partir de técnicas más o menos complejas,

sino también a la obtención de un mayor conocimiento de la realidad económica madrileña, evaluada con el máximo nivel de actualización posible (año 2000), desde el punto de vista del poder adquisitivo medio de los hogares en los 179 municipios de la región. La gran consistencia de la metodología empleada, su gran actualización y la posibilidad de realizar análisis comparativos de los resultados con otras Comunidades Autónomas confieren a este estudio una ventaja sobre otra información similar ya existente.

El contenido de esta investigación se divide en seis capítulos. Tras un capítulo I en el que se plantean los objetivos y estructura de la obra, en el capítulo II se presenta la econometría espacial como una disciplina necesaria para el correcto tratamiento de los datos geográficos en los modelos de regresión de corte transversal y datos de panel, diferenciada de la estadística espacial o geoestadística. Los capítulos III y IV constituyen el corazón de la obra, lo que se conoce con el nombre de análisis exploratorio de datos espaciales (AEDE) y análisis confirmatorio o modelización espacial. El objetivo del AEDE, en el marco de la modelización espacial, sería la detección de los efectos de dependencia y heterogeneidad espacial en las variables de un modelo, como punto de partida en la etapa de modelización. La especificación y el contraste de los efectos espaciales en un modelo de datos geográficos forman parte de lo que se conoce como análisis confirmatorio de datos.

Por último, el capítulo V desarrolla la actividad de predicción-extrapolación de datos espaciales como una última etapa en el desarrollo de la econometría espacial, que debe ser también iluminada por el esquema teórico del problema de la unidad espacial modificable y las inferencias ecológicas. Se propone un proceso de cinco etapas a seguir en todo ejercicio de predicción-extrapolación, que se ilustra con la estimación de la renta familiar disponible municipal de la Comunidad de Madrid, referida a 2000, de cuyos resultados se extraen conclusiones interesantes.

En el capítulo VI, conclusiones y reflexión final, se realiza un compendio ordenado de toda la tesis, por lo que resulta de mucha utilidad para el lector que desee conocer de forma rápida su contenido. Las referencias bibliográficas cierran esta obra que esperamos sea de utilidad para todos los investigadores sociales interesados en el análisis territorial, en cualquiera de sus formas, y que también contribuya a la difusión de las técnicas de econometría espacial en nuestro país.

Para concluir estas consideraciones, quisiera expresar mi agradecimiento más profundo al profesor Pedro Chasco Lafuente, mi padre y codirector de tesis, cuya contribución a esta obra es imposible de medir y al que debo el interés y la vocación por el análisis económico espacial. De igual forma, quiero agradecer el apoyo constante que he recibido de mi madre y hermana, así como de mis amigos más cercanos, que han tenido que soportar mis enormes ausencias y que con su apoyo han acompañado estos años de esfuerzo, no siempre fáciles. Finalmente, dedico este libro a mis sobrinos, Ángel y Nacho, para quienes la vida y el aprendizaje no han hecho más que empezar y a los que esta obra, quizá en el futuro, pueda resultarles de alguna utilidad.

CAPÍTULO I

Planteamiento y Objetivos

I.1. INTRODUCCIÓN Y OBJETIVOS DE LA INVESTIGACIÓN

El análisis del espacio y la localización ha constituido uno de los pilares de la investigación en las ciencias sociales durante las últimas décadas. A ello, han contribuido, sin duda, los profundos cambios geopolíticos recientemente acaecidos en nuestro mundo, entre los que cabe señalar la descomposición de la ex-Unión Soviética (con la creación de los nuevos estados de la República Rusa y la Europa del Este), el proceso de unificación de la Unión Europea, que afianza la realidad de una Europa de las Regiones y, en España, la consolidación del proceso de descentralización territorial, que concede una importancia creciente al ámbito infrarregional.

Por otro lado, la investigación de los mercados en su distribución territorial es una importante necesidad para las instituciones públicas y privadas, que necesitan abordar acciones como la planificación territorial, programación de la actividad comercial e industrial, urbanismo, medioambiente, etc. Por ese motivo, existe una demanda creciente tanto de métodos de análisis espacial como de información geográfica, estimulada por la existencia de un volumen, cada vez más amplio, de datos geocodificados, así como de una tecnología informática y científica que permita su correcto tratamiento.

En este cruce de caminos y posibilidades cabe circunscribir esta tesis cuyo **objetivo** fundamental consiste en proponer la **predicción de datos espaciales** como parte de la econometría espacial, en línea con la intuición de Anselin (2000B, 2001B), presentando una metodología fundamentada en instrumentos exploratorios y confirmatorios propios de la econometría espacial. La denominación “predicción-extrapolación espacial” pretende distinguir esta actividad del conocido método geoestadístico de “interpolación espacial”, para el que existe una amplia literatura (Cressie, 1993; Chica, 1994; Goovaerts, 1997). La extrapolación espacial, en concreto, podría definirse como el conjunto de métodos

inferenciales destinados a obtener datos microterritoriales a partir de información espacial agregada.

Efectivamente, el habitual proceso econométrico de especificación, estimación y contraste puede culminar, también en el ámbito espacial, en un ejercicio de predicción (como sucede con el análisis de series temporales) que, en el caso de la extrapolación, permite la obtención de datos espaciales no disponibles en una escala geográfica inferior. Además de la econometría espacial, la predicción-extrapolación debe asumir los principios del Problema de la Unidad Espacial Modificable (MAUP) y las Inferencias Ecológicas (IE), considerados desde hace años por otros investigadores sociales (geógrafos, sociólogos, políticos). Todo esto implica un esfuerzo añadido, realizado en este trabajo, de **conocimiento interdisciplinar** de la materia.

Por tanto, hablamos expresamente de predicción espacial, es decir, de una actividad en gran parte similar a su homóloga temporal, que ha sido definida por Anselin (2000B) como una “alquimia de la estadística”, lo que recuerda mucho a expresiones utilizadas por algunos especialistas de la predicción temporal, que se refieren a estos procedimientos con términos como “modernas artes adivinatorias de la estadística aplicadas al campo de la economía”, y a los que los utilizan, como “adivinos, brujos, magos y hechiceros de nuestros tiempos” (Pulido y López, 2000; cap. 1).

Evidentemente, en el campo espacial no puede hablarse de “futuro” pero sí tiene sentido plantearse “obtener datos donde no hay datos” (Anselin, 2000B), lo que asemeja mucho esta labor a la predicción temporal. Es decir, puede también afirmarse, como en Pulido y López (2000), que la predicción espacial debe recurrir a “potentes ordenadores y depuradas técnicas matemáticas”, porque sería absurdo desaprovechar “los avances espectaculares del conocimiento humano”. Pero sin olvidar que las verdades económico-sociales no son eternas ni directamente extrapolables de unos ámbitos territoriales a otros, por lo que debe siempre partirse del hecho de que “tenemos muchas posibilidades de equivocarnos”, tanto en la búsqueda de un porvenir incierto (ámbito temporal) como de un territorio desconocido (ámbito espacial).

Este margen de error puede corregirse tanto a través de técnicas adecuadas como de un conocimiento, lo más profundo posible, de la distribución del fenómeno en el territorio al que se va a extrapolar. Es decir, el objetivo planteado en esta tesis consiste no solo en propugnar la utilización de las técnicas de la estadística y la econometría espacial en la actividad de la predicción-extrapolación, sino también la necesidad de una “sensibilidad al dato” sólo posible a partir de información complementaria y experiencia sobre la realidad que se va a extrapolar. Efectivamente, los especialistas en predicción no deben ser observadores desapasionados, sino actores implicados en unos fenómenos que desean conocer con mayor profundidad, porque tan peligroso como ignorar la existencia de instrumentos adecuados para el tratamiento de datos geográficos, sería realizar aplicaciones sofisticadas de un fenómeno cuya distribución microterritorial se desconoce.

Un caso particular de predicción-extrapolación espacial sería el constituido por los llamados métodos indirectos de estimación de renta que, en su mayoría, ignoran problemas como el sesgo de la agregación o los efectos espaciales, con unas consencuencias negativas

sobre las predicciones no siempre bien valoradas. Aunque las extrapolaciones espaciales de renta son las más conocidas en el campo económico, debido al carácter de la renta familiar como indicador de bienestar de un territorio, el campo de aplicación de la predicción-extrapolación de datos espaciales es, en realidad, muy superior y abarcaría temas como la ciencia política (que muchas veces se plantea conocer el sentido del voto en un colectivo de personas o sección censal, extrapolando relaciones establecidas para el nivel de distrito electoral), la sociología (extrapolación de tasas de delincuencia o de ocupación femenina en un distrito o sección), el marketing (extrapolación de hábitos de consumo y ahorro de la población en un municipio o barrio) u otros campos de aplicación, como la sanidad, que en muchas ocasiones necesita estimar las tasas de concentración de determinadas sustancias contaminantes en el microterritorio para tomar medidas que eviten una epidemia.

Dado el papel central que ocupan las herramientas espaciales de la econometría en la predicción-extrapolación, en esta obra se ofrece una **exposición y sistematización** personales del conjunto de técnicas de análisis exploratorio y confirmatorio de la econometría espacial, a través de una **bibliografía** completa y muy actualizada, que ha sido realmente consultada y analizada en su práctica totalidad.

La eficacia e interés de la metodología propuesta puede ser evaluada a través de un ejercicio de predicción-extrapolación, conocido por los investigadores económicos, la estimación de la **renta familiar disponible municipal**, en el que se hacen confluír tanto la utilización de técnicas de econometría espacial (a través de un “software” adecuado), como un amplio conocimiento de la distribución de la renta familiar disponible en los municipios españoles.

Para ello, ha sido determinante la experiencia de la autora como Directora A. del **Área de Economía Espacial Microterritorial** del Instituto Lawrence R. Klein, de la Universidad Autónoma de Madrid, y como profesora del Departamento de Economía Aplicada de la citada Universidad. No en vano, tanto los objetivos planteados como todo el esfuerzo de investigación realizado en este trabajo, tienen como centro su aplicación directa en los Proyectos del Área de Economía Espacial Microterritorial, abriendo también nuevos cauces para la investigación, tanto en la citada Área como en el Instituto Klein, en general.

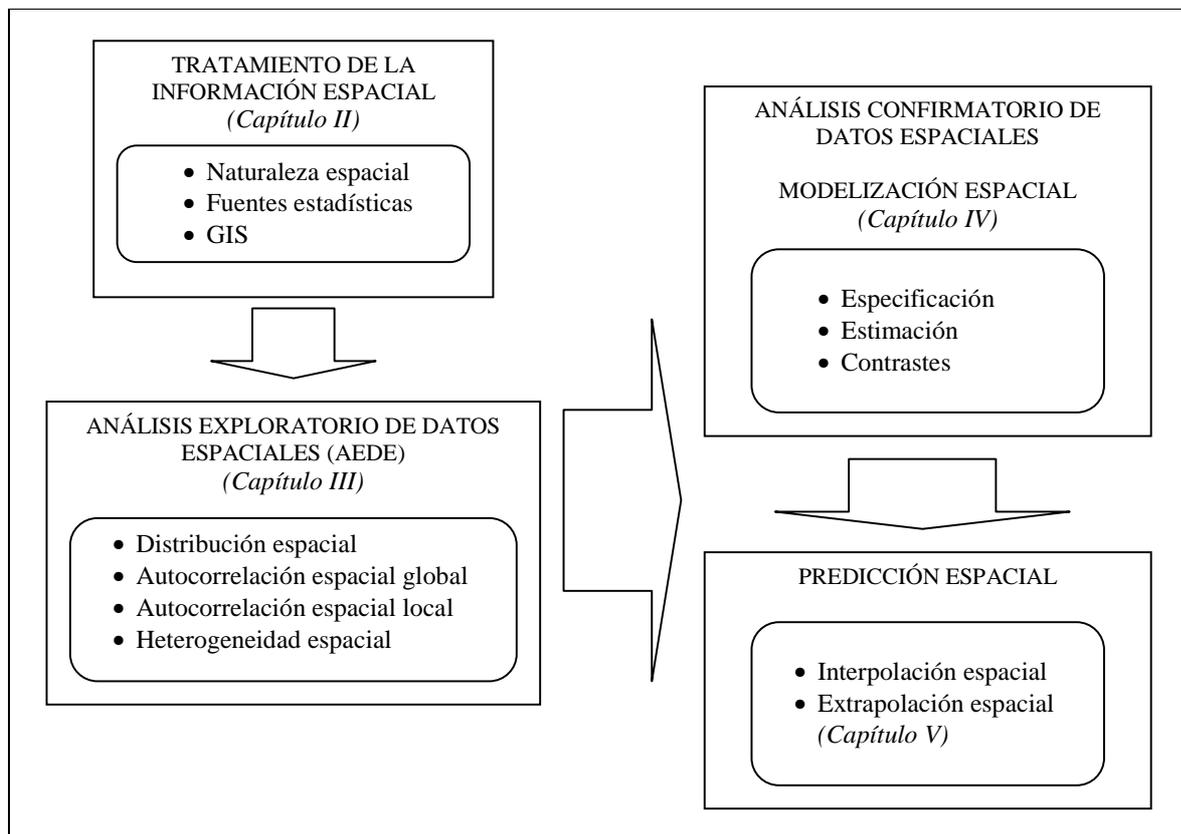
Este trabajo se enmarca en la labor investigadora del Instituto L. R. Klein, su vocación por la econometría (Pulido y Pérez, 2001), la predicción (Pulido y López, 2000), la microeconometría y modelos de datos de panel (Vicéns, 1995, 1996), los modelos de localización e interacción espacial (Chasco, 1988; Mahía, 1994; Chasco, 1997) y, en general, el análisis de la economía regional. En concreto, desde hace años, existe en el Instituto L. R. Klein un gran interés por la economía de la Comunidad de Madrid (López, 2002; Vicéns, 2001; Chasco, 2001A y B; Sur, 2000A, B y C; Pulido, 2000; etc.). Apoyados en esta experiencia, la aplicación propuesta en esta tesis aborda un tema relacionado con la **Comunidad de Madrid**, en concreto, la predicción-extrapolación de la renta familiar disponible de los municipios madrileños, como un ejercicio no circunscrito únicamente a la obtención de unos resultados a partir de técnicas más o menos complejas, sino con el deseo también de aportar un mayor conocimiento de la realidad económica de la Comunidad de Madrid, evaluada con el máximo nivel de actualización posible (año 2000), desde el punto de vista del poder adquisitivo medio de los hogares en los 179 municipios de la región.

La estimación de la renta familiar disponible en los municipios españoles es un proyecto de investigación en el que el Área de Economía Espacial Microterritorial lleva trabajando desde hace años, bajo la dirección del profesor Pedro Chasco, director de esta tesis doctoral. Este trabajo supone (así lo esperamos) la consolidación de esta metodología así como una propuesta para cualquier investigador interesado en este tema de la predicción-extrapolación de datos microterritoriales.

I.2. ESTRUCTURA DE LA INVESTIGACIÓN

El contenido de esta obra se divide en seis capítulos cuya secuencia y concatenación están expresadas en la Figura I.1.

Figura I.1. Etapas del análisis de econometría espacial



Fuente: Elaboración propia a partir de Anselin y Bao (1997).

Tras este capítulo de Planteamientos y objetivos, en el **capítulo II** se presenta la econometría espacial como una disciplina necesaria para el correcto tratamiento de los datos geográficos en los modelos de regresión de corte transversal y datos de panel, diferenciada de

la estadística espacial o geoestadística (apartado II.1.). La difusión de la econometría espacial entre los investigadores en el campo económico es aún escasa aunque los trabajos publicados en los últimos años son, en general, de gran interés y rigor científico. La necesidad de los métodos de econometría espacial surge como consencuencia de las peculiaridades propias de los datos espaciales, resumidas en los llamados efectos de dependencia y heterogeneidad. La creciente demanda de información microterritorial, en concreto, exige del investigador un dominio de las fuentes estadísticas existentes, no siempre bien conocidas. Estas bases de datos, que en el ámbito territorial pueden adquirir grandes dimensiones, demandan programas informáticos, como los sistemas de información geográfica (GIS), capaces de tratar y visualizar datos geocodificados (apartado II.2.).

Los **capítulos III y IV** constituyen el corazón de la econometría espacial, lo que se conoce con el nombre de análisis exploratorio de datos espaciales (AEDE) y análisis confirmatorio o modelización espacial. El AEDE (capítulo III) constituye la etapa previa, tanto al análisis confirmatorio, como a todo ejercicio de predicción, y podría definirse como el grupo de herramientas estadístico-gráficas que describen y visualizan las distribuciones espaciales, identificando localizaciones atípicas, descubriendo formas de asociación (autocorrelación espacial) que, a su vez, pueden ser de carácter global o local, y sugiriendo estructuras en el espacio geográfico (heterogeneidad espacial).

Cuando el AEDE es abordado desde los presupuestos conceptuales de la estadística espacial, los métodos utilizados suelen ser diferentes a los presentados en este capítulo, que siguen el esquema reticular (“lattice”) de la econometría espacial.

El objetivo del AEDE, en el marco de la modelización espacial, sería la detección de los efectos de dependencia y heterogeneidad espacial en las variables de un modelo, como punto de partida en la etapa de modelización. La especificación y el contraste de los efectos espaciales en un modelo de datos geográficos forman parte de lo que se conoce como análisis confirmatorio de datos (**capítulo IV**).

En este capítulo, se analiza la presencia de dichos efectos de dependencia y heterogeneidad espacial en el contexto de los modelos de regresión lineal, sus consecuencias sobre la especificación (apartado IV.1.), estimación y contraste, así como estrategias para la selección de modelos (apartado IV.2.).

El **capítulo V** desarrolla la actividad de predicción-extrapolación de datos espaciales como una última etapa en el desarrollo de la econometría espacial, que debe ser también iluminada por el esquema teórico del problema de la unidad espacial modificable y las inferencias ecológicas (apartado V.I). Los modelos de predicción-extrapolación propuestos en la literatura económica (apartado V.II.) se reducen prácticamente a los métodos indirectos de estimación de la renta en comarcas y municipios.

Por último, se propone un proceso de cinco etapas a seguir en todo ejercicio de predicción-extrapolación, que se ilustra con la estimación de la renta familiar disponible municipal de la Comunidad de Madrid, referida a 2000, de cuyos resultados se extraen conclusiones interesantes (apartado VI.3.).

En el capítulo de Conclusiones y reflexión final (**capítulo VI**) se realiza un compendio ordenado de toda la obra, por lo que resulta de mucha utilidad para el lector que desee conocer de forma rápida su contenido.

Las referencias bibliográficas cierran esta obra que esperamos sea de utilidad para todos los investigadores sociales interesados en el análisis territorial, en cualquiera de sus formas, y que también contribuya a la difusión de las técnicas de econometría espacial en nuestro país.

CAPÍTULO II

Econometría de datos espaciales

En el presente capítulo, se define la econometría espacial como una disciplina de la econometría general necesaria para el correcto tratamiento de los datos geográficos en modelos de regresión de corte transversal y datos de panel, así como en el ejercicio de la predicción espacial. El capítulo se divide, a su vez, en dos apartados. En el primero de ellos, tras una breve presentación del concepto de econometría espacial y de los efectos espaciales de dependencia y heterogeneidad (propios de los datos geográficamente distribuidos), se detallan los antecedentes históricos de la econometría espacial que tienen su origen, como campo específico, durante los años 70, en una Europa de las regiones, necesitada del análisis econométrico de datos de ámbito infranacional, aunque las principales aplicaciones en este campo tuvieron lugar en la última década del siglo XX. En el segundo apartado de este capítulo, se analiza la naturaleza de los datos geográficos, que exige, por un lado, la disponibilidad de programas informáticos capaces de tratarlos y visualizarlos, como los sistemas de información geográfica (GIS), y, por otro, de bases de datos socioeconómicos geocodificados.

El esquema del capítulo es el siguiente:

CAPÍTULO II. ECONOMETRÍA DE DATOS ESPACIALES

II.1. Econometría espacial

- II.1.1. Concepto
- II.1.2. Antecedentes
- II.1.3. Aplicaciones

II.2. Datos espaciales

- II.2.1. Naturaleza de los datos espaciales
- II.2.2. Sistemas de Información Geográfica (GIS)
- II.2.3. Fuentes estadísticas de información microterritorial
- II.2.4. Problemas en el tratamiento de datos espaciales

II.1. ECONOMETRÍA ESPACIAL

II.1.1. Concepto

La econometría espacial es una disciplina de la econometría general que incluye el conjunto de técnicas de especificación, estimación, contraste y predicción necesarias para el tratamiento de los datos espaciales. En palabras de Anselin (2001D), podría también definirse la econometría espacial como “la parte de la econometría que se ocupa del tratamiento de la interacción espacial (autocorrelación espacial) y la estructura espacial (heterogeneidad espacial) en los modelos de regresión de corte transversal y de datos de panel”. En este sentido, se trata de un concepto similar a la geoestadística o la estadística espacial, muy utilizada por físicos y geógrafos (Cressie, 1993), aunque la econometría espacial difiere de la estadística espacial, del mismo modo que la econometría es distinta de la estadística, en general.

Efectivamente, un ejercicio completo en econometría empírica comienza con la especificación de un número de relaciones económicas (dadas en forma cuantitativa), así como de las variables que intervendrán en el modelo y las formas funcionales que relacionarán a las mismas. Estas relaciones son expresiones matemáticas procedentes de modelos teóricos (por ejemplo, la teoría económica) y serán las hipótesis a contrastar. A continuación, esta especificación debe ser identificable, es decir, los parámetros del modelo han de ser susceptibles de estimación a partir de un número suficiente de observaciones. Y, por último, este modelo podría ser utilizado para la obtención de datos espaciales, normalmente de ámbito microterritorial, no existentes, que es lo que se conoce como predicción espacial (Anselin, 2000B). Especificación, estimación, contraste y predicción son los cuatro pasos propios de la labor de la economía cuantitativa, también en el ámbito espacial. Este trabajo encuentra, de hecho y como es bien sabido, muchas dificultades, pero la consideración explícita de la dimensión espacial añade nuevas complicaciones al ejercicio econométrico, aunque también aporta una mayor generalización a su estudio.

Dado que la econometría espacial incluye las técnicas econométricas aplicadas a datos y modelos de naturaleza espacio-geográfica, la cuestión que se plantearía es si este tipo de datos contiene tantas peculiaridades como para demandar unas técnicas de investigación específicas. En este sentido, habría que decir, en primer lugar, que la principal característica de estos datos es su naturaleza **georreferenciada**, es decir, que su posición relativa o absoluta sobre el espacio contiene información valiosa para interpretar las relaciones consideradas. Otra característica básica del espacio geográfico es la **multidireccionalidad** de las relaciones que sobre él se establecen y la **multidimensionalidad**: efectivamente, en un área geográfica no es posible distinguir entre pasado, presente, futuro, sino que todo es presente, todo es pasado y todo es futuro.

Estas peculiaridades de los datos geográficos dan lugar a los llamados “efectos espaciales” de dependencia o autocorrelación espacial y heterogeneidad espacial, que deberían ser considerados explícita y adecuadamente en todo ejercicio de modelización y predicción de datos geográficos. Sin embargo, aún hoy, la mayoría de estas investigaciones suelen ser llevadas a cabo mediante técnicas econométricas

tradicionales e, incluso, muchos análisis estadísticos aplicados de estos modelos en cuestión, tienden a ignorar estos efectos espaciales, de los que se hace un resumen a continuación.

1.- Dependencia o autocorrelación espacial

Se trata del efecto espacial mejor conocido, gracias a los trabajos pioneros de Cliff y Ord (1973, 1981), y consiste en la falta de independencia que se produce a menudo entre las observaciones de datos de corte transversal, en línea con los contenidos formulados por Tobler (1979) en su “primera ley de la geografía”, según la cual “todo tiene que ver con todo, pero las cosas cercanas están más relacionadas entre sí que las cosas lejanas”. La dependencia espacial podría ser definida como la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar (Moreno y Vayá, 2000).

Es decir, una variable se encontrará espacialmente autocorrelacionada cuando los valores observados en un lugar determinado dependen, no sólo de ciertos factores externos (otras variables), sino de los valores observados en regiones vecinas. Por ejemplo, Cox (1969) demostró que existía una continuidad geográfica en las intenciones de voto de los norteamericanos, es decir, que si un estado votaba mayoritariamente al partido demócrata, se producía un aumento en la probabilidad de que los estados vecinos también votaran a los demócratas. Otro ejemplo, propuesto por Cliff y Ord (1981) presenta el análisis espacial como forma de conocimiento de las causas y formas de propagación de epidemias y enfermedades.

En este sentido, la dependencia espacial está determinada por una noción de espacio relativo o de localización relativa, que realza el efecto de la **distancia**. Cuando la noción de espacio va más allá del sentido euclídeo estricto, para incluir el “espacio general” de Isard (1969), que tiene en cuenta el espacio político, la distancia interpersonal, las redes sociales, etc., parece claro que la dependencia espacial es un fenómeno con un amplio campo de aplicación en las ciencias sociales.

A primera vista, la dependencia espacial puede parecer similar a la más conocida dependencia presente en los tests econométricos de correlación de series, en los modelos de distribución de retardos y en otros análisis de series temporales. Sin embargo, esta similitud sólo es real en parte, debido a la naturaleza multidireccional de la dependencia en el espacio que, frente a la clara situación unidireccional del tiempo, hace necesario el uso de una estructura metodológica diferente, como se verá más adelante (ver Apartado III.2).

2.- La heterogeneidad espacial.

Este segundo efecto espacial está relacionado con la diferenciación espacial o regional y viene definido por la ausencia de estabilidad en el espacio del comportamiento humano o de otras relaciones en estudio. Esto implicará que, en los modelos espaciales, las formas funcionales y los parámetros variarán con la localización geográfica no siendo homogéneos para toda la matriz de datos. Esto

es lo que ocurre, por ejemplo, en los modelos econométricos estimados con datos de corte transversal procedentes de unidades espaciales no similares, como es el caso de regiones ricas del norte y regiones pobres del sur.

A diferencia de lo que sucede con la dependencia espacial, el problema causado por la heterogeneidad espacial podría en gran parte ser resuelto mediante procedimientos de la econometría estándar (como el análisis cluster). Sin embargo, en algunos casos, la compleja interacción resultante de la estructura y los flujos espaciales pueden generar dependencia espacial combinada con heterogeneidad espacial, haciéndose altamente complicado distinguir entre ambos efectos.

Aunque dependencia y heterogeneidad son características propias tanto de los datos georreferenciados como de las series temporales, no pueden ser abordadas con los mismos instrumentos estadístico-econométricos. La diferencia existente entre ambos tipos de datos, temporales y espaciales, se encuentra en el tipo de instrumentos necesarios para investigar relaciones que tienen lugar sobre el espacio, dado que los desarrollados en un ámbito de series temporales presentan problemas de adecuación, dado que la dimensión espacio no es homologable a la dimensión tiempo.

II.1.2. Antecedentes

El conocimiento de los problemas causados por la dependencia y heterogeneidad espacial y sus efectos en la validez de los métodos estadísticos tradicionales podría remontarse a Student (1914), aunque los comienzos en el desarrollo de un campo separado de la estadística espacial se atribuyen a la obtención de los primeros índices formales para detectar la presencia de autocorrelación espacial en los años 50, con los trabajos de **Moran (1948)** y **Geary (1954)**, así como la obra de Isard (1956) sobre localización y economía espacial.

Tras esta fase inicial de reconocimiento del problema, es en la década de los setenta cuando **Paelinck y Klaasen (1979)** bautizaron estas técnicas con el nombre explícito de “econometría espacial”. Además, los trabajos de los geógrafos **Cliff y Ord (1972, 1973)**, junto con otras contribuciones publicadas en diferentes monográficos y revistas sobre el tema, sirvieron para poner de manifiesto la relevancia de un campo que no se restringe sólo a la ciencia regional, sino que se extiende también a la geografía, estadística espacial, biología, psicología, ciencia política y otras ciencias sociales.

Efectivamente, en 1979, Jean Paelinck y Luc Klaasen publican un texto donde recopilan algunos de sus últimos resultados y lo titulan “Spatial Econometrics”. El contenido de este texto es denso y su lectura nada fácil. Su importancia estriba en que ambos autores saben establecer el nexo de unión entre teoría y evidencia empírica en el análisis regional. Este reto fue asumido entonces por una parte minoritaria de la comunidad científica, que se esforzó en dar contenido a este término, con resultados quizá parcos e insuficientes.

El gran desarrollo de la econometría espacial se produce en los años ochenta y noventa, que se inicia con los trabajos de Cliff y Ord (1981). Los textos posteriores de **Anselin (1980, 1988A)** y **Arbia (1989)** serán capitales en la fundamentación de una

metodología de análisis econométrico de datos espaciales. De una manera especial, durante los años noventa, el libro “Spatial Econometrics: Methods and Models” de Luc Anselin (1988A) fue considerado como el manual por excelencia de esta disciplina, en una década en la que fueron apareciendo en las revistas de economía regional contribuciones concretas en el campo de la econometría espacial, así como varias colecciones de artículos, como en **Anselin y Florax (1995, 2000)** y **Ficher y Getis (1997)**. De esta forma, la importancia y relevancia de los métodos que analizan los efectos espaciales en los modelos econométricos se ha ido incrementando de forma notable.

Aún así, es inevitable recordar que en 1970, nueve años antes de la aparición de la obra “Spatial Econometrics” de Paelinck y Klaasen, Box y Jenkins habían publicado “Time Series Analysis. Forecasting and Control”, provocando un desarrollo vertiginoso del análisis de series temporales. Mientras que, en el otro extremo, el amplio espectro que cubre la econometría espacial ha sido ampliamente ignorado en la literatura econométrica tradicional hasta nuestros días, a excepción de Kmenta (1971), Johnston (1984) y Novales (1993), que se limitan a un breve apunte. En general, como ya se ha dicho, los efectos espaciales no son tenidos en cuenta incluso en los métodos recientemente desarrollados para el análisis de datos de panel, donde el punto de mira se pone más en la dimensión temporal, mientras que la potencial dependencia a través de los datos de corte transversal no es tenida en cuenta. Se trata, sin duda, de un desequilibrio sorprendente si se tiene en cuenta que los agentes objeto del estudio económico están dispersos sobre el espacio (económico, geográfico, social) y se esfuerzan por superar sus distancias.

Dada la utilidad innegable del análisis econométrico espacial, cabría preguntarse por las razones que han condenado al estudio econométrico espacial a un cierto olvido por parte de algunos investigadores. Por un lado, es bien sabido que la importancia y urgencia de la actividad de la predicción, propiamente dicha, en un campo incierto como es la economía, ha sido una de las causas que ha motivado el gran avance experimentado por el análisis econométrico de series temporales. Esto se une a que la escasez de información estadística microterritorial y la ausencia de un “software” útil y de bajo costo capaz de tratarla adecuadamente, han relegado a un segundo plano, hasta hace pocos años, el análisis de datos geográficos en el proceso de toma de decisiones.

Actualmente, la disponibilidad de nuevas fuentes estadísticas de información microterritorial (censos, directorios, encuestas, etc.), así como de mejores herramientas informáticas de tipo estadístico-econométrico para su correcto tratamiento y visualización, han dado lugar a una demanda creciente de este tipo de análisis espacial, tanto desde las instituciones privadas como públicas. Según Anselin y Florax (1995), podrían apuntarse tres grandes razones que justificarían este creciente interés por las técnicas econométricas espaciales:

- a) La importancia que el papel del espacio y de la interacción espacial ha cobrado últimamente en las teorías de las ciencias sociales y en el marketing. En concreto, la teoría económica clásica muestra gran interés en lo que se ha dado en llamar la “nueva geografía económica”, que encuentra su máximo exponente en Krugman (1991, 1993), entre otros. Por otro lado, hay autores que hablan ya abiertamente del “geomarketing” como una disciplina propia de la investigación de mercados. Además, esta atención renovada por el tema

espacial no se limita a la economía, sino que también tiene sus exponentes en otras ciencias sociales y en el estudio de las relaciones internacionales (por ejemplo, en la nueva geopolítica de Ward, 1992).

- b) La creciente disponibilidad de grandes bancos de datos socioeconómicos con observaciones georreferenciadas. Tanto en los Estados Unidos como en la Unión Europea, los institutos de estadística oficiales ofrecen grandes bases de datos, con información de ámbito regional y local, en formatos accesibles y de bajo costo. Por ejemplo, en los EEUU existe la base REIS (“Regional Economic Information System”), que incluye una gran cantidad de datos de ámbito provincial, publicada en formato CD-ROM por el Ministerio de Comercio; de forma similar, el banco de datos REGIO, publicado por Eurostat, contiene datos regionales de los países integrantes en la UE. Este análisis sin precedentes de grandes cantidades de datos de corte transversal o espacio-temporal hace aflorar, más con carácter de regla que de excepción, problemas de autocorrelación espacial y/o temporal, que deben ser abordados con precisión.
- c) El desarrollo de una tecnología eficiente y de bajo coste capaz de manejar observaciones georreferenciadas: sistemas de información geográfica (GIS, en la terminología anglosajona) y, en general, software para el análisis de datos espaciales. El uso de un GIS hace posible el almacenamiento eficaz y una rápida recuperación y visualización interactiva de las formas correspondientes a conjuntos de datos geográficos. Esto implica la posibilidad de aplicación de los últimos avances en las técnicas analíticas. Con relación a este tema, cabe destacar el programa “SpaceStat” que Anselin (1992D, 1995A), ha puesto a disposición de la comunidad investigadora, en el que implementa bajo el entorno Gauss los diferentes contrastes de detección de autocorrelación espacial en los modelos de regresión así como los métodos de estimación que incorporan dicha problemática. Recientemente además, es posible la integración entre programas estadísticos y GIS como SpaceStat y ArcView (Anselin, 1998, 1999B), haciendo posible, entre otras cosas, la visualización geográfica (mapas) de los resultados obtenidos en el análisis econométrico espacial (ver Apartado III.1).

Pese a todo, actualmente es todavía escasa la difusión de la econometría espacial en el terreno aplicado. Así, la mayor parte de la literatura publicada en los últimos años sobre estos temas ha tenido poco eco entre la mayoría de los analistas de economía empírica.

II.1.3. Aplicaciones

La mayor parte de los análisis que, de hecho, tienen en cuenta los efectos espaciales raramente superan los métodos de regresión lineal propuestos por Cliff y Ord (1981), es decir, no suelen tener suficientemente en cuenta la literatura econométrica espacial formulada en los últimos diez o quince años. Estos trabajos son todavía poco conocidos por los analistas de la economía aplicada y podrían agruparse, por su contenido, en los siguientes grandes temas o cuestiones: análisis de los efectos espaciales (dependencia espacial y heterogeneidad espacial) y su tratamiento en

modelos de regresión lineal, modelos de variables dependientes limitadas y modelos espacio-temporales (algunos de estos estudios se exponen en la Tabla II.1.1).

1. Efectos espaciales

Los efectos espaciales de dependencia y heterogeneidad espacial implican, de un lado, la falta de independencia y, por tanto, la existencia de una estructura entre datos espaciales determinada por la localización absoluta o relativa de dichos datos y, de otro, la existencia de inestabilidad estructural que determina diferentes localizaciones o sub-regiones en los datos. En los últimos años, los estudios se han centrado en una mejor interpretación de las matrices de pesos espaciales o del parámetro autorregresivo (Smirnov y Anselin, 2001; Kelejian y Robinson, 1995; Florax y Rey, 1995), en la formulación de nuevos contrastes de dependencia y heterogeneidad espacial (Sohn, 2002; Baltagi y Li, 2001; Graaff *et al.*, 2001; Rey, 2001B, Mur, 2000; Anselin *et al.*, 1996) o en las consecuencias que la ignorancia de dichos efectos producen sobre diversos aspectos de modelización, como sería el caso de las inferencias ecológicas (Anselin, 2000B; Anselin y Cho, 2002A y B; Cho, 2001).

Además, algunos autores han avanzado en el análisis exploratorio de datos espaciales (AEDE) integrado en el entorno GIS (Anselin, 1998, Anselin y Bao, 1997; Wise *et al.*, 1997), aplicado al análisis regional clásico (Chasco y Vicéns, 2000; Vayá, 1998) o en problemas aún pendientes de resolver (Anselin, 2001A).

2. Efectos y datos espaciales en modelos de regresión lineal

Cuando se trabaja con variables espacialmente distribuidas es necesario introducir y tratar convenientemente los efectos espaciales de dependencia y heterogeneidad espacial para una correcta aplicación de los modelos de regresión lineal. Por ejemplo, suele ser necesaria la especificación de los efectos de vecindad o “spillover” en los modelos de convergencia regional (López-Bazo *et al.*, 2001; Rey, 2001A; Toral, 2001A; Baumont *et al.*, 2000; Magalhães *et al.*, 2000; Rey y Montouri, 1999; Vayá, 1996), crecimiento económico (Thomas, 2001; Pons y Viladecans, 1999, Goicolea *et al.*, 1998), productividad (Moreno *et al.* 2001; Moreno y Artís, 1996;), actividad I+D (Acs *et al.*, 2001; Fischer y Varga, 2001; Anselin *et al.*, 2000), demanda de transportes (Haynes *et al.*, 2000), modelos hedónicos del precio de la vivienda (Sandberg y Johansson, 2001; Van der Kruk, 2001) y otros fenómenos sociales, como los actos delictivos (Baller *et al.*, 2001; Ceccato *et al.*, 2002).

Asimismo, la dependencia espacial existente en las variables explicativas omitidas en un modelo tendrá su reflejo en la perturbación aleatoria, como se expone en el modelo de flujos de viaje propuesto por Bolduc, Laferrière y Santarossa (1995), en Kelejian y Robinson (1999) o en los modelos espaciales de componentes del error (Anselin y Moreno, 2001; Tsutsumi *et al.*, 2000).

Los estudios realizados sobre este tema han sido numerosísimos y tratan gran cantidad de aspectos, como la veracidad de los diferentes contrastes de dependencia espacial en el trabajo con muestras pequeñas (Anselin y Florax, 1995), eliminación del problema de dependencia espacial en las variables mediante métodos de filtrado (Alañón, 2001; Stough *et al.*, 1998; Getis, 1995, 2002), cointegración espacial de series (Fingleton, 1999; Lauridsen, 1999), aplicación de nuevos contrastes de autocorrelación espacial (Anselin, 2001C; Baltagi y Li, 2001), introducción de dinamicidad espacial en los modelos (Lauridsen, 2000), estimación de modelos espaciales de ecuaciones simultáneas (Herny *et al.*, 2001; Rey y Boarnet, 1998).

Asimismo, algunos autores han determinado las consecuencias sobre los modelos de regresión lineal de algunos problemas que plantean los datos espaciales (Haining, 1995), la no estacionariedad o la inestabilidad espacial de los parámetros (Caseti y Poon, 1995, Fotheringham *et al.*, 1997) o la heterogeneidad espacial en los coeficientes de autocorrelación espacial (Kristensen, 1998).

3. Efectos espaciales en modelos de variables dependientes limitadas

A diferencia del caso del análisis de regresión lineal con efectos espaciales, para el que existe ya gran cantidad de estudios, la literatura acerca de los modelos logit, probit y tobit es aún muy reciente y limitada en su producción. Algunos estudios se dirigen a la incorporación de los estadísticos de autocorrelación espacial a la estimación de este tipo de modelos.

Por ejemplo, algunos autores han introducido los efectos espaciales en modelos logit (como Dubin, 1995, para un modelo de difusión de una innovación tecnológica entre empresas de un entorno), probit (McMillen, 1995, demuestra los problemas que el efecto de heterogeneidad espacial produce en estos modelos) y tobit (Snell, 1999, o Kelejian y Prucha, 2001, que han estimado la distribución del estadístico I de Moran en estos modelos).

4. Efectos espaciales en modelos espacio-temporales

El tratamiento de los efectos espaciales en modelos que combinan datos de corte-transversal y series temporales puede abordarse dentro del esquema estadístico clásico (Buettner y Zew, 1999; Henry *et al.*, 1999), así como desde una perspectiva bayesiana (LeSage, 1995; LeSage y Krivelyova, 1999; Pace y LeSage, 2002), de grandes potencialidades en el contexto de la predicción espacio-temporal.

En paralelo con el desarrollo de una metodología adecuada para el tratamiento de modelos espaciales, se ha producido en los últimos años una demanda creciente de información estadística territorial, así como de técnicas capaces de manejar datos geográficos, a veces de gran volumen. La importancia de los sistemas de información

geográfica (GIS) y la creciente producción de bases de datos de ámbito regional y microterritorial será objeto de tratamiento, entre otras cuestiones, en el apartado siguiente.

Tabla II.1.1. Algunos trabajos de econometría espacial publicados en revistas y libros de perfil económico en los años 2000 a 2002

Autor	Año	Publicación
<i>Anselin, L. y Cho</i>	2002	Polical Analysis
<i>Anselin, L. y Cho</i>	2002	Political Analysis
<i>Ceccato, V., R. Haining y P. Signoretta</i>	2002	Annals of the American Association of Geographers
<i>Getis, A.</i>	2002	Geographical Analysis
<i>Pace, R.K. y J. LeSage</i>	2002	Geographical Analysis
<i>Acs, Z., L. Anselin y A. Varga</i>	2001	Research Policy
<i>Anselin, L.</i>	2001	American Journal of Agricultural Economics
<i>Anselin, L.</i>	2001	Journal of Statistical Planning and Inference
<i>Anselin, L.</i>	2001	Capítulo de libro. Ed. Oxford: Basil Blackwell
<i>Baller, R., L. Anselin, S. Messner y D. Hawkins</i>	2001	Criminology.
<i>Baltagi, B. y D. Li</i>	2001	International Regional Science Review
<i>Cho, W.</i>	2000	Electoral Estudios
<i>Graaff, T., R. Florax, P. Nijkamp y A. Reggiani</i>	2001	Journal of Regional Science
<i>Herny, M, B. Schimitt y V. Piguet</i>	2001	International Regional Science Review
<i>Kelejian, H. e I. R..Prucha</i>	2001	Journal of Econometrics
<i>López-Bazo, E., E. Vayà y R. Moreno.</i>	2001	Capítulo de libro. Ed. Mundi-Prensa
<i>Moreno, R., E. López-Bazo, E. Vayá y M. Artís</i>	2001	Capítulo de libro Ed. Springer-Verlag
<i>Ord, K. y A. Getis</i>	2001	Journal of Regional Science
<i>Rey, S.J.</i>	2001	Geographical Analysis
<i>Rey, S.J. y L. Anselin.</i>	2001	International Regional Science Review
<i>Smirnov, O. y L. Anselin</i>	2001	Computational Statistics and Data Analysis

Autor	Año	Publicación
<i>Vayá, E., R. Moreno, E. López-Bazo y J. Suriñach</i>	2001	Capítulo de Libro (Ed. Springer-Verlag)
<i>Wheeler, C.</i>	2001	Journal of Regional Science
<i>Anselin, L.</i>	2000	Journal of Geographical Systems
<i>Anselin, L.</i>	2000	Annals of the Association of the American Geographers
<i>Anselin, L., A. Varga y Z. Acs</i>	2000	Papers in Regional Science
<i>Bao, S., L. Anselin, D. Martin y D. Stralberg</i>	2000	Journal of Geographical Systems
<i>Florax, R., H. Folmer y S. Rey</i>	2000	Capítulo de libro. Ed. Springer-Verlag
<i>Haining, R., S. Wise y P. Signoretta</i>	2000	Journal of Geographical Systems
<i>Moreno, R. y E. Vayá</i>	2000	Edicions Universitat de Barcelona, col·lecció UB 44, manuals.
<i>Unwin, A.</i>	2000	Computational Statistics & Data Analysis
<i>Walker, R., E. Moran y L. Anselin.</i>	2000	World Development

Fuente: Elaboración propia.

II.2. DATOS ESPACIALES

II.2.1. Naturaleza de los datos espaciales

El espacio geográfico es una realidad de naturaleza compleja y distribución continua imposible de manejar sin realizar un esfuerzo de abstracción encaminado a reducir o dividir esta continuidad en entidades numéricas discretas, observables y susceptibles de medición matemática. En esta línea, podría definirse el dato espacial como la observación de una variable asociada a una localización del espacio geográfico. Esta localización puede ser expresada mediante tres tipos de objetos espaciales: puntos, líneas y polígonos (Anselin, 2001B, MapInfo, 1995).

- Los **puntos** se encuentran determinados por las coordenadas terrestres de latitud y longitud, pudiendo corresponderse con individuos, empresas, ciudades, delitos cometidos o accidentes acontecidos.
- Las **líneas** son objetos abiertos que cubren una distancia dada y comunican varios puntos o nodos (dada la naturaleza esférica de la Tierra, las líneas son en realidad “arcos”). Éste sería el caso de las líneas de transmisión telefónica, infraestructuras viarias, calles de una ciudad, etc.
- Por último, los **polígonos** son figuras planas conectadas por distintas líneas u objetos cerrados que cubren un área determinada, como es el caso de países, provincias, secciones censales o áreas comerciales.

Figura II.2.1. Ejemplo de objetos espaciales en el municipio de Madrid



Fuente: Elaboración propia

En la Figura II.2.1, se han representado a modo de ejemplo algunos objetos espaciales en el municipio de Madrid: los polígonos son barrios de la ciudad (Goya, Lista, Guindalera, etc.), los puntos son las entidades bancarias presentes en los mismos y la línea que atraviesa de norte a sur la Figura representa la vía de circunvalación M-30. Para cada tipo de objeto es posible medir variables como población residente en los barrios, depósitos bancarios o flujo de tráfico rodado en la vía de circunvalación.

Como ya se ha indicado, los datos espaciales se caracterizan por su naturaleza georreferenciada, multidireccional y multidimensional.

- La **georreferenciación** pone de manifiesto que la posición relativa o absoluta de cualquier elemento sobre el espacio contiene una información valiosa, pues la localización (el lugar donde acontecen los fenómenos) debe considerarse explícitamente en cualquier análisis de datos espaciales.
- La **multidireccionalidad** aflora en situaciones de dependencia o interacción espacial, así como en otros fenómenos como externalidades, efecto contagio, desbordamiento (“spillovers”) o réplica (“copy-cattig”). La primera ley de la geografía de Tobler (1979) indica que “todo tiene que ver con todo, pero las cosas cercanas están más relacionadas entre sí que las cosas lejanas”.
- La **multidimensionalidad** de las áreas geográficas significa que, en ellas, no es posible distinguir entre pasado, presente, futuro, sino que todo es presente, todo es pasado y todo es futuro.

En la geografía, el proceso de “discretización” o partición del espacio es más sencillo cuando se trata de objetos con localización o fronteras claramente identificadas (edificios, carreteras o provincias), a diferencia de otros fenómenos que varían de forma continua, como la elevación del terreno, la temperatura del suelo o la densidad de la vegetación. En estos casos, la continuidad del fenómeno puede “aproximarse” mediante una muestra de puntos situados en lugares representativos de la superficie que se desea registrar, de forma que sea posible obtener los valores situados entre dos puntos a partir de alguna función de alisado de los mismos. Existe además otro modo de expresar algunas realidades continuas sobre el espacio geográfico que consiste en determinar un grupo de polígonos contiguos con características homogéneas en su interior, pero diferentes entre sí. Los sistemas de información geográfica (GIS), como se verá a continuación, utilizan dos estructuras de datos, vectoriales y “raster”, para representar esta compleja realidad espacial.

II.2.2. Sistemas de Información Geográfica (GIS)

Los sistemas de información geográfica (GIS o “Geographic Information System”) hacen posible la visualización, exploración, almacenamiento eficaz, recuperación rápida y visualización interactiva de las formas correspondientes a conjuntos de datos geográficos. Los nuevos productos GIS no requieren de una gran especialización informática, pues son de fácil manejo para cualquier usuario que desee trabajar con datos espaciales. A partir de ficheros estándar grabados, por ejemplo, en programas de texto (.txt, .prn), bases de datos (.dbf, .mdb) u hoja de cálculo (.xls, .wks), es posible introducir en el GIS la información geográfica en forma de tablas cuyos datos podrán ser visualizados, manejados o tratados por un sistema de “query” (mediante lenguaje “SQL”), resumidos y organizados geográficamente. Así, con un fácil aprendizaje, podrán utilizarse los datos geográficos de un modo completamente nuevo, siendo posible el descubrimiento de estructuras o relaciones hasta ahora ocultas, que ofrecen nuevas perspectivas y conclusiones en cualquier investigación o negocio.

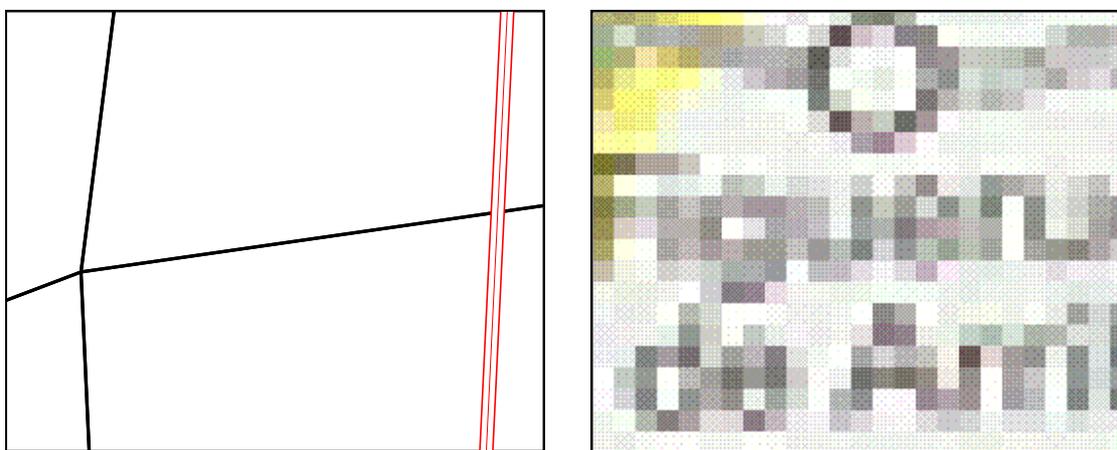
Los avances recientes en “hardware” y “software” informático han modificado la naturaleza del análisis espacial llevado a cabo por los GIS. Nuevos conceptos como almacenamiento de datos y minería de datos espaciales superan ya con creces la revolución cuantitativa que sufrió la geografía en los años 60, cuando se establecieron la mayoría de los fundamentos del actual análisis espacial.

Las ubicaciones geográficas de los datos suelen ser almacenadas en dos tipos de estructuras, vectorial o de datos “raster”, mientras que los datos propiamente dichos se guardan en un conjunto de tablas relacionadas geográficamente con las formas o figuras a que corresponde, lo que suele ser conocido como estructura de datos georrelacionados.

- Los **datos vectoriales** registran la información a modo de coordenadas geográficas (x,y) en un sistema plano o rectangular de coordenadas, de forma que las formas o figuras de puntos se registran como localizaciones simples del tipo (x,y), mientras que las formas lineales, incluyendo los contornos de los polígonos, se recogen como series ordenadas de coordenadas (x,y). Este tipo de datos vectoriales son muy apropiados para registrar la ubicación de formas geográficas discretas, con localizaciones precisas, como calles, ríos, límites fronterizos de parcelas, postes telefónicos, etc. Los datos vectoriales son altamente dependientes de la estructura de coordenadas (x,y), por lo que la representación gráfica de los vectores tiene forma de segmentos de líneas rectas, que pueden ser visualizados realizando una aproximación (“zoom”) sobre este tipo de formas, como puede apreciarse en la Figura II.2.2 (imagen izquierda), con los límites de dos barrios de Madrid, en negro, y la representación de la M-30, en rojo.
- Los **datos “raster”**, por su parte, registran la información espacial en una cuadrícula regular o matriz organizada en filas y columnas, de forma que cada celda contiene un número que representa una forma geográfica determinada, como un tipo de suelo, la elevación, el uso de la tierra, una pendiente, etc. Este

tipo de datos “raster” suele ser utilizado para almacenar información sobre formas geográficas que varían de forma continua sobre una superficie, como la elevación del terreno, reflectancia, profundidades marinas, etc. Los datos procedentes de imágenes son también un modo de datos “raster” en los que cada celda o “pixel” almacena un valor registrado como un dispositivo óptico o electrónico. Los datos “raster” son muy dependientes de la resolución del entramado regular en el que se graban, siendo fijo el tamaño de las celdas del entramado, de modo que si se realiza una aproximación sobre estos datos, podrá observarse la forma de las celdas, tal como se aprecia en la imagen de la derecha de la Figura II.2.2, correspondiente a la imagen de un mapa de carreteras Michelin, incorporada a un GIS.

Figura II.2.2. Ejemplo de datos vectoriales (izquierda) y raster (derecha)



Fuente: Elaboración propia.

La utilización de un GIS, junto con el análisis de datos y modelos espaciales, es cada vez más común en los campos de la economía aplicada, política económica, economía urbana y medioambiental, y economía del desarrollo. En los últimos años, se ha prestado especial atención a la integración del análisis espacial en el entorno GIS, tanto desde el campo académico como desde el comercial, planteándose diversos enfoques sobre el contenido concreto que debería tener el paquete de herramientas de análisis espacial de un sistema de información geográfica (ver, entre otros, Haining *et al.*, 2000; Anselin, 1998; Fischer y Getis, 1997), tal como se expondrá en el Apartado III.1.

La aplicación de las herramientas de análisis espacial y el GIS hace posible la obtención y tratamiento estadístico-econométrico de una gran variedad de datos geográficos de diferentes escalas o ámbitos territoriales. Efectivamente, como se verá a continuación, en los últimos años se está produciendo una auténtica explosión en la disponibilidad de bases de datos socioeconómicas geocodificadas, es decir, con información acerca de la localización geográfica de las observaciones, lo que impulsa el desarrollo del análisis

espacial en el conjunto de las ciencias sociales y del conjunto de la denominada Ciencia GI o Ciencia de la Información Geográfica (Anselin, 1999A).

II.2.3. Fuentes estadísticas de información microterritorial

La creciente disponibilidad de fuentes estadísticas de información geográfica (censos, directorios, encuestas, etc.), así como de mejores herramientas informáticas de tipo estadístico-econométrico para su correcto tratamiento y visualización, han dado lugar a una demanda creciente de este tipo de análisis espacial, sobre todo en el campo microterritorial. Hasta donde las leyes de la privacidad lo permiten, va siendo posible la distribución de datos de una amplia gama de variables socioeconómicas, desde la actividad económica o el mercado de trabajo a la salud pública, la educación o la religión, distribuidos en formatos accesibles al análisis geográfico. Todo esto está motivando una actividad cada vez mayor, en el sector privado, en forma de venta de datos, más o menos públicos, o indicadores propios, especialmente depurados y geocodificados.

En concreto, tanto en los Estados Unidos como en la Unión Europea, los institutos de estadística oficiales ofrecen grandes bases de datos, con información de ámbito regional y local, en formatos accesibles y de bajo costo. Por ejemplo, en los EEUU, la base REIS (“Regional Economic Information System”), incluye una gran cantidad de datos de ámbito provincial, publicada en formato CD-ROM por el Ministerio de Comercio; además, existe en este país una extensa infraestructura de redes viarias espacialmente georreferenciadas, los ficheros “Tiger” de la Oficina del Censo de los EEUU, y mapas digitales de una extensa gama de unidades administrativas. De forma similar, el banco de datos REGIO, publicado por Eurostat, contiene datos regionales de los países integrantes en la UE.

En España, el Instituto Nacional de Estadística dispone de bases de datos regionales fundamentalmente, aunque está realizando un esfuerzo importante por proporcionar también variables de ámbito provincial. Sin embargo, exceptuando la información procedente de los censos de población y padrones, existe aún un importante déficit de información oficial de ámbito microterritorial para todo el territorio nacional. Algunas instituciones llevan realizando, desde hace unos años, un esfuerzo en esta línea, como es el caso de los Anuarios Económico y Social elaborados por el Instituto L. R. Klein para “la Caixa” (“laCaixa”, 2001A y B). También debe destacarse la meritoria labor que realizan algunos institutos de estadística regionales para proporcionar información microterritorial de sus correspondientes comunidades autónomas (ver Tabla II.2.2). Debe advertirse, no obstante, que las bases de datos disponibles de forma gratuita para el ámbito microterritorial, en España, no siempre se ofrecen convenientemente depurados ni geocodificados, lo que dificulta especialmente el trabajo a este nivel.

A continuación, se ofrecen algunas direcciones de Internet importantes donde es posible encontrar, de forma más o menos geocodificada, información española de ámbito microterritorial. Como puede observarse, para toda España sólo es posible obtener datos estadísticos municipales en las “páginas web” del Instituto Nacional de Estadística (INE), la Dirección General del Catastro y “la Caixa”, que publica los datos del Instituto L. R. Klein. Mucho más amplia es la información municipal disponible para una única

comunidad autónoma, aunque como todas ellas no elaboran el mismo tipo de estadísticas, no es posible completar el conjunto nacional a partir de la suma de datos suministrada por cada una de las mismas.

Tabla II.2.1. Información microterritorial de España disponible en Internet con acceso gratuito

Institución	Contenidos	Página web
Información microterritorial de ámbito nacional		
Instituto Nacional de Estadística, INE	<ul style="list-style-type: none"> • Información municipal de España. 	www.ine.es
Dirección General del Catastro	<ul style="list-style-type: none"> • Información municipal de España (excepto País Vasco y Navarra). 	www.catastro.meh.es
“la Caixa”-Instituto L. R. Klein (UAM)	<ul style="list-style-type: none"> • Información municipal de España (municipios de más de 1000 habitantes). 	www.estudios.lacaixa.es/anuarioeconomico www.estudios.lacaixa.es/anuariosocial
Información microterritorial de ámbito regional (comunidades autónomas)		
Instituto de Estadística de Andalucía, IEA	<ul style="list-style-type: none"> • Información municipal de Andalucía. 	www.iea.junta-andalucia.es
Instituto de Estadística de Aragón	<ul style="list-style-type: none"> • Información municipal de Aragón. 	www.aragob.es
Instituto Balear de Estadística, IBAE	<ul style="list-style-type: none"> • Información municipal de Illes Balears. 	http://web2.caib.es/owa/g01.unidad_organica?p_coduo=994
Instituto Canario de Estadística, ISTAC	<ul style="list-style-type: none"> • Información municipal de Canarias • Callejero de Canarias. • Directorio de unidades económicas de Canarias y censo comercial. 	www.istac.rcanaria.es
Dir. Gral. de Estadística de la Junta de Castilla y León	<ul style="list-style-type: none"> • Información municipal de Castilla y León. 	www.cyl.es/jcyl/ceh/dge/
Instituto de Estadística de Cataluña, IDESCAT	<ul style="list-style-type: none"> • Información municipal de Cataluña. • Información por distritos de Cataluña. 	www.idescat.es
Instituto Valenciano de Estadística, IVE	<ul style="list-style-type: none"> • Información municipal de la Comunidad Valenciana 	www.ive.infocentre.gva.es
Junta de Extremadura	<ul style="list-style-type: none"> • Información municipal de Extremadura. 	http://www.juntaex.es/consejerias/eic/home.html
Instituto Galego de Estadística, IGE	<ul style="list-style-type: none"> • Información municipal de Galicia. 	www.ige.xunta.es
Instituto de Estadística de la Comunidad de Madrid	<ul style="list-style-type: none"> • Información municipal de la Comunidad de Madrid. • Información para los distritos y barrios del municipio de Madrid. 	www.comadrid.es/iestadis/index.htm

Institución	Contenidos	Página web
Ctro. Reg. de Estadística de la Región de Murcia	• Información municipal de la Región de Murcia.	www.carm.es/econet/
Instituto de Estadística de Navarra	• Información municipal de la Comunidad Foral de Navarra	www.cfnavarra.es/estadistica/confindex.asp?i=El+I.E.N.&p=presentacion.html&d=ienmenu.html
Instituto Vasco de Estadística, EUSTAT	• Información municipal del País Vasco.	www.eustat.es
Ceuta	• Información municipal de la Ciudad Autónoma de Ceuta.	www.ciceuta.es/orgecohac/procesa/Atlas/Index.htm
Melilla	• Información municipal de la Ciudad Autónoma de Melilla.	http://www.promesa.net/info/melilla/
Información microterritorial de ámbito municipal		
Ayuntamiento de Madrid	• Información para los distritos y barrios del municipio de Madrid.	http://www.munimadrid.es/principal/index.html

Fuente: Elaboración propia.

II.2.4. Problemas en el tratamiento de datos espaciales

Como indican Anselin (1999A) y Haining (1995), muchos de los análisis de tipo espacial que se realizan actualmente gracias a la creciente disponibilidad de datos geográficos, tienen que contrastarse con la literatura científica existente, dado que la gran dimensión de las bases de datos y la multitud de formas de obtenerlas dan lugar a un número de cuestiones que no siempre se resuelven satisfactoriamente y requieren de un mayor esfuerzo de investigación. Además, como ya se ha indicado, en la mayoría de los casos, el tratamiento de las bases de datos espaciales (sobre todo, las grandes bases microterritoriales) no debe llevarse a cabo directamente, tal como es proporcionada por los organismos o fuentes originales, sino que exige un tratamiento que haga posible la obtención de una información lo más precisa, depurada y operativa posible.

Por eso, se resumen a continuación algunos problemas relativos al tratamiento de datos espaciales, entre los que deben destacarse la elección de escala o nivel de agregación territorial, el proceso de obtención y almacenamiento informático de los datos y el análisis exploratorio y confirmatorio de la información microterritorial.

➤ **Elección de la escala o nivel de agregación territorial:**

La elección de la escala o ámbito espacial adecuados se ha convertido en un tema fundamental en la investigación espacial, máxime cuando en la actualidad se dispone con relativa facilidad de grandes bases de datos geocodificadas con información recogida desde ámbitos más o menos individuales (individuos, hogares, secciones censales) hasta niveles globales (regiones, países). Además, las cada vez más poderosas herramientas de los GIS

hacen posible al usuario pasar de una escala de análisis a otra, así como la integración de datos recogidos en diferentes ámbitos geográficos. Muchas veces, las observaciones procedentes de un nivel de análisis agregado (por ejemplo, países) no proporcionan necesariamente información útil sobre ámbitos inferiores (municipios, hogares), sobre todo cuando existe heterogeneidad espacial en el fenómeno que se analiza.

Otras veces sucede que algunos estadísticos cambian de valor según se obtengan en un ámbito u otro de agregación territorial, tal como se ha demostrado en los coeficientes de correlación y las medidas de autocorrelación espacial (Arbia, 1989). Es éste un problema metodológico ya viejo y familiar (que se desarrolla con más extensión en el Apartado V.II), conocido con diferentes denominaciones: los políticos y sociólogos hablan de “inferencia ecológica” (King, 1997), mientras que en geografía (Openshaw y Taylor, 1979) se trata del “problema de la unidad espacial modificable” (MAUP, “Modifyable Areal Unit Problem”), en estadística es la “estimación indirecta de unidades pequeñas” (Rao, 2000), y los geólogos lo denominan problema del “aumento de escala” o “cambio de soporte” (Chilès y Delfiner, 1999).

Otra cuestión relacionada tiene que ver con la diferencia existente entre el ámbito espacial del proceso a considerar y la escala geográfica concreta en la que se mide la observación, cosa que suele producirse muchas veces en estudios que combinan fenómenos físicos y económicos (Anselin, 2001A). Por ejemplo, en el análisis de los factores socioeconómicos que subyacen al proceso de deforestación tropical (Wood y Skole, 1999), la variable dependiente, la deforestación, se calcula a partir de la interpretación de imágenes vía satélite que dan lugar a datos correspondientes a enrejados regulares de cuadrados o “pixels”. Este enrejado no se corresponde, ni en el tamaño ni en las fronteras geográficas, con las unidades administrativas de las que proceden las variables estadísticas de censos o encuestas. Como normalmente las unidades administrativas son, en estos casos, las de mayor tamaño, son también las que determinan el mínimo ámbito geográfico de análisis, lo que conlleva la agregación (o alisado) de los valores “pixel” y la pérdida de variabilidad espacial en cada unidad geográfica finalmente utilizada. Y ya se sabe que, cuanto mayor es la variabilidad (es decir, cuanto menor es la escala espacial en la que se trabaja), menos precisa será la información agregada como estimador de la variable dependiente. Es más, la variabilidad interna no siempre permanece constante entre unidades, lo que da lugar a problemas de heteroscedasticidad.

Es éste un problema aún abierto y de difícil solución en muchas situaciones reales, sobre todo cuando el ámbito geográfico de disponibilidad de datos es muy agregado (nivel regional o provincial), como sucede con algunas macromagnitudes económicas de difícil cálculo directo, como el VAB o la Cuenta de los Hogares. En estos casos, la necesidad de estadísticas para ámbitos territoriales inferiores (municipios, distritos o secciones censales de grandes ciudades) constituye un verdadero reto por encontrar una solución al problema de la predicción espacial, como se verá en el Capítulo V.

➤ **Proceso de obtención y almacenamiento informático de los datos:**

Una vez decidido el ámbito espacial de trabajo, deberán afrontarse los problemas relacionados con el proceso de recogida y tratamiento informático de los mismos. En esta

línea, Haining (1995) apunta como problemas a solucionar la identificación y corrección de errores, valores poco fiables o no disponibles para algunas unidades espaciales, los posibles errores de muestreo y la correcta asignación de datos a su correspondiente unidad geográfica.

Efectivamente, en la mayoría de los casos, la información procedente de un determinado organismo público o privado adolece de errores u omisiones que deben ser previamente detectados y depurados, según unos determinados criterios. Cuando sea posible habrá de procurarse la obtención de la información correcta, solicitándola al propio organismo generador o a otro que disponga también de dicha estadística (como sucede en el caso de la información municipal con los ayuntamientos). En caso contrario, podrán aplicarse modelos ingenuos que estimen los datos incorrectos o inexistentes. En concreto, deberán estimarse también los datos estadísticos correspondientes a entidades geográficas de nueva creación (municipios, distritos, secciones, etc.), dado que muchas fuentes de información ignoran la existencia de dichas entidades.

Otras veces se producen errores de muestreo espacial, como se pone de manifiesto en Anselin (1999A). La estratificación de las encuestas en las que se basan las estadísticas y estudios socioeconómicos habituales, suelen ignorar el papel de los efectos espaciales y pueden, por tanto, ser inadecuadas para los propósitos del análisis espacial. En el muestreo espacial debe tenerse en cuenta que el fenómeno de autocorrelación espacial suele estar presente en muchas variables geográficas, lo que obligará, en estos casos, a seleccionar en la muestra observaciones “suficientemente alejadas” unas de otras (o incorrelacionadas), para que puedan ser consideradas como representativas de una población.

Además, el tratamiento de la información microterritorial procedente de muchos organismos generadores de estadísticas, exige muchas veces la identificación y agrupación de direcciones y localidades pertenecientes a una determinada área poblacional, tomada como unidad de investigación (municipio, distrito, barrio, etc.).

Por último, pueden producirse también problemas relacionados con el proceso de informatización de datos, como los errores de teclado, redondeo, cálculo, geocodificación, etc. El análisis de las grandes bases de datos microterritoriales exige un conocimiento suficiente de los programas informáticos gestores de bases de datos (Microsoft Access, dBase IV) y hojas de cálculo (Microsoft Excel), a los que el investigador deberá recurrir siempre que sea posible. La introducción manual de datos deberá ser la última opción, debido al riesgo más que probable de errores de teclado cuando se trabaja con un volumen grande de información. Hay que tener en cuenta que, en muchas ocasiones, la información facilitada por las fuentes estadísticas se encuentra sin codificar o con una codificación diferente de la establecida por el Instituto Nacional de Estadística (INE), en cuyo caso, habrá de ser geocodificada para hacer posible un tratamiento posterior por programas gestores de bases de datos o GIS.

➤ **Análisis exploratorio y confirmatorio de la información microterritorial**

El tamaño de las bases de datos geoespaciales disponibles constituye también un reto para el análisis espacial. La mayor parte de las técnicas tradicionales de análisis de

datos espaciales fueron inicialmente desarrolladas para situaciones en las que los conjuntos de datos tuvieran menos de 100 observaciones¹. Sin embargo, lo normal hoy día es trabajar con bases de datos muy superiores, como sucede con el análisis exploratorio denominado “minería de datos espaciales”, fundamental para la detección de estructuras, agrupaciones, asociaciones y otras interesantes situaciones de no aleatoriedad geográfica. Aunque muchas de las técnicas actualmente en uso, como los indicadores LISA, son en principio aplicables, su implementación en bases de datos muy grandes puede constituir un problema desde el punto de vista informático o conceptual. Por este motivo, el papel de la geocomputación o geoinformática ha llegado a ser más importante que nunca, dado que muchos problemas espaciales son intrínsecamente de orden N^2 (por ejemplo, las matrices de pesos espaciales) por lo que no pueden ser abordados de una forma eficaz a través de los sistemas informáticos habituales.

Otra cuestión relacionada con las bases de datos microterritoriales de gran tamaño es la elección del paradigma inferencial. Es decir, la teoría asintótica clásica sobre la que se asientan los métodos de estadística y econometría espacial ha sido desarrollada para obtener las propiedades de los estimadores y contrastes estadísticos en muestras finitas, pero no tienen sentido cuando la muestra utilizada prácticamente tiende a infinito. Por eso, en los últimos años, se han desarrollado otros paradigmas prometedores basados en nociones bayesianas o de puro cómputo, como los estimadores de simulación, los métodos de remuestreo o las aproximaciones permutativas, aunque su implementación para bases de datos territoriales muy extensas está aún lejos de conseguirse satisfactoriamente.

Todas estas cuestiones ponen, por tanto, de manifiesto la necesidad de una adecuada formación para llevar a cabo cualquier análisis de datos espaciales, sobre todo, en el ámbito microterritorial. Como se ha expuesto, quedan aún problemas técnicos o científicos sin una solución satisfactoria. Sin embargo, la demanda de información y análisis microterritorial es cada vez mayor, por lo que debe estimularse la investigación interdisciplinar sobre en temas relacionados con esta materia: econometría y estadística espacial, geoinformática, geografía cuantitativa, geomarketing, geopolítica, etc.

¹ Por ejemplo, el clásico ejemplo de datos provinciales irlandeses, utilizado por Cliff y Ord (1973) para ilustrar los contrastes de autocorrelación espacial y los modelos autorregresivos, sólo tenía 25 observaciones.

CAPÍTULO III

Análisis Exploratorio de Datos Espaciales

Este capítulo está dedicado al análisis exploratorio de datos espaciales (AEDE), que es la etapa previa al análisis confirmatorio o modelización econométrica espacial. Para ello, a modo de introducción, en el primer apartado, se presenta el concepto de análisis exploratorio de datos (AED), en el que se encuentra inscrito el AEDE, entendido este último como grupo de técnicas que describen y visualizan las distribuciones espaciales, identificando localizaciones atípicas, descubriendo esquemas de asociación (autocorrelación espacial) y sugiriendo estructuras en el espacio geográfico (heterogeneidad espacial). A continuación, el apartado segundo del capítulo estará dedicado a la exposición, formulación y contraste del efecto espacial de dependencia o autocorrelación espacial “univariante”, es decir, sobre una variable propiamente dicha. El tratamiento de los efectos de dependencia y heterogeneidad espacial en modelos de regresión será presentado con detalle en el Capítulo IV.

El esquema del capítulo es el siguiente:

CAPÍTULO III. ANÁLISIS EXPLORATORIO DE DATOS ESPACIALES

III.1. Introducción al análisis exploratorio de datos espaciales (AEDE)

III.1.1. Análisis exploratorio de datos

III.1.2. AEDE y paquetes informáticos especializados

III.1.3. Métodos gráficos del AEDE

III.2. Dependencia o autocorrelación espacial

III.2.1. Concepto

III.2.2. Formulación matemática del fenómeno de dependencia espacial

III.2.3. Perspectivas en el análisis de autocorrelación espacial: global y local

III.1. INTRODUCCIÓN AL ANÁLISIS EXPLORATORIO DE DATOS ESPACIALES (AEDE)

En muchas ocasiones, el análisis de datos se realiza con el objetivo de contrastar determinadas hipótesis o modelos como, por ejemplo, cuando se quiere conocer a través de una encuesta si la edad de las personas está relacionada positivamente con su capacidad de asumir riesgos o si, en situaciones más complicadas, se desea conocer si la relación entre un determinado número de variables es lineal o no lineal. Sin embargo, otras veces, el investigador no dispone de una hipótesis definida a contrastar, por lo que la naturaleza del estudio es “exploratoria” y se deben utilizar ciertas técnicas especialmente diseñadas para encontrar algún “sentido” en los datos, es decir, para la detección en ellos de estructuras o conglomerados (“clusters”).

III.1.1. Análisis exploratorio de datos

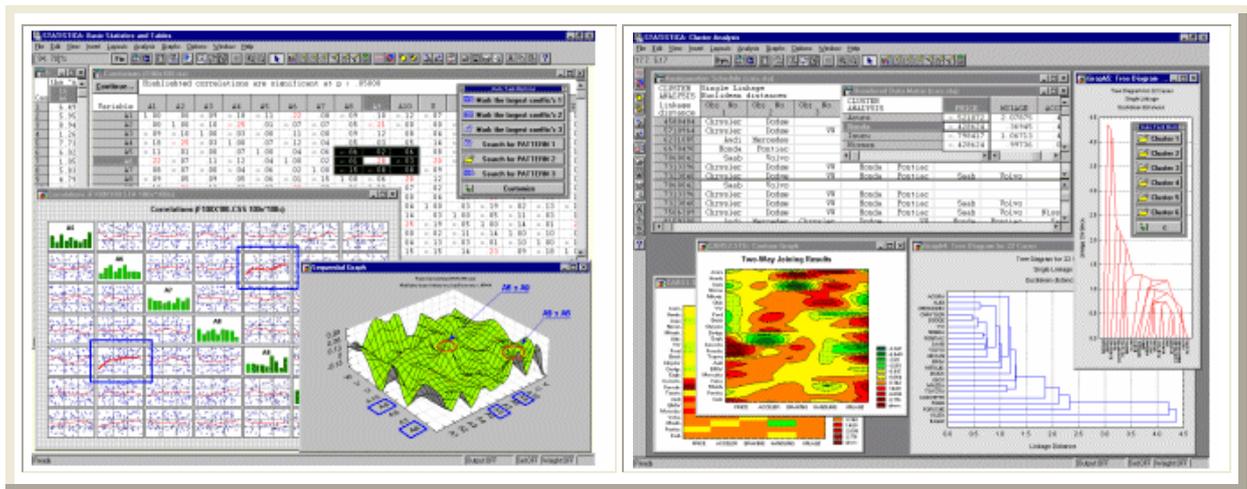
El análisis exploratorio de datos (AED), a diferencia de los contrastes de hipótesis tradicionales diseñados para la verificación a priori de conexiones entre variables, se utiliza para identificar relaciones sistemáticas entre variables cuando no existen expectativas claras sobre la naturaleza de estas relaciones. El clásico AED consiste en el tratamiento y comparación de una batería de variables utilizando técnicas que identifican en ellas formas estables. El AED podría definirse como “el conjunto de herramientas gráficas y descriptivas utilizadas para el descubrimiento de patrones de comportamiento en los datos y el establecimiento de hipótesis con la menor estructura posible”. Esta definición fue propuesta por Tukey (1977), que hizo posible la extensión de este tipo de análisis multivariante y del “software” estadístico moderno (Anselin, 1998; Anselin y Bao, 1997; Moreno y Vayá, 2000).

De este modo, es posible obtener una estructura explicativa de los datos a través de unas técnicas que combinan herramientas de la estadística básica (descriptivos, correlaciones, tablas de frecuencias o de correlación cruzada) con análisis multivariante avanzado, especialmente diseñado para identificar formas en grandes bases de datos (análisis cluster, escalas multidimensionales, análisis logit, correlación canónica, análisis de correspondencias, árboles de clasificación, etc.).

Las técnicas de AED suelen estar acompañadas de métodos de visualización gráfica capaces de identificar conexiones, tendencias o sesgos presentes en las bases de datos inicialmente desestructuradas. Una de las más conocidas técnicas propias del AED es el “cepillado” o “brushing”, un método interactivo que permite la selección en pantalla de un determinado número de datos para la identificación de características comunes o el examen de efectos o relaciones entre variables relevantes. Estas relaciones entre variables suelen visualizarse mediante el ajuste de funciones (líneas, en un plano de dos dimensiones o superficies tridimensionales) y sus intervalos de confianza, de forma que, por ejemplo,

puedan examinarse los cambios que producen sobre dichas funciones la eliminación o añadido (temporal) de un determinado número de datos. En la Figura III.1.1 (izquierda), se presenta una de las muchas aplicaciones de la técnica del “brushing” que consiste en la selección destacada en una matriz de dispersión de datos pertenecientes a una determinada categoría.

Figura III.1.1. Ejemplo de herramientas de AED: grandes matrices de correlación (izquierda) y análisis multivariante (derecha)

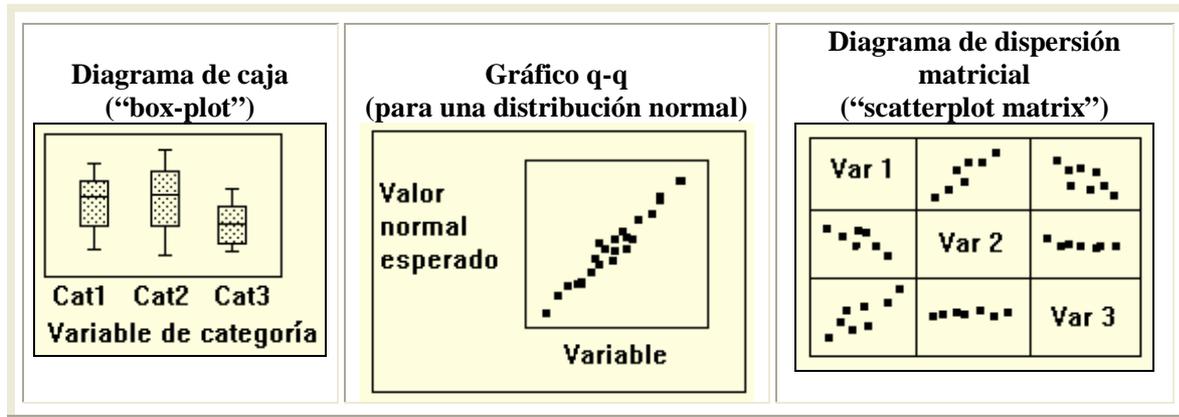


Fuente: Elaboración propia a partir de Statistica (2000).

Otras técnicas gráficas del AED incluyen el ajuste y representación gráfica de funciones, alisado de datos, superposición o fusión de múltiples “vistas” (“views”), añadido de datos en gráficos, identificación y señalización de subconjuntos de datos que cumplen determinadas condiciones, representación gráfica de intervalos o áreas de confianza, técnicas de reducción de imágenes, etc. La exploración de los datos es sólo un primer paso y sus resultados deben ser considerados como una etapa previa a su “confirmación” (modelización). Si los resultados de la fase exploratoria sugieren la utilización de un modelo determinado, entonces su validación puede verificarse aplicando dicho modelo a un nuevo conjunto de datos y contrastando el ajuste, es decir, su capacidad predictiva.

Es decir, los modernos métodos de AED hacen hincapié en la interacción existente entre el conocimiento humano y la informática en forma de gráficos estadísticos dinámicos que permiten al usuario la manipulación directa de diversas “vistas” de los datos. Ejemplos de este tipo de “vistas” son algunos instrumentos gráficos como los histogramas, diagramas de caja (“box plot”)², gráficos q-q³ o los diagramas de dispersión matricial (“scatterplot matrices”)⁴, tal como se muestra en la Fig. III.1.2.

² Los diagramas de caja resumen una única variable numérica en categorías de otra variable, de forma que en cada cuadro se muestra la mediana, los cuartiles y los valores extremos contenidos en una categoría. También pueden resumir una o más variables numéricas.

Figura III.1.2. Ejemplos de diagramas del análisis exploratorio de datos (AED)

Fuente: Elaboración propia a partir de SPSS (1999).

Algunos programas clásicos de AED, como Matlab o SPSS han ido incorporando mapas como nuevas "vistas" integradas de forma dinámica con el resto de información, a través de sus extensiones Mapping Toolbox y SPSS Mapas, respectivamente. Sin embargo, la importancia concedida por ellos al espacio se reduce prácticamente al tratamiento de la localización como un simple plano de coordenadas X-Y, dado que las herramientas propias del AED no suelen ser diseñadas para el tratamiento específico de datos geográficos, ignorando totalmente los efectos espaciales de dependencia y heterogeneidad. Por este motivo, surge el análisis exploratorio de datos espaciales (AEDE), como disciplina dentro del AED, para el tratamiento específico de los datos geográficos.

III.1.2. AEDE y paquetes informáticos especializados

El AEDE puede definirse como el conjunto de técnicas que describen y visualizan las distribuciones espaciales, identifican localizaciones atípicas o "atípicos espaciales" ("spatial outliers"), descubren esquemas de asociación espacial, agrupamientos ("clusters") o puntos calientes ("hot spots") y sugieren estructuras espaciales u otras formas de heterogeneidad espacial (Anselin, 1999A). Por tanto, el AEDE tiene un carácter descriptivo (estadístico) más que confirmatorio (econométrico) pero, tras la detección de estructuras espaciales en las variables geográficas, hará posible la formulación de hipótesis previas

³ Gráfico con los cuantiles de distribución de una variable respecto a los cuantiles de una distribución de referencia (por ejemplo, beta, chi-cuadrado, exponencial, gamma, semi-normal, normal, t de Student). Los gráficos de probabilidad suelen emplearse para determinar si la distribución de una variable coincide con otra dada: si la variable seleccionada coincide con la distribución de referencia, los puntos se concentran en torno a una línea recta (SPSS, 1999).

⁴ En un diagrama de dispersión matricial el número de filas y columnas es igual al número de variables de la matriz seleccionadas. Se muestran todas las combinaciones posibles de variables (Var 1 vs. Var 2, y también "intercambiadas" Var 2 vs. Var 1).

para la modelización econométrica y, en su caso, posible predicción espacial de nuevos datos.

Es decir, el AEDE debe constituir la fase previa a toda modelización econométrica espacial, sobre todo cuando no exista un marco formal o teoría previa acerca del fenómeno que se pretende explicar, como es el caso de algunos análisis interdisciplinarios realizados en las ciencias sociales, sobre todo en el campo de la economía regional (Tabla III.1.1), así como en los ejercicios de predicción-extrapolación de datos (Capítulo V) en los que puede no existir una identificación entre relaciones económicas establecidas en diversos ámbitos o escalas territoriales.

Tabla III.1.1. Algunos ejemplos recientes de AEDE aplicado a las ciencias sociales

Referencia	Materia	Fenómeno explorado
Ceccato <i>et al.</i> , 2001	Criminología	Tasa de delitos cometidos en la ciudad de Estocolmo (1982).
Baller <i>et al.</i> , 2001	Criminología	Tasa de homicidios en los condados americanos (1960-1990).
Button y Kulkarni, 2001	Economía regional	Actividad de las empresas de transporte y logística en el área metropolitana de Washington (1999).
López <i>et al.</i> , 2001	Economía regional	Tasa de paro en las provincias españolas (1990).
Rey, 2001	Economía regional	VAB por ocupado en los estados americanos (1929-1999).
Stough <i>et al.</i> , 2001	Economía regional	Actividad de las empresas TIC en el área metropolitana de Washington.
Chasco y Vicéns, 2000	Economía regional	Renta disponible per cápita de las provincias españolas (1986-1998).
LeSage y Kelley, 2000	Economía regional	Tasa de ocupación en grandes industrias presentes en algunos distritos postales de Ohio (1956-1987)
Moreno y Vayá, 2000	Economía regional	VAB por ocupado en las regiones europeas (1975-1992).
Wall y Devine, 2000	Sanidad	Tasas de mortalidad por cáncer de pulmón, en los condados de Ohio (1968-1992).
Kristensen, 1998	Economía regional	Precio del suelo urbano en las secciones de la ciudad de Copenhague (1977).

Fuente: Elaboración propia.

El AEDE combina el análisis estadístico con el gráfico, dando lugar a lo que podría denominarse una “visualización científica” (Haining *et al.*, 2000) que, a los contrastes

estadísticos sobre los efectos espaciales de dependencia y heterogeneidad, un amplio marco de gráficos o “vistas” múltiples y dinámicas sobre la información geográfica (Unwin, 2000). Por eso, estos métodos de visualización científica son mucho más que simples mapas o gráficos estáticos de representación, como podrá apreciarse con mayor detalle en el Apartado siguiente, siendo fundamental el papel jugado por los paquetes informáticos especializados.

En los últimos años, los esfuerzos desarrollados por la investigación en este campo del AEDE se han venido centrando en la conexión de los GIS disponibles en el mercado (Grassland, ESRI, MapInfo) con paquetes estadísticos tradicionales (S-PLUS, SPSS) o específicos de AEDE (ver Tabla III.1.2). De este modo, a la potente capacidad de visualización y análisis de los GIS se le une la especialización propia del análisis espacial estadístico y gráfico (ver Bao *et al.*, 2000).

Tabla III.1.2. Paquetes informáticos especializados en el análisis de datos espaciales exploratorio y confirmatorio

Análisis	Empresa/autor	Página web	Paquete informático
AEDE Perspectiva geoestadística	TerraSeer/L. Anselin	www.spacestat.com	SpaceStat
	D. Swayne, D. Cook y A. Buja	www.xplore-stat.de/tutorials/rpcframe7.html	XGobi/ ArcView/ XploRe
	ESRI Geosistemas	www.esri.com www.esri-españa.es	ArcGIS Network y Spatial Analyst
	J. Ma, B. Haining y S. Wise	www.shef.ac.uk/~scgisa	Sage
	J.P. LeSage	www.spatial-econometrics.com	Econometrics Toolbox
	MathSoft, Inc.	www.mathsoft.com	S+SpatialStats
	MIMAS/J. Dykes	www.geog.le.ac.uk/jad7/cdv/	cdv
	Northwood Geoscience Ltd.	www.northwoodgeo.com	Vertical Mapper
	Spatial Insights, Inc.	www.spatialinsightful.com	TrendMap
	TerraSeer Enviromental Insight P. Wall y O. Devine	www.terraser.com -	BoundarySeer ClusterSeer MapSpat
AEDE Perspectiva retícula o “lattice”	Biomedware/L. Anselin	www.spacestat.com (en desarrollo)	SpaceStat SpaceStat para ArcView DynESDA DynESDA2.0
	D. Swayne, D. Cook y A. Buja	www.xplore-stat.de/tutorials/rpcframe7.html	XGobi/ ArcView/ XploRe
	Insightful Corporation	www.insightful.com	S-Plus para ArcView GIS

Análisis	Empresa/autor	Página web	Paquete informático
	J. Ma, B. Haining y S. Wise	www.shef.ac.uk/~scgisa	Sage
	J.P. LeSage	www.spatial-econometrics.com	Econometrics Toolbox
	K. Pace y R. Barry	www.spatial-statistics.com	SpaceStatPack Spatial Statistics Toolbox
	MathSoft, Inc.	www.mathsoft.com	S+SpatialStats

Fuente: Elaboración propia.

En la Tabla III.1.2 se presentan algunos importantes programas de “software” relacionados con el análisis geográfico de datos: todos ellos tienen en común el tratamiento específico de variables espaciales (AEDE) que, de acuerdo con Cressie (1993), puede ser abordado desde dos puntos de vista, según que se trate del análisis de una muestra de datos puntuales procedentes de distribuciones espaciales continuas (perspectiva geoestadística) o de localizaciones geográficas discretas de puntos o polígonos (perspectiva de retícula o “lattice”⁵).

Además, en Okabe y Funamoto (2000), pueden encontrarse referencias sobre otros paquetes de AEDE desarrollados en el campo de la ingeniería o la sanidad, como VCM (Variable Clumping Method), GAM (Geographical Analysis Machine), SPIDER (Spatial Interactive Data Explorer), SIMPLE (Spatial and Inductive Modelling Package por Land Evaluation), MINITAB o GLIM⁶.

Cada uno de estos enfoques implica métodos de exploración diferentes aunque los conceptos generales que manejan sean los mismos (ver artículos recientes en Anselin, 2001A, Anselin *et al.*, 2001; Bao *et al.*, 2000; Haining *et al.*, 2000; Wall y Devine, 2000). Como puede observarse en la Tabla III.1.2, hay programas de AEDE, como ArcGIS, Cdv, MapSpat, Network Analyst, Spatial Analyst, TrendMap o Vertical Mapper, que están orientados únicamente a la estadística espacial o perspectiva geoestadística, mientras que hay otros paquetes que abordan conjuntamente ambos métodos (geoestadístico y reticular o “lattice”), como los paquetes de TerraSeer, que desde hace meses incluyen también SpaceStat y sus extensiones, Cdv, Econometrics Toolbox, Sage, S+SpatialStats y XGobi.

Debe, sin embargo, destacarse la marcada orientación hacia el análisis econométrico espacial (exploratorio y confirmatorio) del “software” SpaceStat y sus extensiones para AcView (Anselin), Econometrics Toolbox (LeSage), Sage (Ma, Haining y Wise), SpaceStatPack y Spatial Statistics Toolbox (Pace y Barry), S+SpatialStats (MathSoft) y S-Plus para ArcView (Insightful Corporation), que permiten también la estimación, contraste y validación de diversos tipos de modelos espaciales de regresión lineal, modelos de

⁵ Haining *et al.* (2000) define los datos reticulares o “lattice” como aquéllos referidos a regiones de formas no regulares en el espacio (provincias, municipios, etc.).

⁶ Efectivamente, podrían realizarse otras muchas aplicaciones especializadas en el análisis de datos espaciales, como la extensa relación que, para el campo de la ecología propone Legendre (1993).

variables dependientes limitadas (logit, probit, tobit), modelos de expansión o estructuras espaciales, modelos bayesianos, espacio-temporales, etc.

El programa “SpaceStat”, desarrollado por Luc Anselin (1992D, 1995A), como ya se ha indicado, destaca en todo este panorama por su dedicación específica a los métodos exploratorios y confirmatorios propios de la econometría espacial. SpaceStat no es un paquete estadístico amplio, en el sentido tradicional del término, dado que se han omitido intencionadamente muchos tests y métodos fácilmente disponibles en programas actualmente en el mercado (SPSS, E-Views, SAS, SCA, etc.). El objetivo de este programa es centrarse en aquellas técnicas estadísticas espaciales que implican una inversión considerable en términos de programación, en vez de incluir las que ya se encuentran disponibles. Este programa se encuentra escrito y compilado en GAUSS, que es un lenguaje matricial para técnicas estadísticas de Aptech Systems, Inc., de forma que puede desarrollarse tanto sobre el propio programa GAUSS como sobre un módulo o “runtime” del mismo incluido en la propia versión de SpaceStat.

El menú principal de SpaceStat 1.90 consta de cuatro puntos, cada uno de los cuales está a su vez subdividido: DATA (creación y tratamiento de bases de datos), TOOLS (tratamiento de pesos espaciales), EXPLORE (medidas de asociación espacial) y REGRESS (análisis de regresión espacial). El contenido básico de este programa incluye algunos estadísticos descriptivos básicos, el análisis espacial de componentes principales, los principales contrastes estadísticos globales y locales de autocorrelación espacial, métodos de estimación de modelos de regresión lineal con contrastes para los efectos espaciales, estimación de modelos con autocorrelación espacial, así como otros modelos espaciales (superficie tendencial, estructuras espaciales, expansión espacial, ANOVA espacial, modelos de interacción espacial).

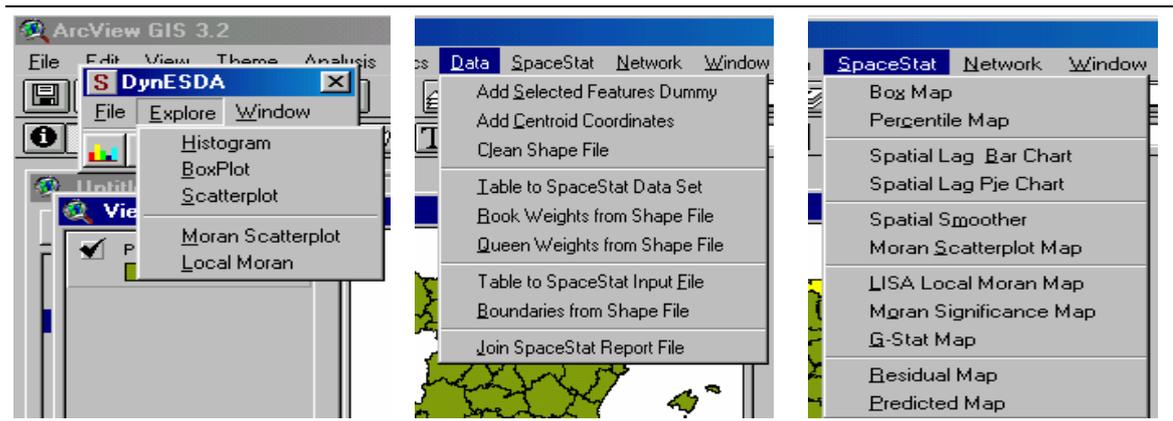
Además, Anselin ha desarrollado dos extensiones para el GIS ArcView, la extensión de SpaceStat para ArcView propiamente dicha y DynESDA (Anselin, 1999B, 2000A), así como una nueva versión DynESDA2.0, para MapObjects de ESRI (Anselin *et al.*, 2001). Estas aplicaciones tienen como objetivo facilitar el intercambio y visualización de datos de algunos resultados obtenidos por SpaceStat susceptibles de representación geográfica, haciendo especial hincapié en los métodos del AEDE.

La extensión de SpaceStat para ArcView consiste en un conjunto de librerías (DLL) escritas en lenguaje C++ que se presenta al usuario en la barra superior de ArcView, en forma de dos menús adicionales denominados “Data” y “SpaceStat” (ver Figura III.1.3).

El menú “Data” tiene por principal misión la obtención automática, por parte de ArcView, de información (ficheros de datos y matrices de pesos espaciales) procedente de SpaceStat, incluyendo también un pequeño número de funciones destinadas al cálculo de variables con valor espacial, como las coordenadas X-Y de los centroides de las regiones geográficas, necesarias para la construcción de algunas matrices de pesos espaciales. Por último, este menú permite también la exportación a formato Ascii (accesible desde SpaceStat) de bases de datos procedentes de ArcView.

El menú “SpaceStat” permite la visualización dinámica, en ArcView, de resultados obtenidos por SpaceStat, como los puntos atípicos espaciales, percentiles de distribuciones espaciales, retardos espaciales, funciones de alisado espacial, estadísticos de asociación espacial global y local, valores estimados y residuos procedentes de un análisis de regresión espacial.

Figura III.1.3. Extensión de SpaceStat para el GIS ArcView y DynESDA



Fuente: Elaboración propia a partir de la extensión de SpaceStat para ArcView.

Por su parte, la extensión DynESDA para ArcView ha dotado de dinamicidad el AEDE, de forma que un mapa constituya una de las varias “vistas” posibles de una base de datos, junto con una tabla, histograma, gráfico de caja u otro gráfico de carácter estadístico. Estas vistas se encuentran conectadas de forma que cualquier observación señalada en una de ellas queda simultáneamente destacada en el resto.

Además de gráficos estadísticos tradicionales, DynESDA incluye un visualizador para el fenómeno de asociación espacial que permite el recálculo interactivo de los estadísticos de asociación para subconjuntos de datos, así como contrastes de influencia y atípicos espaciales. Esta aplicación ha sido diseñada para el tratamiento de datos reticulares o “lattice” (unidades poligonales), en lugar de puntos, como sucede en la perspectiva geoestadística.

Por último, la extensión DynESDA2.0, aún en desarrollo, está construida en lenguaje C++ a partir de elementos del programa MapObjects, de ESRI, de forma que no sea necesaria la utilización de ArcView ni de cualquier otro GIS, de mucho más complejo uso (Anselin *et al.*, 2001). El entorno de MapObjects permite la visualización (no la manipulación) de datos espaciales, procedentes de diversos formatos, a través de ventanas unidas de forma dinámica, así como la combinación de éstos con gráficos estadísticos clásicos, como histogramas, diagramas de caja y diagramas de dispersión. Además, se incluyen herramientas del AEDE reticular o “lattice”, como los diagramas de dispersión de Moran y los mapas LISA.

III.1.3. Métodos gráficos del AEDE

En los últimos años, se han propuesto gran cantidad de métodos gráficos para el AEDE aunque, como ponen de manifiesto Haining *et al.* (2000), existen pocos estudios que valoren la utilidad y efectividad de todos ellos. En línea con Wise *et al.* (1999), tal como propone Tukey (1977) para el AED, podría afirmarse que un buen método gráfico de AEDE es aquél capaz de analizar y representar dos características fundamentales en toda distribución espacial: alisado (“smooth”) y asperezas (“rough”).

Por un lado, la propiedad de alisado, que en el campo temporal serían la tendencia central de la variable (medida a través de la mediana) y su dispersión (recorrido intercuartílico), en las variables geográficas incluiría elementos globales, referidos a todo el mapa, como la tendencia espacial y autocorrelación espacial global. Por otro, la propiedad que podría denominarse como asperezas (“rough”) de la distribución se refiere a aquellos datos localizados a cierta distancia del elemento de alisado (mediana), como los llamados atípicos (“outliers”), situados bajo/sobre el primer/tercer cuartil de un diagrama de caja.

Esta última propiedad es local, a diferencia de la propiedad de alisado que es global, por lo que incluye, por ejemplo, casos que se revelan como muy distintos de sus correspondientes valores vecinos en el mapa (atípicos espaciales), regiones que se encuentran agrupadas en forma de valores altos o bajos de una variable produciendo autocorrelación espacial local (agrupamientos o “clusters” de zonas calientes o frías) o incluso, líneas de discontinuidad geográfica (heterogeneidad espacial).

En la Tabla III.1.3, se presentan las principales técnicas del AEDE que, como puede apreciarse, contemplan los dos citados elementos de alisado (global) y asperezas (local), desde una perspectiva geoestadística y reticular o “lattice”.

Esta taxonomía de métodos del AEDE se inspira en otra similar propuesta por Anselin (1998) y recogida por Moreno y Vayá (2000), en la que se diferencian los métodos de representación de distribuciones espaciales, por un lado y, por otro, los específicos de visualización del fenómeno de asociación espacial (global, local y multivariante), a los que se han añadido otros gráficos y el análisis exploratorio propio del efecto de heterogeneidad espacial.

Estos métodos, que han recibido gran atención en la literatura científica reciente, han sido también implementados en varios de los programas informáticos expuestos en la Tabla III.1.2. Dada la profusión de gráficos y métodos diseñados, se ha optado por dar prioridad al enfoque reticular (“lattice”) propio de los métodos de econometría espacial, dejando el desarrollo de las técnicas propias de la perspectiva geoestadística para futuras investigaciones.

Tabla III.1.3. Técnicas del análisis exploratorio de datos espaciales (AEDE)

		Perspectiva geoestadística	Perspectiva reticular o “lattice”
Distribución espacial	Univariante	<ul style="list-style-type: none"> • Función de distribución acumulada espacial • Diagrama/mapa de puntos • Mapa de círculos 	<ul style="list-style-type: none"> • Diagrama/mapa de caja
	Multivariante	<ul style="list-style-type: none"> • Cartograma • Gráfico de coordenadas paralelas 	<ul style="list-style-type: none"> • Diagrama dispersión-caja
Asociación espacial	Global	<ul style="list-style-type: none"> • Diagrama de dispersión espacialmente retardado • Nube del variograma • Diagrama de caja del variograma 	<ul style="list-style-type: none"> • Mapa de contigüidades espaciales • Gráfico del retardo espacial • Diagrama/mapa de dispersión de Moran
	Local	<ul style="list-style-type: none"> • Puntos atípicos en el diagrama de caja del variograma • Puntos atípicos en la nube del variograma 	<ul style="list-style-type: none"> • Puntos atípicos en el diagrama de dispersión de Moran • Mapas LISA • Diagrama de caja LISA
	Multivariante	<ul style="list-style-type: none"> • Nube del variograma multivariante 	<ul style="list-style-type: none"> • Diagrama de dispersión multivariante de Moran
Heterogeneidad espacial			<ul style="list-style-type: none"> • Mapa del histograma de frecuencias • Diagrama de dispersión

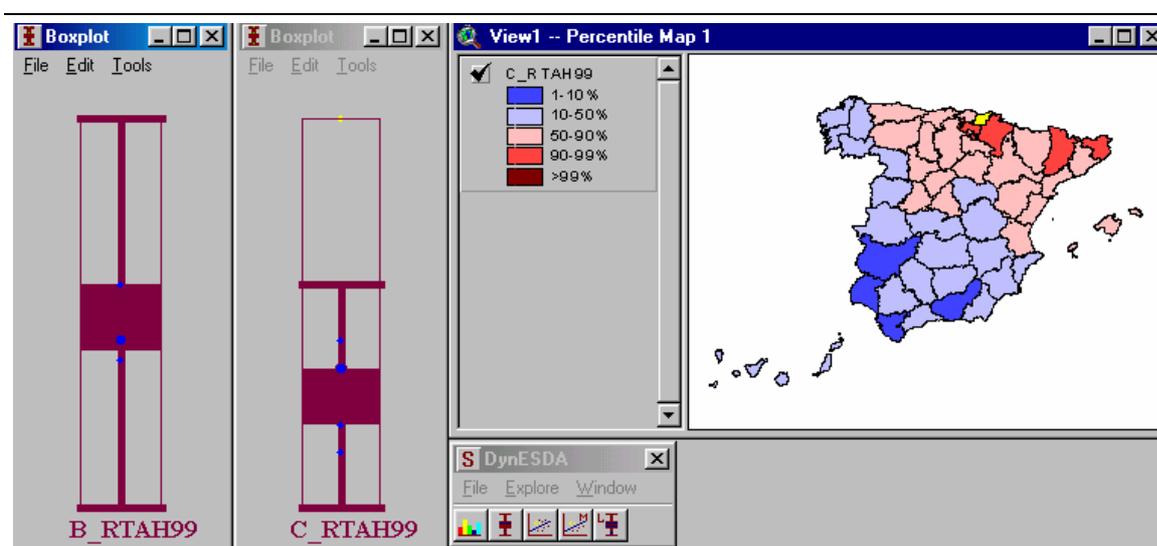
Fuente: Elaboración propia.

III.1.3.1. Métodos de representación de distribuciones espaciales

Las técnicas de visualización de distribuciones geográficas son muy parecidas a las herramientas de representación cartográfica, aunque el punto de partida es ligeramente distinto: para las segundas, el mapa es el elemento central, mientras que los procedimientos del AEDE se basan en elementos gráficos de la estadística clásica. En la primera fila de la Tabla III.1.3, se exponen los métodos de representación univariante más destacados, según que sean utilizados por la perspectiva geoestadística o por el enfoque reticular o “lattice” (diagrama/mapa de caja) Además, se han incluido algunos métodos novedosos de representación multivariante como el diagrama de dispersión-caja, para el que se ha utilizado el programa Cvd 1.1, especializado en el AEDE multivariante.

- **El diagrama/mapa de caja**⁷ es un método de representación univariante utilizado por la perspectiva reticular o “lattice”, que considera que los datos espaciales son realizaciones individuales de un proceso estocástico (como en el análisis de series temporales), haciendo posible la identificación de puntos atípicos o “outliers” espaciales. En el caso de trabajar en un sistema de AEDE dinámico, el usuario puede seleccionar varios puntos atípicos en un diagrama de caja dando lugar a zonas destacadas en el mapa.

Figura III.1.4. Diagrama y mapa de caja de la variable renta disponible por habitante de las provincias españolas en 1999



Fuente: Elaboración propia a partir de las extensiones SpaceStat y DynESDA para ArcView

La construcción del diagrama y mapa de caja parte del cálculo de los cuartiles y la media de una variable, así como de la obtención de las llamadas cotas o valores adyacentes superior e inferior, que se obtienen, a su vez, como el producto de los valores del tercer (primer) cuartil por 1,5 veces el recorrido intercuartílico. De esta forma, se consideran como valores atípicos aquéllos situados por encima (o por debajo) de dichas cotas (en la Figura III.1.4, se trata del diagrama de la variable C_RTAH99).

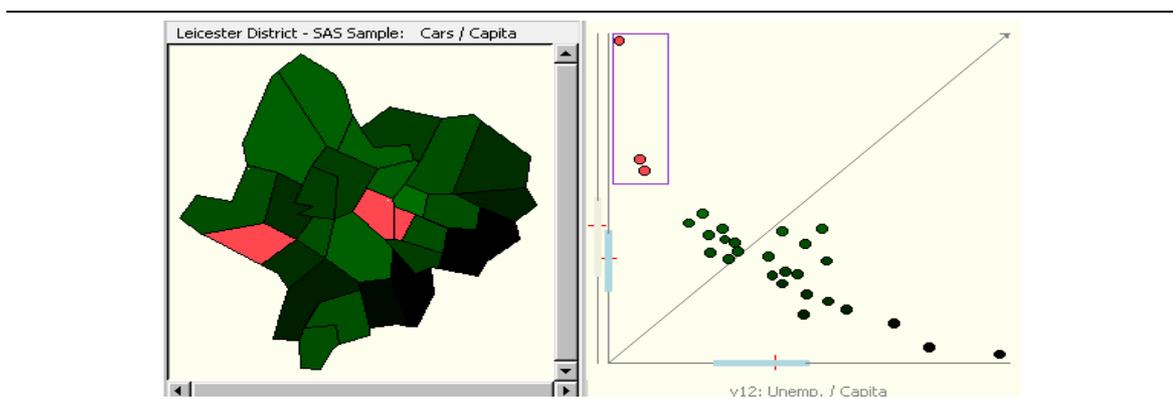
Un criterio un poco más estricto que el anterior consistiría en multiplicar por tres el recorrido intercuartílico para la fijación de las cotas (B_RTAH99). En el ejemplo de la Figura III.1.4, el único valor atípico (superior) de renta disponible por habitante sólo se produce en el caso menos estricto (C_RTAH99) correspondiéndose con la provincia de Guipúzcoa (destacada en color amarillo).

⁷ Los mapas de caja representan mediante distintos colores las unidades geográficas cuyos datos en una variable coinciden con la mediana, el rango intercuartil y valores atípicos.

tanto en el gráfico como en el mapa), no existiendo atípicos por debajo de la cota inferior.

- Por su parte, el **diagrama de dispersión/caja** es una técnica propia del AEDE multivariante que consiste en un diagrama de dispersión para dos variables en el que se representan también los correspondientes diagramas de caja en los ejes, indicándose los valores de las medianas (en la Figura III.1.5 con una línea roja) y los cuartiles (caja azul). En esta Figura, el diagrama de dispersión, que muestra una fuerte relación negativa entre las variables tasa de paro y tasa de automóviles, se han destacado tres puntos atípicos, con altas tasas de paro y bajas tasas de automóviles, correspondientes a los distritos de Spinney Hill, North Braunstone y Wycliffe, del condado de Leicester.

Figura III.1.5. Diagrama de dispersión/caja (derecha) de las variables tasa de paro y tasa de automóviles de los distritos de Leicester



Fuente: Elaboración propia a partir del programa cvd 1.1.

III.1.3.2. Representación del fenómeno de dependencia espacial

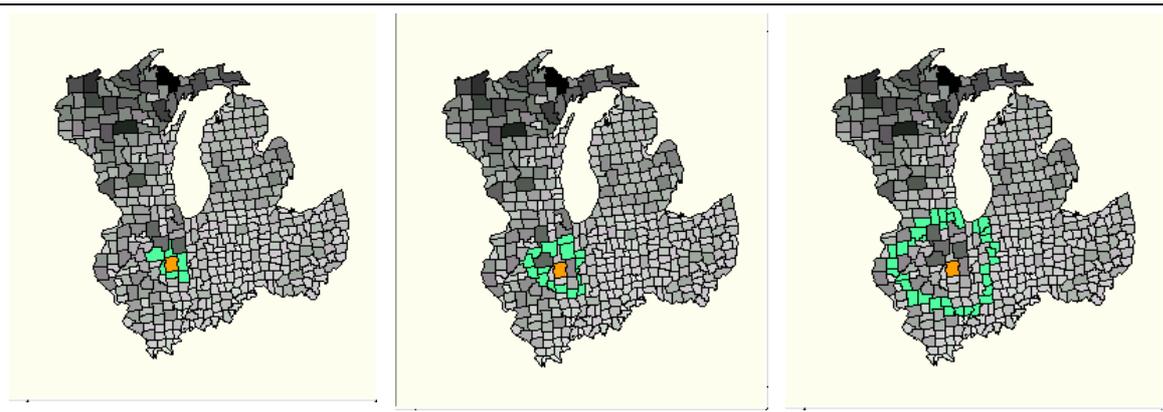
Los otros tres grupos de técnicas expuestas en la Tabla III.1.3 se encuentran dentro del concepto de asociación, dependencia o autocorrelación espacial (global, local y multivariante). En el Apartado II.1.1, la dependencia espacial era definida como el efecto que consiste en la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y en lugares cercanos o vecinos. Es decir, una variable se encontrará espacialmente autocorrelacionada cuando los valores observados en un punto o región dependan de los valores observados en regiones vecinas, de forma que se produzca una cierta continuidad geográfica en la distribución de esta variable, por ejemplo, sobre un mapa.

Las principales técnicas del AEDE de asociación espacial, pertenecientes a la perspectiva reticular o “lattice”, se encuentran en paquetes informáticos como SpaceStat, siendo posible también el dinamismo de dicho análisis con un “software” especial de

vinculación con ArcView, como DynESDA. Estas técnicas son las siguientes: mapa de contigüidades espaciales, gráfico del retardo espacial, diagrama y mapa de dispersión de Moran (asociación global), puntos atípicos en el diagrama de dispersión de Moran, mapas LISA, diagrama de caja LISA (asociación local) y diagrama de dispersión multivariante de Moran (asociación multivariante).

- Los contrastes de dependencia o autocorrelación espacial propios del enfoque reticular o “lattice” se basan en la noción de contigüidad binaria entre las unidades espaciales, según la cual dos unidades espaciales se consideran como contiguas si poseen una frontera común de longitud no nula. Esta definición de contigüidad requiere obviamente de la existencia de un mapa, a partir del cual se puedan obtener las fronteras entre unidades espaciales. De acuerdo con esta definición, podrían considerarse varios órdenes de contigüidad de forma recursiva, tal como se expone en el **mapa de contigüidades** de la Figura III.1.6.

Figura III.1.6. Representación de las relaciones de contigüidad de orden 1, 2 y 4 para el condado de Champaign, en el medio oeste americano



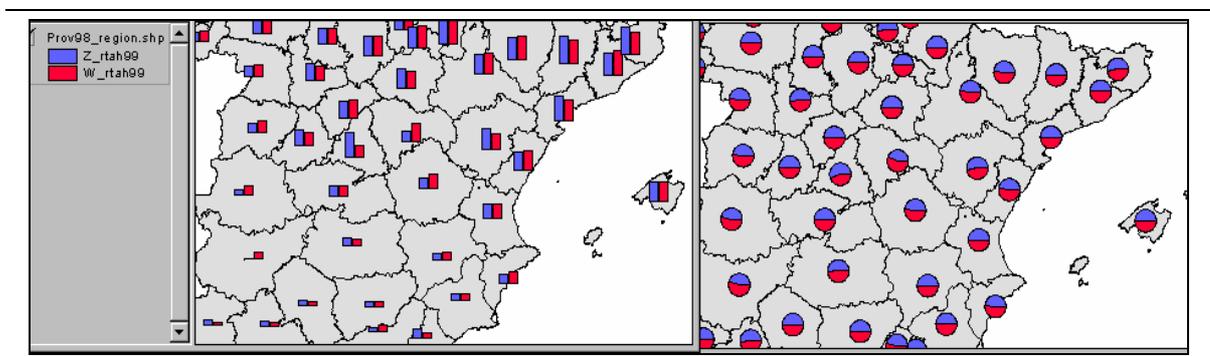
Fuente: Elaboración propia a partir del programa cvd 1.1.

Así, por ejemplo, en la Figura III.1.6, se representan varios órdenes de contigüidad para el condado americano de Champaign (señalado en naranja): los condados contiguos de primer orden al considerado (en color aguamarina), serían los que comparten con él una frontera común (McLean, Ford, Vermilion, Edgar, Douglas y Piatt), mientras que los 14 condados contiguos de segundo orden (Woodford, Livingstone, Kankakee, Iroquois, Benton, Warren, Vigo, Clark, Coles, Moultrie, Macon, De Witt, Logan y Tazewell), son los que, manteniendo una frontera común con los seis anteriores, no son contiguos de primer orden con Champaign.

- En el **gráfico del retardo espacial** el valor que adopta una variable en una determinada región se representa mediante un diagrama de barras o sectores,

frente a su retardo espacial⁸, de forma que el predominio de barras de igual altura, dos a dos, es signo claro de fuerte autocorrelación espacial y, por el contrario, alturas desiguales en las barras resultan indicativas de atípicos espaciales. En la Fig. III.1.7, se han representado dos gráficos del retardo espacial de la renta disponible per cápita de una selección de provincias españolas, advirtiéndose bastante equilibrio entre la altura de las barras dentro de cada par de variables, excepto en el caso de las provincias de Madrid y Guadalajara que constituyen atípicos espaciales.

Figura III.1.7. Gráfico del retardo espacial de la renta per capita provincial para 25 provincias españolas



Fuente: Elaboración propia a partir de la extensión de SpaceStat para ArcView.

- En el **diagrama de dispersión de Moran**, se estandariza la variable a analizar y se obtiene el retardo espacial de dicha variable estandarizada, representándose ambos valores en un eje cartesiano. La pendiente de la recta de regresión es el valor del estadístico I de Moran de autocorrelación espacial global⁹, de forma que cuanto mayor sea el ángulo que forme ésta con el eje de abscisas, más fuerte será el grado de autocorrelación espacial, y viceversa.

Este diagrama de dispersión suele dividir el tipo de asociación espacial en cuatro categorías: dos para autocorrelación espacial positiva (valores altos de una variable rodeados de valores altos o valores bajos rodeados de valores bajos) y dos para autocorrelación espacial negativa (valores altos rodeados por valores bajos, y viceversa). Las categorías de asociación espacial positiva se

⁸ Las herramientas del AEDE reticular (“lattice”) se basan en la representación gráfica de la asociación existente entre variables y retardos espaciales, entendiéndose por retardo espacial el promedio ponderado de los valores que adopta una variable en el subconjunto de observaciones vecinas a una dada. Por ejemplo, el retardo espacial de la variable renta per cápita de la provincia de Madrid podría obtenerse como una media aritmética simple de los valores de renta per cápita en las provincias limítrofes (Segovia, Ávila, Toledo, Cuenca y Guadalajara).

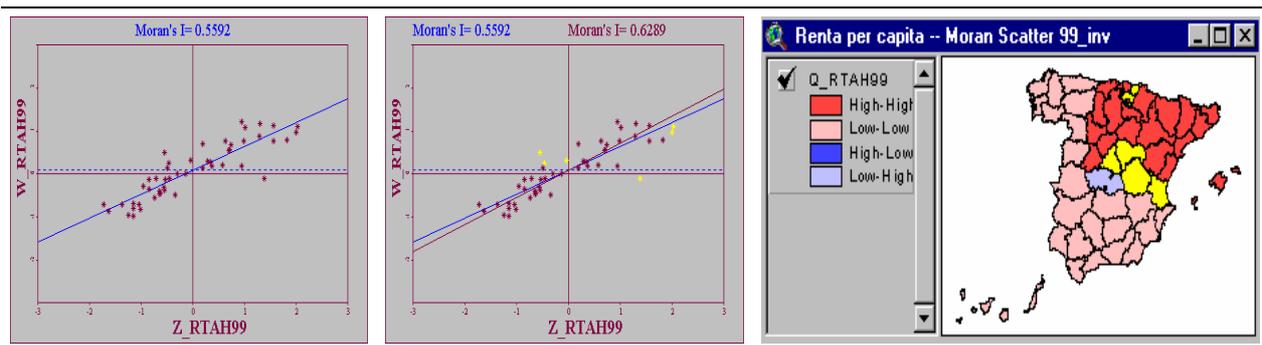
⁹ Las medidas estadísticas de autocorrelación espacial que se mencionan en este Apartado, se presentarán con mayor detalle en el Apartado III.2.2.

corresponden con los cuadrantes I y III (ver gráfico izquierdo de la Fig. III.1.8), en los que se presentan aquellas provincias que, con una renta disponible por habitante superior/inferior a la media nacional (valor estandarizado positivo/negativo de la variable) se encuentran rodeadas de provincias que también disponen de rentas superiores/inferiores a la media nacional (retardo espacial positivo/negativo), respectivamente. Por el contrario, las categorías de asociación negativa vienen dadas por los cuadrantes II y IV de este diagrama, en los que se representan las provincias con valores bajos/altos de renta per cápita rodeadas por provincias con valores altos/bajos de dicha variable, respectivamente.

Estas cuatro categorías dan lugar a una cierta forma de alisado espacial que puede visualizarse fácilmente en un **mapa** (Anselin y Bao, 1997), como el de la Fig. III.1.8.

Debe advertirse, además, que la regresión del retardo espacial sobre la variable correspondiente se realiza con todas las hipótesis clásicas del análisis de regresión. Así, la interpretación del test I de Moran claramente permite conocer el grado en que este estadístico resume la estructura global de asociación lineal existente en un fenómeno espacial que, en el caso que se expone en la Fig. III.1.8, para el total de la distribución, sería del 56% (valor del estadístico I de Moran en términos porcentuales), aunque si se excluyen los casos atípicos situados en el primero, segundo y cuarto cuadrantes (puntos destacados en color amarillo), se eleva al 63%.

Figura III.1.8. Diagrama y mapa de dispersión de Moran de la renta disponible per capita provincial



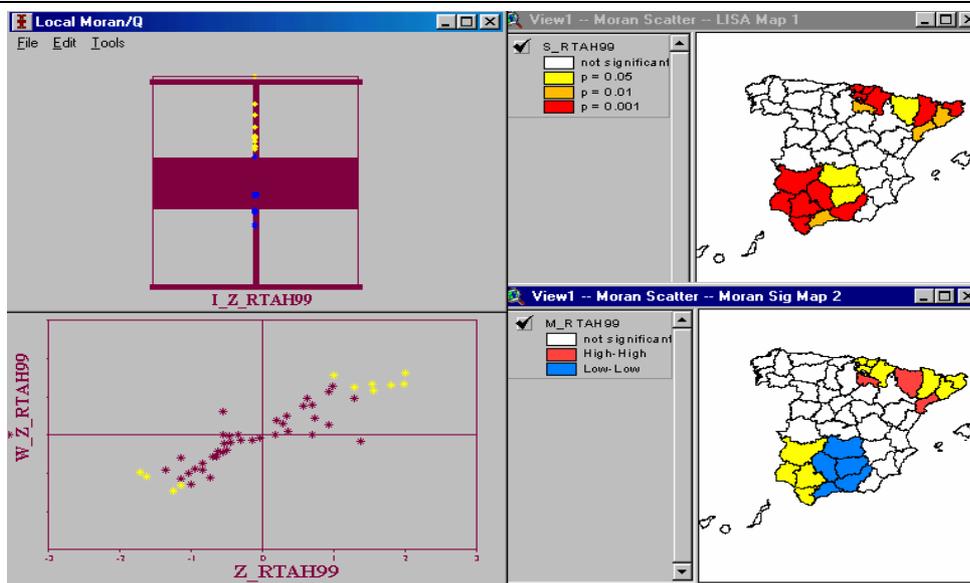
Fuente: Elaboración propia a partir de las extensiones de SpaceStat y DynESDA para ArcView.

- Dado que los valores de la variable se encuentran estandarizados en el diagrama de dispersión de Moran, es posible conseguir información tanto de la asociación espacial global (pendiente de la recta de regresión) como de la local. De este modo, los valores en el diagrama de dispersión de Moran situados a más de dos unidades del origen, según el llamado criterio “2-sigma”, pueden considerarse

como **puntos atípicos en el diagrama de dispersión de Moran**, es decir, importantes “agujeros” locales de no estacionariedad espacial que, en el caso de la Figura III.1.8, se trataría de las provincias de Guipúzcoa y Álava (en color amarillo y en el primer cuadrante). Se da además la circunstancia de que algunas provincias se localizan en los cuadrantes de asociación espacial negativa, como Cuenca, Guadalajara y Valencia/València (segundo cuadrante), que poseen un valor de renta per cápita inferior a la media pero están rodeadas de provincias con valores superiores (“Low-high”) y, en el cuarto cuadrante, la provincia de Madrid, que presentando un valor alto de renta per cápita, se encuentra rodeada de valores de dicha variable ligeramente inferiores a la media nacional (“High-low”).

- En cuanto a los **mapas LISA** (“Local Indicator of Spatial Association”), se trata de mapas en los que se representan aquellas localizaciones con valores significativos en indicadores estadísticos de asociación espacial local (Getis y Ord, 1992; Anselin, 1995B; Ord y Getis, 1995; Unwin, 1996), poniendo así de manifiesto la presencia de puntos calientes (“hot spots”) o atípicos espaciales, cuya mayor o menor intensidad dependerá de la significatividad asociada de los citados estadísticos.

Figura III.1.9. Mapas LISA de la renta disponible per capita provincial (derecha) y diagrama de dispersión multivariante de Moran (izquierda)



Fuente: Elaboración propia a partir de las extensiones de SpaceStat y DynESDA para ArcView.

En la Fig. III.1.9, se han representado dos mapas LISA: en el superior, a través de una gradación de colores, se representan los distintos niveles de significación del estadístico de asociación espacial I local de Moran para las provincias con

alta concentración de valores de renta per cápita alta rodeadas de vecinas también con renta alta (en el diagrama de dispersión de Moran, zona “High-high”), y viceversa para valores bajos de renta per cápita (“Low-low”). En el mapa LISA inferior de la Figura, se representan las zonas calientes de alto valor (color rojo) y bajo valor (azul) de la distribución de la renta disponible per cápita provincial. La zona caliente de alta renta disponible per cápita está formada por las provincias vascas, catalanas, Navarra, La Rioja y Huesca, mientras que la zona caliente de bajo valor de renta se localiza en el suroeste peninsular y comprende las provincias andaluzas (excepto Almería), Badajoz y Ciudad Real.

- En la Figura III.1.9, también se ha representado el **diagrama de caja LISA** para la distribución del estadístico I local de Moran de asociación espacial. En un entorno de AEDE dinámico, se han destacado (en color amarillo) las provincias con un nivel alto en este estadístico (superior al segundo cuartil), que también aparecen destacadas en el diagrama de dispersión de Moran inferior. Es decir, se trataría de aquellas provincias en las que se produce una especial concentración de valores extremos (altos rodeados de altos o bajos rodeados de bajos) de renta per cápita.

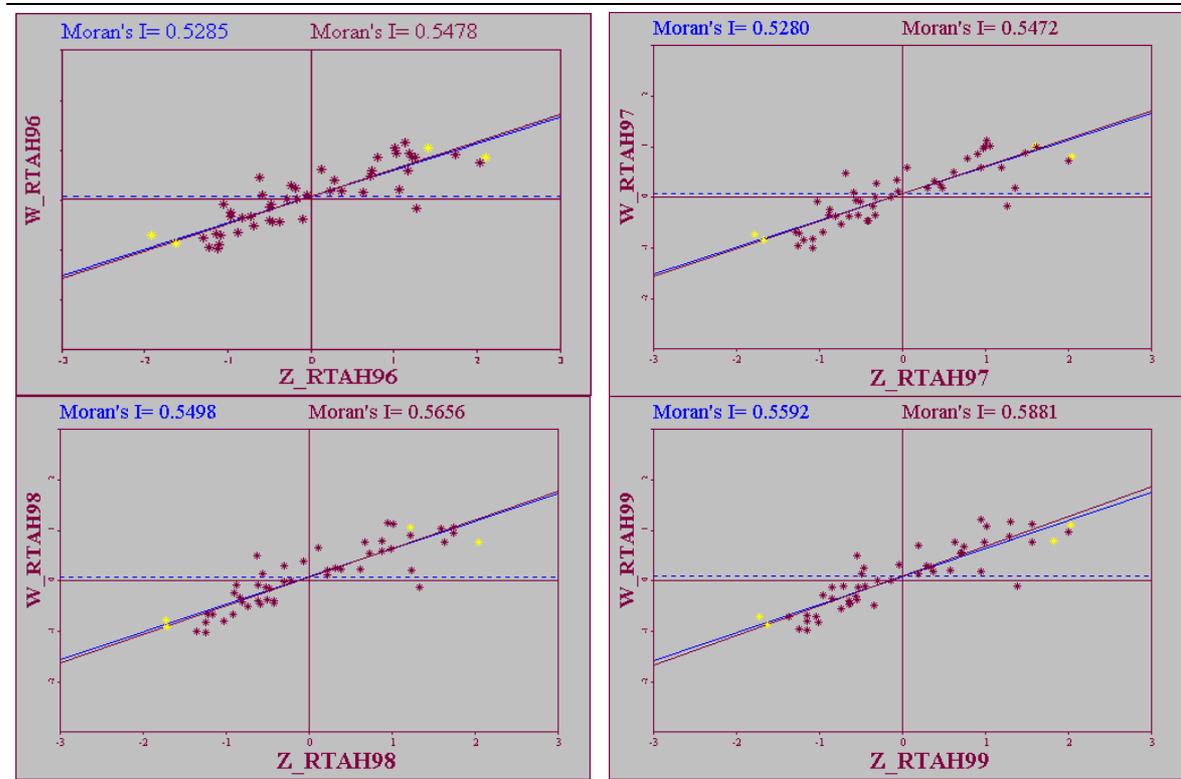
Así, en la zona caliente, el centro de la misma se localizaría en las provincias más septentrionales (excepto Huesca), mientras que el “núcleo duro” de la zona fría vendría definido por las provincias suroccidentales de Cádiz, Huelva, Sevilla y Badajoz.

- Por último, el **diagrama de dispersión multivariante de Moran** es una técnica de exploración de asociación espacial multivariante derivada del clásico estadístico I de Moran de asociación espacial. Este diagrama multivariante, implantado en un entorno dinámico, permite comparar el comportamiento del fenómeno de asociación espacial en varios indicadores.

Como puede apreciarse en la Figura III.1.10, durante el período 1996-99, la variable renta disponible por habitante de las provincias españolas mantiene un nivel alto de autocorrelación espacial positiva, siendo máximo en el último año considerado.

Además, este método permite analizar el comportamiento de determinadas unidades geográficas que, por ejemplo, se presentan con valores atípicos o extremos. Este último sería el caso de las provincias de Guipúzcoa y Álava, de altos niveles de renta per capita (destacadas en amarillo en el cuadrante primero), que han ido avanzando posiciones en el ranking de renta per capita en los cuatro años considerados, mientras que, por el contrario (ver cuadrante tres), las provincias de Cádiz y Badajoz se han mantenido siempre, a lo largo de este período, en los últimos puestos.

Figura III.1.10. Diagrama de dispersión multivariante de Moran de las variables de renta disponible per capita provincial para los años 1996-1999.



Fuente: Elaboración propia a partir de las extensiones de SpaceStat y DynESDA para ArcView.

La existencia de asociación espacial (global y/o local), puesta de manifiesto por las técnicas del AEDE en una variable que va a ser explicada en un modelo de regresión (así como en alguna de las variables explicativas o en el propio residuo), pone de manifiesto la necesidad de contrastar estadísticamente este fenómeno, a través de las diversas medidas estadísticas de autocorrelación espacial que se han propuesto en la literatura. De confirmarse esta hipótesis, el fenómeno de asociación espacial deberá ser considerado explícitamente a través de modelos de regresión específicos, como el modelo del retardo espacial (“spatial lag model”) y el modelo del error espacial (“spatial error model”), tal como se expondrá en el Capítulo IV con mayor detalle.

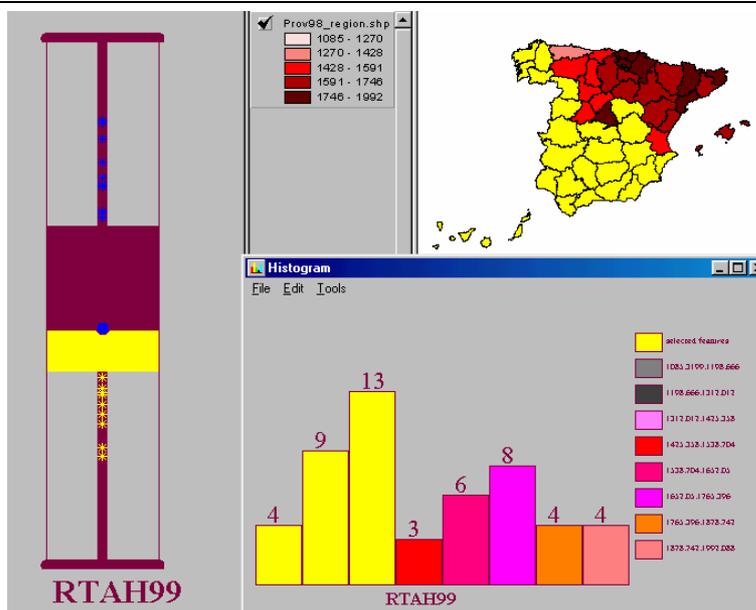
III.1.3.3. Representación del fenómeno de heterogeneidad espacial

La heterogeneidad espacial es, junto con el fenómeno de asociación espacial, uno de los efectos espaciales, definido por la ausencia de estabilidad en el espacio de la variable en estudio, lo que suele ocurrir con situaciones del tipo centro-periferia, norte-sur, este-oeste, etc. Como ya se puso de manifiesto en el capítulo anterior, esto implicará que, en los modelos espaciales, las formas funcionales y los parámetros variarán con la localización geográfica, no siendo homogéneos para toda la matriz de datos. En este caso, aunque no

suela ser un tema muy citado en la literatura, la utilización del AEDE dinámico resulta también particularmente útil para su exploración. Anselin (1999B) sugiere la utilización conjunta del mapa del histograma de frecuencias y el diagrama de dispersión, en el modo que se presenta a continuación.

- El **mapa del histograma de frecuencias** consiste en la selección geográfica de las unidades situadas en las “barras” o intervalos de un histograma que, a su vez, puede ser complementado con un diagrama de caja, como en el caso de la Figura III.1.11. La interacción dinámica existente entre el histograma y el mapa permitirán seleccionar el número más adecuado de intervalos de cara a una mejor identificación de diferentes estructuras o zonas de comportamiento en una variable. En concreto, en esta Figura se ha dividido la distribución de la renta disponible por habitante en 8 intervalos, de forma que la selección de los tres primeros intervalos del histograma de frecuencias de la variable de renta per cápita de 1999 (con valores inferiores a la mediana, como se pone de manifiesto en el diagrama de caja), destaca en color amarillo las provincias localizadas en la mitad oeste-centro-sur de España, pudiéndose trazar una línea inclinada de discontinuidad geográfica en la distribución de esta variable.

Figura III.1.11. Mapa del histograma de la variable renta disponible por habitante de las provincias españolas en 1999.

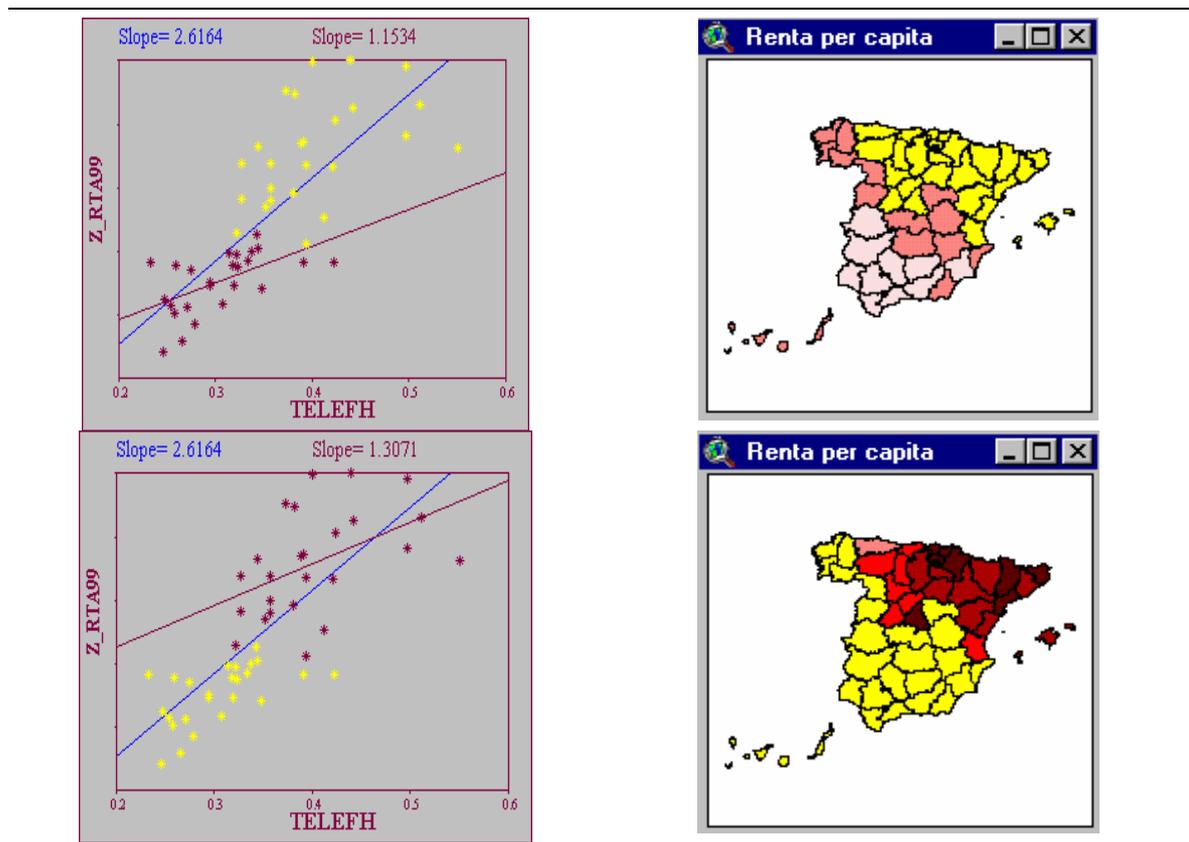


Fuente: Elaboración propia a partir de las extensiones de SpaceStat y DynESDA para ArcView.

- Además, Anselin (1999B) propone comprobar la existencia de relaciones o coeficientes diferentes entre la variable a explicar en un modelo y sus explicativas, en las dos estructuras detectadas, lo que podrá llevarse a cabo

mediante un **diagrama de dispersión**. Efectivamente, como se observa en la Figura III.1.12, el valor del coeficiente estimado para la variable explicativa de un modelo de regresión lineal en las estructuras espaciales anteriormente detectadas no varía significativamente: la regresión de la variable de líneas telefónicas por habitante sobre la renta disponible por habitante, que estima un coeficiente de 2,62 para el conjunto global de provincias españolas, obtiene un coeficiente de 1,15 para la muestra de datos correspondiente a la mitad occidental-centro-sur de España (valores bajos de renta per capita), siendo el coeficiente correspondiente a la más rica mitad nororiental de 1,31. Estos resultados estarían poniendo de manifiesto una mayor elasticidad de la variable de teléfonos por habitante sobre la renta per cápita en la mitad nororiental que en la occidental-centro-sur.

Figura III.1.12. Diagrama de dispersión de la variable teléfonos por habitante vs renta disponible por habitante de las provincias españolas en 1999.



Fuente: Elaboración propia a partir de las extensiones de SpaceStat y DynESDA para ArcView.

Cuando el AEDE pone de manifiesto la posible hipótesis de heterogeneidad espacial en un modelo, deberían aplicarse los contrastes estadísticos propuestos en la literatura que,

de confirmar esta hipótesis, conducirán a la utilización de modelos específicos que contemplen adecuadamente este problema, como los modelos de análisis de la varianza espacial (SANOVA) y de estructuras espaciales, para situaciones de heterogeneidad espacial discreta, o los modelos de superficie tendencial y expansión espacial, en los casos de heterogeneidad espacial continua.

III.2. DEPENDENCIA O AUTOCORRELACIÓN ESPACIAL

III.2.1. Concepto

El fenómeno de dependencia o autocorrelación espacial¹⁰ ha sido ya introducido, de forma más o menos general, en los Apartados II.1.1 (concepto de econometría espacial) y III.1.3.2 (representación del fenómeno de dependencia espacial) y definido como la relación funcional existente entre los valores que adopta un indicador en una zona del espacio y en zonas vecinas. Recientemente, Anselin (2001D) también define este fenómeno como “coincidencia entre valores similares (de una variable) y localizaciones cercanas”. Aunque la dependencia espacial puede parecer similar a la dependencia en las series temporales, ya se ha indicado que este parecido sólo es real en parte, debido a la naturaleza multidireccional de la dependencia en el espacio frente a la clara situación unidireccional del tiempo¹¹.

La dependencia o autocorrelación espacial ha sido un tema mucho más estudiado que el otro efecto de heterogeneidad espacial, desde los primeros trabajos estadísticos de Cliff y Ord (1973), aunque ya otros anteriormente habían constatado la existencia de continuidad geográfica en muchos fenómenos, como las intenciones de voto de los norteamericanos (Cox, 1969). Por su parte, Cliff y Ord (1981) presentan el análisis de la dependencia espacial como una técnica eficaz para el conocimiento de las causas y formas de propagación de epidemias y enfermedades. Además, este efecto espacial ha sido objeto de aplicaciones interesantes en otros campos, como los recursos naturales y medio ambiente (Anselin, 2001), delincuencia (Baller et al., 2001; Ceccato et al., 2001), actividad de I+D (Fischer y Varga, 2001; Anselin *et al.*, 2000), etc.

En el terreno específicamente económico, se ha demostrado como muy común la existencia de autocorrelación espacial en fenómenos como la producción, paro, renta disponible, procesos de convergencia regional, demanda de transporte, impuestos, precio de la vivienda, etc., tal como han puesto de manifiesto múltiples autores, sobre todo en el ámbito internacional (por citar los más recientes, Rey, 2001A, B y C; Rey y Montouri, 1999; Sandberg y Johanson, 2001; Thomas, 2001; Van der Kruk, 2001; Haines *et al.*, 2000; Magalhães *et al.*, 2000), aunque también se han realizado estudios aplicados a provincias y regiones españolas, por ejemplo, en Alañón (2001), Toral (2001), Chasco y Vicéns, (2000), López y Palacios (2000), Moreno *et al.* (2000), Goicolea *et al.* (1998), Cano *et al.*, 1997, Vayá y Suriñach (1996) y Chica (1994), entre otros.

¹⁰ En un sentido estricto, los conceptos de dependencia y autocorrelación espacial no son sinónimos. En efecto, la autocorrelación espacial es una forma de dependencia espacial definida estadísticamente de forma más débil, a través sólo de los primeros momentos de la distribución conjunta de una variable espacial. Sin embargo, la mayoría de los autores utiliza ambos términos indistintamente (Anselin, 1999B; Moreno y Vayá, 2000), con algunas excepciones como en Brett y Pinkse (1997).

¹¹ Propio del contexto espacial es el problema, aún sin resolver, del efecto frontera según el cual la dependencia espacial no se limita a las regiones incluidas en la muestra analizada, sino que se extiende a unidades espaciales para las que no se dispone de información (Anselin, 1988A).

El fenómeno de autocorrelación espacial obliga a tener siempre en cuenta la configuración espacial de una muestra de valores. Así, Vayá (1998) llama la atención sobre ciertas medidas estadísticas como los índices de desigualdad de la renta (Gini, convergencia sigma, etc.), incapaces de detectar situaciones de re-localización de las regiones más dinámicas cuando la dispersión de la variable en estudio no varía. Efectivamente, estas medidas obtendrán resultados idénticos en muestras distintas en las que los valores de la variable en estudio sean iguales, pero localizados en puntos distintos del espacio.

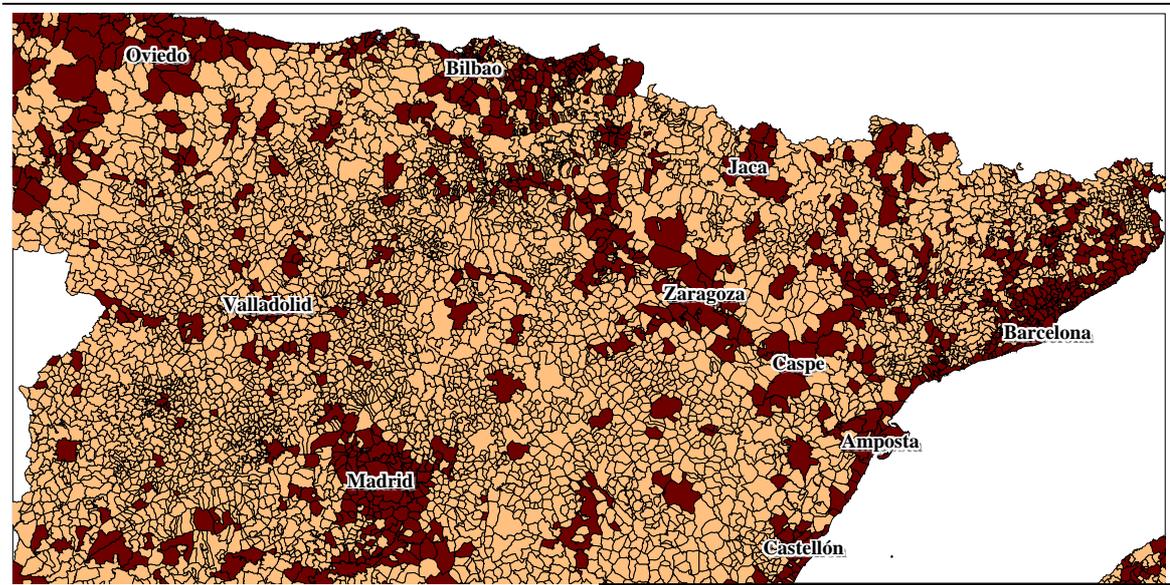
Parece claro que, en muchas ocasiones, resulta fundamental conocer la respuesta a la pregunta formulada por Cliff y Ord (1981): ¿Es de algún modo significativa la estructura espacial propia de una variable y, por tanto, digna de ser conocida e interpretada? Evidentemente, cuando una variable se distribuye de forma sistemática en un espacio geográfico se dice que el fenómeno en cuestión presenta autocorrelación espacial, para lo que se han desarrollado diversas medidas o contrastes estadísticos que permiten determinar si la estructura espacial definida por una variable es o no significativa, es decir, si merece o no la pena su conocimiento y estudio, como se verá en el próximo Apartado.

Tal como se observaba en el diagrama de dispersión de Moran (Apartado III.1.3), el efecto de autocorrelación espacial puede ser de signo positivo o negativo, así como nulo.

- Se entiende por **autocorrelación espacial positiva** el fenómeno de asociación entre valores similares de una variable y localizaciones cercanas; es decir, cuando, en el espacio geográfico, los valores altos de una variable están rodeados por valores altos y viceversa. Éste sería el caso, del llamado efecto contagio o desbordamiento (“spillover”) que se produce en muchos fenómenos socioeconómicos de renta y desarrollo humano, en general, en los que su presencia en una región es causa de su extensión a regiones vecinas, favoreciendo la concentración del fenómeno en la zona.
- Por el contrario, existe **autocorrelación espacial negativa** en un espacio cuando los valores altos de una variable se encuentran rodeados por valores bajos de la misma, y viceversa, como ocurre con la disposición de las casillas blancas y negras en un tablero de ajedrez. Esta configuración, en la que se produce una mayor disimilitud entre unidades geográficas cercanas que entre las lejanas es, por ejemplo, la que se produciría en fenómenos de jerarquías espaciales del tipo centro-periferia, como el representado en la Figura III.2.1, en la que los centros de mayor actividad económica están rodeados de regiones periféricas próximas con baja actividad.
- Por último, se produce **ausencia de autocorrelación espacial** en una variable geográfica cuando ésta se distribuye de manera aleatoria sobre el espacio.

Sin embargo, como se verá a continuación, no siempre que se produce un efecto de autocorrelación espacial en una variable geográfica será consecuencia de la existencia de una interacción espacial entre regiones, como los citados fenómenos de desbordamiento y jerarquías espaciales. Algunas veces, las variables espaciales pueden estar correlacionadas de forma espuria o falsa.

Figura III.2.1. Autocorrelación espacial negativa en la distribución del índice de actividad económica de los municipios del nordeste peninsular en 1999



Fuente: Elaboración propia a partir de MapInfo Professional.

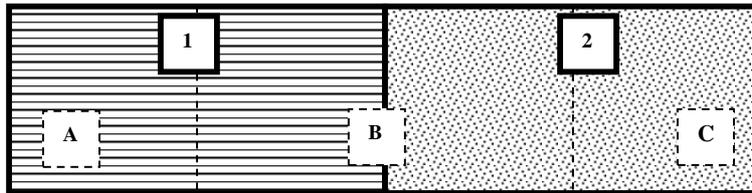
Efectivamente, según Anselin (1988A), la aparición de dependencia espacial en una variable puede ser debida a una falsa causa, como ciertos errores de medida, y a verdaderos fenómenos de interacción espacial.

- **Errores de medida procedentes de observaciones en regiones contiguas.** En muchas ocasiones, los datos son recogidos de la realidad en forma más o menos agregada, por lo que puede haber poca correspondencia entre el ámbito del fenómeno en estudio y el nivel de agregación de las unidades espaciales en observación, tal como se presentó en el Apartado II.2.4. Es probable que, por este motivo, se produzcan ciertos errores de medida. Es más, estos errores tienden a ir más allá de los límites de las propias unidades espaciales afectadas; es decir, que los errores cometidos en el dato correspondiente a la región *i*, probablemente se transferirán a los errores de la región vecina *j*. Este trasvase espacial de los errores de medida es una de las causas obvias de existencia de dependencia espacial.

En un contexto de análisis de regresión, esto puede fácilmente producir perturbaciones aleatorias no esféricas, así como problemas de error en las variables, tal como se ilustra en la situación hipotética de la Fig. III.2.2. En este ejemplo, se supone que la correcta escala o delimitación espacial de una variable *Y* se corresponde con las áreas señaladas en línea discontinua, A, B y

C, mientras que las observaciones disponibles de Y se han obtenido en forma agregada a dos niveles 1 y 2 (dibujados en trazo grueso).

Figura III.2.2. Dependencia espacial y agregación



Fuente: Anselin (1988A).

En este sistema, la variable observable en el nivel de agregación 1 (Y_1) será la agregación de Y_A y parte de Y_B , mientras que la variable observable en 2 (Y_2) será la agregación de Y_C y lo que resta de Y_B . Una consecuencia probable sería que la variable Y se encontrara correlacionada espacialmente de forma falsa o espuria, por haber sido medida en un nivel geográfico incorrecto.

Un ejemplo muy común en el ámbito del marketing geográfico, es el hecho de que muchas empresas utilizan información propia disponible para las zonas determinadas por los códigos postales, cuando es bien sabido que estas áreas no han sido determinadas con un criterio económico. Cuando se pretenden establecer áreas de influencia de un negocio a partir de los códigos postales, pueden producirse importantes errores de medida. Efectivamente, la distribución del público objetivo de muchos negocios suele estar más relacionada con las secciones censales, determinadas para una población no superior a 2.000 habitantes (equivalente, en los núcleos urbanos, a una o varias manzanas de casas), que con el territorio delimitado por los códigos postales que están establecidos, en muchos casos, con criterios más o menos desfasados (antiguas rutas de distribución postal).

- **Interacciones espaciales propias del comportamiento humano.** Ésta sería la verdadera causa de autocorrelación espacial, positiva o negativa, en los fenómenos propios de la ciencia regional y la geografía humana, en los que, como ya se ha indicado, la localización y la distancia son fundamentales a veces, porque dan lugar a una gran variedad de relaciones de interdependencia tanto en el espacio como en el tiempo. Las teorías de interacción espacial, los modelos de difusión y las jerarquías espaciales se han formulado para expresar matemáticamente estas relaciones de dependencia existentes entre diferentes puntos o regiones del espacio geográfico.

III.2.2. Formulación matemática del fenómeno de dependencia espacial

Las interacciones espaciales que suelen producirse en muchos fenómenos humanos son la causa verdadera del efecto de dependencia o autocorrelación espacial y podrían expresarse matemáticamente como una relación funcional entre los valores que adopta una variable Y en la localización i (y_i) y los valores de dicha variable en un conjunto n de localizaciones situadas en el territorio en estudio, del modo siguiente:

$$y_i = f(y_1, y_2, \dots, y_N) \quad \text{Eq. III.2.1.}$$

Esta expresión no resulta operativa en la práctica porque da lugar a un sistema no identificable, con muchos más parámetros a estimar, hasta un total de $(N^2 - N)$, que observaciones disponibles (N). Por este motivo, deben establecerse estructuras comunes, subyacentes al fenómeno de interacción que se estudia, de forma que sólo se estime un pequeño número de características propias de la dependencia espacial existente. En los fenómenos humanos, la influencia que ejercen ciertas unidades geográficas sobre una dada, suele expresarse matemáticamente a través de conceptos topológicos de vecindad, en los que juega un papel fundamental la distancia métrica (euclídea, bloque de Manhattan o general de Minkovski).

Obviamente, tal como corresponde a una perspectiva reticular o “lattice”, propia del análisis econométrico espacial (ver Apartado III.1.3), se supone que las observaciones se encuentran organizadas en unidades espaciales discretas, tanto puntos situados en una red regular (o irregular) como regiones de un mapa. El conjunto de unidades vecinas a una dada suele representarse gráficamente como una estructura gráfica reticular y, de forma matemática, como una matriz de interacciones.

III.2.2.1. Matriz de interacciones espaciales

Los contrastes de dependencia o autocorrelación espacial pueden basarse en una noción de contigüidad binaria entre las unidades espaciales. De acuerdo con este concepto, una situación de vecindad entre dos unidades espaciales se podría expresar mediante valores de tipo 0-1. Es decir, si dos unidades espaciales tienen una frontera común de longitud no nula, se considera que son contiguas y se les asigna el valor 1¹².

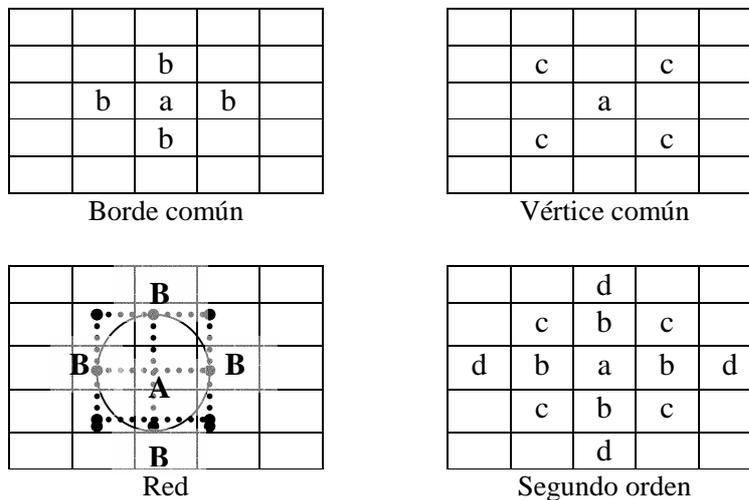
Esta definición de contigüidad requiere de la existencia de un mapa, a partir del cual se puedan obtener las fronteras entre unidades espaciales o de un “software” que realice, de forma automática, esta operación, como la extensión de SpaceStat para ArcView (ver Apartado III.1.2). Cuando dichas unidades se posicionan de forma irregular, puede ser fácil conocer las fronteras entre las distintas unidades geográficas. Sin embargo, cuando las

¹² En el caso de que las unidades espaciales sean puntos (por ejemplo, ciudades), la contigüidad puede consistir en la noción del camino más corto. Así, dos puntos serían contiguos si la distancia que los separa es inferior a un valor determinado (Moreno y Vayá, 2000).

unidades pertenecen a una cuadrícula regular la determinación de la contigüidad no es única.

Por ejemplo, consideremos la cuadrícula regular y los centroides de la Figura III.2.3. Hay varias formas de establecer una frontera común entre la celda “a” y las que le rodean: puede considerarse como frontera la existencia de un borde común, siendo entonces las celdas “b” las contiguas a “a” y, alternativamente, podría ser considerado como frontera la existencia de un vértice común, siendo en este caso las celdas “c” las contiguas a “a”. Además, podría tenerse también en cuenta una combinación de ambos conceptos. En analogía con el juego del ajedrez, estas situaciones han sido denominadas, respectivamente, el criterio de la torre, el criterio del alfil y el criterio de la reina.

Figura III.2.3. Contigüidad en un entramado regular



Fuente: Anselin (1988A).

Cuando las unidades espaciales son puntos regular o irregularmente distribuidos sobre el sistema, como el caso de las ciudades en una jerarquía urbana, el concepto de contigüidad se define sobre la noción de camino más corto en una red como la determinada por las líneas discontinuas de la Figura: los nudos de la red son considerados como vecinos si se encuentran dentro de una distancia máxima (camino más corto) unos de otros. Por ejemplo, en la Figura III.2.3, los nudos B se encuentran en un radio “d” del centroide A, y pueden ser considerados como contiguos para esta concreta distancia crítica.

De forma similar, pueden considerarse varios órdenes de contigüidad de forma recursiva, definiendo el orden de contigüidad k-ésimo cuando las unidades espaciales son de primer orden contiguas a una unidad espacial de orden k-ésimo, y no contiguas de orden inferior. Por ejemplo, en un sistema de red cuadrada, se trataría de series de franjas concéntricas alrededor de una unidad espacial considerada, como en la Figura III.2.3, en el

que las celdas “c” y “d” son contiguas de 2º orden con “a”, según el criterio de la torre, siendo también contiguas de primer orden con b.

La estructura espacial suele expresarse formalmente a través de una matriz de interacciones espaciales, también llamada “matriz de pesos, ponderaciones, distancias o contactos espaciales” (Aznar *et al.*, 1996; Vayá y Suriñach, 1996)¹³. En esta matriz, cada unidad espacial se representa a la vez mediante una fila y una columna. En cada fila, los elementos no nulos de las columnas se corresponden con las unidades espaciales contiguas. Por ejemplo, para las 9 celdas situadas en el centro de la Figura III.2.4, es posible construir una matriz (9x9) correspondiente a estos datos (cuyas celdas están numeradas de izquierda a derecha y de arriba abajo) utilizando el criterio de contigüidad de la torre.

Figura III.2.4. Matriz de interacciones para nueve datos situados en una cuadrícula

	1	2	3						
	4	5	6						
	7	8	9						

	1	2	3	4	5	6	7	8	9
1	0	1	0	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	0	1	0	0	0	1	0	0	0
4	1	0	0	0	1	0	1	0	0
5	0	1	0	1	0	1	0	1	0
6	0	0	1	0	1	0	0	0	1
7	0	0	0	1	0	0	0	1	0
8	0	0	0	0	1	0	1	0	1
9	0	0	0	0	0	1	0	1	0

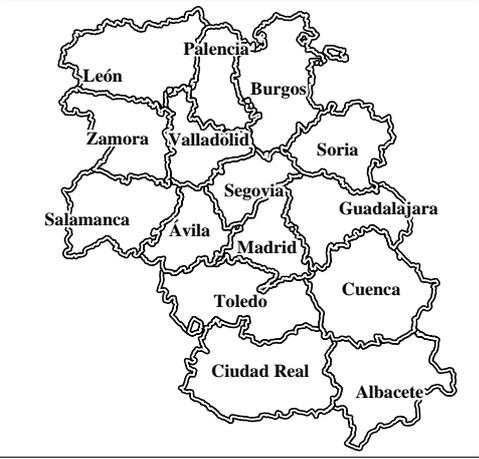
Fuente: Anselin (1988A) y elaboración propia.

En los primeros análisis de estadística espacial, esta matriz de interacciones espaciales (también denominada “matriz de booleana o de contigüidad”) recibía la notación $[\delta_{ij}]$, siendo $\delta_{ij} = 1$, cuando las regiones i, j se encuentren vinculadas, y $\delta_{ij} = 0$ en los demás casos, estableciéndose que $\delta_{ij} = \delta_{ji}$, para todo i, j y, por convenio, $\delta_{ii} = 0$. Es decir, la matriz de interacciones espaciales $[\delta_{ij}]$ es simétrica y, por tanto, incapaz de incorporar influencias no recíprocas (no siempre la influencia que j recibe de i es la misma que la que i recibe de j) y su diagonal principal está constituida por ceros. Además, considera como único determinante de las interdependencias regionales la adyacencia física, descuidando con ello posibles influencias mutuas entre regiones que, aun estando alejadas, mantienen, por ejemplo, estrechas relaciones comerciales. Pese a todo, esta matriz de interacciones espaciales es muy utilizada habitualmente por su simplicidad.

¹³ En términos generales, en esta obra se utilizarán los términos “matriz de interacciones espaciales o matriz booleana”, $[\delta_{ij}]$, para designar a la matriz más sencilla de términos binarios (0-1) que se presenta en este Apartado, y “matriz de ponderaciones o pesos espaciales”, $W = [w_{ij}]$, para aquellas matrices generalizadas de términos no necesariamente binarios, que se presentarán en el Apartado siguiente.

En la realidad regional, no suelen darse situaciones de cuadrícula regular como la que se acaba de presentar. Los países, regiones, municipios, etc. tienen formas irregulares y las relaciones de vecindad deben definirse de acuerdo con el fenómeno que se desea estudiar. A modo de ilustración, se presenta una posible matriz de interacciones para las 15 provincias pertenecientes a las autonomías de Castilla y León, Castilla-La Mancha y Madrid (Figura III.2.5). En este ejemplo, se considerará que dos provincias son contiguas de primer orden cuando tienen una frontera no nula en común, es decir, Palencia, Valladolid y Zamora se considerarán contiguas de primer orden a León, porque tienen con ella, en común, una frontera de 90, 50 y 120 km., respectivamente.

Figura III.2.5. Matriz de interacciones para las provincias castellanas y Madrid



	Alb	Ávi	Bur	CRe	Cue	Gua	Leó	Mad	Pal	Sal	Seg	Sor	Tol	Vall	Zam
Alb	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
Ávi	0	0	0	0	0	0	0	1	0	1	1	0	1	1	0
Bur	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0
CRe	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
Cue	1	0	0	1	0	1	0	1	0	0	0	0	1	0	0
Gua	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0
Leó	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
Mad	0	1	0	0	1	1	0	0	0	0	1	0	1	0	0
Pal	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
Sal	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1
Seg	0	1	1	0	0	1	0	1	0	0	0	1	0	1	0
Sor	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
Tol	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0
Vall	0	1	1	0	0	0	1	0	1	1	1	0	0	0	1
Zam	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0

Fuente: Elaboración propia.

En un intento de superar la excesiva simplificación de la matriz de interacciones, siguiendo el criterio de proximidad física, diversos autores han propuesto otras definiciones más complejas de esta matriz, como se verá a continuación.

III.2.2.2. Matriz de pesos espaciales

La matriz de pesos espaciales es una generalización de la matriz de interacciones o contigüidades, por lo que suele ser designada también con estos nombres, aunque se la conoce habitualmente como matriz de pesos, ponderaciones, retardos o contactos espaciales. Hay autores que también la denominan “matriz de ponderaciones Cliff-Ord”, en honor a quienes la formularon por primera vez (Anselin, 1988A)¹⁴. El uso de esta matriz de

¹⁴ Esta matriz constituye un caso particular del concepto más general de matriz de contigüidades (“adjacency matrix”) propia de la teoría de grafos, utilizada para describir las conexiones entre elementos en un gráfico o red (ver, por ejemplo, Garrison y Marble, 1963).

pesos espaciales, a diferencia de la matriz binaria de interacciones, permite al investigador la elección del conjunto de ponderaciones que él considere apropiado para cada fenómeno, lo que supone una mayor flexibilidad en la definición de la estructura de interdependencias de un sistema regional y permite considerar cuestiones como las barreras naturales o el tamaño de las regiones. Es más, cuando sea necesaria la consideración de hipótesis acerca del grado de vinculación existente entre áreas vecinas, deben utilizarse distintos conjuntos de ponderaciones que permitan contrastar dichas hipótesis.

Esta matriz de pesos generalizada (W), en lugar de utilizar las ponderaciones binarias δ_{ij} como cuantificación del concepto de vínculo, recoge el efecto de la región i sobre la región j a través de un peso o ponderación w_{ij} , de forma que $W \equiv [w_{ij}]$. La interpretación de estos elementos es la siguiente:

- $w_{ij} = 0$ sería indicativo de ausencia de autocorrelación espacial entre las observaciones i, j (por convenio, los elementos de la diagonal principal de la matriz de pesos serán, como en la matriz de interacciones, igual a cero).
- $w_{ij} \neq 0$ sería indicativo de existencia de una interacción espacial entre las observaciones i, j , que podría ser expresada como simple contiguidad binaria (teniendo una frontera común), como contiguidad de distancias, a través de centroides, dentro de una banda de distancia mínima (considerándose contiguos dos puntos cuando estén situados a una distancia inferior a la prefijada como mínima), o como función inversa de la distancia simple o cuadrática.

La especificación adecuada de los elementos de esta matriz, w_{ij} , es uno de los puntos metodológicos más difíciles y controvertidos en la econometría espacial. Efectivamente, será necesario tener cuidado a la hora de elegir los pesos espaciales para evitar posibles correlaciones espurias. Los factores más importantes a tener en cuenta para determinar las ponderaciones dependerán de cada estudio en particular.

Por ejemplo, el grado de vinculación existente entre dos regiones puede depender de la distancia entre sus centros geográficos o demográficos, o también puede estar relacionada con la longitud de la frontera común de dichas regiones, etc (Stetzer, 1982, Anselin y Rey, 1991, Florax y Rey, 1995). Por ejemplo, en áreas urbanas, la relación entre dos zonas podrá depender de la frecuencia de los servicios de transporte público.

Anselin (1995A) incluye en el paquete informático SpaceStat la posibilidad de obtener matrices sencillas de pesos espaciales, basadas en el modelo de gravitación, con la distancia entre regiones como única variable fricción, del modo siguiente:

$$w_{ij} = d_{ij}^{-2} \quad \text{Eq. III.2.2.}$$

donde w_{ij} : peso que recoge la influencia de la región j sobre la región i
 d_{ij} : distancia euclídea entre las regiones i, j

Este modelo, que otorga una menor influencia sobre una región dada a las regiones situadas a mayor distancia, constituye un caso particular del modelo de Cliff y Ord (1973). Estos autores habían propuesto anteriormente, a modo de ejemplo, que las variables más importantes para medir el grado de relación existente entre cualquier par de regiones eran la distancia entre los centros regionales y la longitud de la frontera común de ambas.

$$w_{ij} = d_{ij}^{-a} [\beta_{i(j)}]^b \quad \text{Eq. III.2.3.}$$

donde $\beta_{i(j)}$: proporción del perímetro de la región i que está en contacto con j
 a, b : parámetros con valor positivo, que conceden un mayor peso a aquellos pares de regiones cuyos centros se encuentran a menor distancia y tengan en común fronteras más extensas (tal como comenta Anselin, 1980); para evitar una mayor complejidad estos parámetros deberían ser dados a priori y no estimados conjuntamente).

Muchas veces se suele realizar la operación de estandarización o escalamiento de los pesos, de forma que cada elemento i de una fila j sea dividido por la suma de los pesos de dicha fila, de la siguiente manera:

$$\frac{w_{ij}}{\sum_{j \in J} w_{ij}} \quad \text{Eq. III.2.4.}$$

siendo J el grupo de regiones relacionadas con i .

En este tipo de matrices, la suma de los pesos de una fila es igual a la unidad y, como $i = 1, 2, \dots, n$, la suma de todos los pesos de la matriz es igual al tamaño de la muestra, n .

$$\sum_{j \in J} w_{ij} = w_{i \bullet} = 1 \quad \text{Eq. III.2.5.}$$

Evidentemente, tras la estandarización de W , la matriz resultante será asimétrica (siempre que $\sum_j w_{ij} \neq \sum_i w_{ij}$), complicando así los cálculos de algunos estadísticos y estimadores. Esta nueva matriz de ponderaciones espaciales es denominada como matriz estandarizada por filas. Algunos autores recomiendan su utilización (Anselin, 1995A), aunque no existe una razón contundente que lo justifique, excepto 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 (Moreno y Vayá, 2000).

Además, estas medidas también resultan menos eficaces cuando el fenómeno de interacción espacial en consideración viene determinado por variables, como las puramente económicas, que tienen poco que ver con la configuración espacial de las fronteras en un mapa físico. Por este motivo, algunos autores han sugerido la utilización de ponderaciones más directamente relacionadas con el fenómeno particular que se está analizando, introduciendo en las formulaciones variables de accesibilidad (relativas, por ejemplo, a

medios de comunicación entre regiones) o conceptos de la teoría de redes sociales (Bodson y Peeters, 1975; Case *et al.*, 1993; Goicolea *et al.*, 1998; Vayá *et al.*, 1998A y B; López-Bazo *et al.*, 1999A, Toral, 2000A y B). A continuación, se presenta un resumen con algunas especificaciones para la matriz de pesos espaciales sugeridas en la literatura.

Tabla III.2.1. Diversas especificaciones de la matriz de pesos espaciales

Referencia	Modelo	Descripción
Dacey (1968)	$w_{ij} = d_{ij} \cdot \alpha_i \cdot \beta_{i(j)}$. d_{ij} : distancia entre los puntos o regiones (i, j) . α_i : proporción de i sobre el área total de regiones. . $\beta_{i(j)}$: proporción del perímetro de i en contacto con j
Cliff y Ord (1973)	$w_{ij} = d_{ij}^{-a} [\beta_{i(j)}]^b$. a, b: parámetros positivos
Bodson y Peeters (1975)	$w_{ij} = \sum_{n=1}^N K_n \left\{ \frac{a}{1 + b \cdot e^{-c_j d_{ij}}} \right\}$. K_n : importancia del medio de comunicación n . N: total de medios de comunicación considerados . a, b, c _j : parámetros a estimar.
Anselin (1980)	$w_{ijt} = d_{ij}^{-2}$	
Cliff y Ord (1981)	$w_{ij} = (c + d_{ij})^{-a}$. c: término constante positivo
Case <i>et al.</i> (1993)	$w_{ij} = \frac{1}{ x_i - x_j }$. x: variable socioeconómica (ej., PIB per cápita).
Molho (1995)	$w_{ij} = \frac{E_j^{-ad_{ij}}}{\sum_{k \neq i} E_k^{-ad_{ik}}}$; $\forall i \neq j$. E: volumen de empleo
Ma <i>et al.</i> (1997)	$w_{ij} = e^{-d_{ij}^a}$ $w_{ij} = (l_{ij}/l_i)^a$ $w_{ij} = \frac{(l_{ij}/l_i)^a}{d_{ij}^{-b}}$. l_{ij} : longitud de frontera entre las regiones (i,j) . l_i : perímetro de la región i
Toral (2000A,B)	$w_{ij} = \delta_{ij} \frac{k_i k_j p_i p_j}{d_{ij}^a}$. $\delta_{ij} = 1$, si las unidades espaciales i, j tienen una frontera en común y cero, si no la tienen. . p: población; k: longitud (km) de carreteras . d_{ij} : distancia por carretera entre las capitales de i, j . a: parámetro positivo, con valores 0, 1 ó 2.
Van der Kruk (2001)	$W = \sum_{d=1}^D W_d$. d: orden de vecindad . D: número máximo de órdenes de vecindad existentes

Fuente: Elaboración propia.

III.2.2.3. Operador retardo espacial

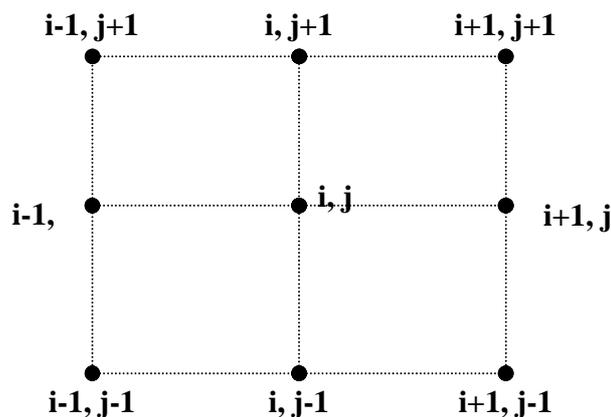
La matriz de ponderaciones espaciales hace posible la conexión entre el valor de una variable en un punto del espacio geográfico y las observaciones de dicha variable en otros puntos del sistema. En el contexto de series temporales, esto sería posible mediante un operador retardo que desfasara la variable uno o más períodos en el tiempo (Box y Jenkins, 1976). Por ejemplo, la expresión:

$$y_{t-s} = B^s y_t \quad \text{Eq. III.2.6.}$$

presenta la variable y desplazada s períodos en el pasado (a partir de t) en términos del operador retardo B aplicado a dicha variable y .

En el espacio, las cosas no son tan sencillas, dadas las muchas direcciones en que el desfase puede tener lugar. A modo de ilustración, se presenta la estructura reticular regular de la Figura III.2.6, en la que se obtendrán los retardos espaciales de la variable Y , observada en la localización (i,j) .

Figura III.2.6. Retardos espaciales en una red regular



Fuente: Anselin (1988A).

Siguiendo el criterio de simple contigüidad binaria, el valor de Y en (i,j) , y_{ij} podría ser desplazado o retardado de varias formas:

- Utilizando el criterio de contigüidad de la torre, los desplazamientos espaciales (“spatial shifts”) serían:

$$y_{i-1,j} ; y_{i,j-1} ; y_{i+1,j} ; y_{i,j+1} \quad \text{Eq. III.2.7.}$$

- Utilizando el criterio de contigüidad del alfil, los desplazamientos serían:

$$y_{i-1,j-1} ; y_{i+1,j-1} ; y_{i+1,j+1} ; y_{i-1,j+1} \quad \text{Eq. III.2.8.}$$

- Con el criterio de la reina, el número de posibles localizaciones, consideradas como desplazamientos o “shifts” aumentaría hasta el total de 8.

En la mayoría de las situaciones reales, en las que la localización espacial de las observaciones no suele presentarse en una retícula o malla regular (“regular lattice”), sino que suelen estar irregularmente repartidas en un mapa, no se pueden utilizar los criterios formales de desplazamiento espacial de las *Eq. III.2.7* y *III.2.8*: el número de localizaciones vecinas a una dada, sería diferente para cada caso, multiplicándose hasta el infinito el número de desplazamientos direccionales posibles en un sistema¹⁵.

Este problema se resuelve con el concepto de retardo espacial que consiste en un promedio ponderado de los valores de una variable en las localizaciones vecinas, con unos pesos o ponderaciones fijas y dadas de forma exógena. Es decir, el retardo espacial considera la suma ponderada de todos los valores de una variable que pertenecen a un mismo criterio de contigüidad (torre/alfil/reina), en vez de tomar individualmente cada uno de ellos. Los términos de esta suma se obtienen multiplicando las observaciones en cuestión (y_i) por sus correspondientes pesos de la matriz de ponderaciones W del modo siguiente:

$$B^s y_i = \sum_j w_{ij} y_j \quad ; \quad \forall j \in J_i \quad \text{Eq. III.2.9.}$$

donde B^s : operador retardo asociado con el criterio de contigüidad s
 j : subíndice correspondiente al conjunto J_i de unidades relacionadas con i , según el criterio de contigüidad s
 w_{ij} : ponderaciones espaciales.

De esta forma, cada elemento del retardo espacial es igual a un promedio ponderado de los valores de la variable Y en el subgrupo de observaciones vecinas a ella, J_i , dado que, $w_{ij} = 0, \forall j \notin J_i$. En el caso de estandarizar por filas la matriz de pesos espaciales, la variable retardada espacialmente representaría un suavizado de los valores vecinos, dado que la suma de todos los pesos de una determinada fila, en una matriz de este tipo, debe ser igual a la unidad.

De forma similar a lo que sucede con los retardos temporales en el análisis de series temporales, el concepto de retardo espacial puede ser extendido a órdenes superiores aunque, en el contexto espacial, los órdenes superiores están referidos a diferentes clases de contigüidad más que a un retardo mayor. Además, este concepto sólo tiene un significado preciso para matrices de interacciones binarias. La contigüidad de orden superior suele

¹⁵ Este problema se agudiza cuando en el contexto de los modelos de regresión espacial, se desea utilizar el retardo espacial como variable explicativa. En estos casos, el número de parámetros asociados a todas las direcciones posibles de desplazamiento en seguida se vuelve intratable, siendo imposible cualquier análisis significativo porque, a menos que el conjunto de datos sea muy grande y esté estructurado de forma regular, los grados de libertad resultantes serían insuficientes para permitir una estimación eficiente de dichos parámetros.

utilizarse en la elaboración de correlogramas espaciales, necesarios para la aplicación de los tests de autocorrelación espacial que se exponen en el siguiente Apartado.

La contigüidad de orden superior entre unidades espaciales es definida en un estilo recursivo como una contigüidad de primer orden con unidades que son contiguas a las siguientes de orden inferior. Es decir, una unidad espacial será contigua de 2º orden a una observación dada si es contigua de primer orden a otra unidad que es, a su vez, contigua de primer orden a dicha observación. Blommestein (1985) puso de manifiesto que este procedimiento puede dar lugar fácilmente a rutas circulares o trayectorias redundantes, de forma que las potencias de la matriz de interacciones incluyan trayectorias que ya están parcialmente contenidas en una matriz de interacciones de orden inferior. Dado que esta redundancia tendría implicaciones importantes en los procesos de estimación e inferencia de modelos espaciales, deben eliminarse previamente las trayectorias circulares presentes en las matrices de pesos antes de que sean utilizadas para la construcción de variables espacialmente retardadas.

III.2.3. Perspectivas en el análisis de autocorrelación espacial univariante: global y local¹⁶

En la taxonomía de técnicas del AEDE propuesta por Anselin (1998), se ponía de manifiesto la existencia de una doble perspectiva en el análisis del fenómeno de asociación o dependencia espacial, lo que, también se ha denominado, en Wise *et al.* (1999), como alisado o “smooth” (perspectiva global) y asperezas o “rough” (perspectiva local). La perspectiva global del fenómeno de autocorrelación espacial tiene por objeto el contraste de la presencia de tendencias o estructuras espaciales generales en la distribución de una variable sobre un espacio geográfico completo, mientras que el fenómeno de dependencia local vendría definido por una concentración, en un lugar del espacio global analizado, de valores especialmente altos o bajos (“puntos calientes/fríos”, “picos”, “bolsas” de valores atípicos) de una variable en comparación con el valor medio de la misma.

Aunque la mera observación en un mapa de la distribución de una variable espacial permite captar de forma intuitiva la existencia o no de patrones de comportamiento en dicha variable, esta información será siempre subjetiva y altamente dependiente de, por ejemplo, el número de intervalos establecido para la representación de dicha variable en el mapa. Por eso, resulta fundamental contar con una combinación de medidas o instrumentos gráficos y estadísticos, capaces de detectar la presencia significativa de autocorrelación (global y/o local) en una variable espacialmente distribuida. En esta línea, los contrastes de autocorrelación espacial, en combinación con las herramientas del AEDE¹⁷, ofrecen criterios estadísticos objetivos que permiten confirmar o rechazar la presencia de tendencias o estructuras espaciales en la distribución de una variable.

¹⁶ Se entiende como análisis de autocorrelación espacial “univariante” el contraste de este fenómeno en una variable propiamente dicha. En el Capítulo IV, se presentarán nuevos contrastes de autocorrelación espacial aplicables en un contexto de análisis de regresión espacial.

¹⁷ En el Apartado III.1.3.2, se incluyen las principales técnicas del AEDE reticular o “lattice” aplicadas al análisis y contrastación del fenómeno de dependencia espacial global y local.

Los estadísticos de autocorrelación espacial global son las primeras formulaciones propuestas en la literatura como medida estadística del efecto de autocorrelación espacial. Estos contrastes tienen la capacidad de resumir el esquema general de dependencia presente en una variable espacial en un único indicador (Moreno y Vayá, 2000). Recientemente, se han propuesto también los tests de autocorrelación espacial local, muy útiles para la detección de agrupaciones (“clusters”) de este fenómeno en el espacio geográfico general. En la Tabla III.2.2., se incluyen los estadísticos de autocorrelación, global y local, más utilizados y mencionados en la literatura.

De todos ellos, los tests I de Moran (1948) y c de Geary (1954) son probablemente los contrastes de autocorrelación espacial más conocidos. En el trabajo con datos ordinales y de intervalo¹⁸, ambos contrastes son de uso generalizado, aunque también puede añadirse el estadístico $G(d)$, que fue propuesto por Getis y Ord (1992) junto con otros indicadores de autocorrelación local, que han sido denominados conjuntamente como “familia de estadísticos G de Getis y Ord”. Por su parte, los tests de autocorrelación espacial local permiten contrastar la presencia de subzonas de dependencia espacial dentro de un espacio general dado, que Anselin (1995B) denomina Indicadores LISA (Indicadores Locales de Asociación Espacial), para los que establece una serie de buenas propiedades estadísticas.

Tabla III.2.2. Principales contrastes de autocorrelación espacial univariante

Autocorrelación global	Autocorrelación local
<i>I de Moran (1948)</i>	<i>I_i de Moran (Anselin, 1995B)</i>
<i>c de Geary (1954)</i>	<i>c_i de Geary (Anselin, 1995B)</i>
<i>Γ de Mantel (1967)</i>	<i>Γ_i de Mantel (Anselin, 1995B)</i>
<i>G(d) de Getis y Ord (1992)</i>	<i>G_i(d) de Getis y Ord (1992)</i>
	<i>G_i[*](d) de Getis y Ord (1992)</i>
	<i>Nueva-G_i de Ord y Getis (1995)</i>
	<i>Nueva – G_i[*] de Ord y Getis (1995)</i>
	<i>O_i de Ord y Getis (2001)</i>

Fuente: Elaboración propia.

III.2.3.1. Contrastes de autocorrelación espacial global

La perspectiva global del fenómeno de autocorrelación espacial tiene por objeto el contraste de la presencia de tendencias o estructuras espaciales generales en la distribución de una variable sobre un espacio geográfico completo. Es decir, se trata de contrastar la

¹⁸ Para un análisis más detallado sobre los tests de autocorrelación espacial con datos nominales (variables cualitativas), en Cliff y Ord (1973, 1981) se detalla el conocido test de recuento de vínculos de Moran, que es la medida de autocorrelación espacial más simple, así como el proceso inferencial asociado a dicho estadístico.

hipótesis de que una variable se encuentre distribuida de forma totalmente aleatoria en un sistema espacial o, si por el contrario, existe algún tipo de asociación significativa de valores similares o distintos entre regiones vecinas. Para ello, se han propuesto en la literatura varios estadísticos de dependencia espacial, como los tests I de Moran (1948), c de Geary (1954) y G(d) de Getis y Ord (1992), que son los contrastes más utilizados y fáciles de computar, aunque su interpretación no siempre sea muy directa.

1. Test I de Moran

El test I de Moran fue inicialmente formulado como función de una variable (Y), considerada en los puntos del espacio (i,j), en desviaciones a la media, y los elementos de la matriz binaria de interacciones espaciales [δ_{ij}]. Esta expresión inicial de Moran podría ser generalizada, sustituyendo la matriz de interacciones por la más general matriz de pesos espaciales, w_{ij} (Cliff y Ord, 1973, 1981; Anselin, 1995A), de la manera siguiente:

$$I = \frac{N}{S_0} \frac{\sum_{(2)} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad \text{Eq. III.2.10.}$$

donde: w_{ij} : elemento de la matriz de pesos espaciales correspondiente al par (i, j).
 $S_0 = \sum_i \sum_j w_{ij} = \sum_{(2)} w_{ij}$, es decir, la suma de los pesos espaciales.
 \bar{y} : valor medio o esperado de la variable y
 N: número de observaciones o tamaño muestral.

Cuando se utiliza una matriz de interacciones espaciales estandarizada por filas, que es la situación óptima de aplicación de este test, el término $S_0 = N$, dado que la suma de los valores de cada fila es igual a la unidad (ver Eq.III.2.5). De este modo, el estadístico I queda reducido al cociente del producto espacial cruzado de los valores de la variable partido por la varianza:

$$I = \frac{\sum_{(2)} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad \text{Eq. III.2.11.}$$

Adviértase que I está basado en los productos cruzados de las desviaciones de y_i respecto de \bar{y} (de forma análoga al test de recuento de vínculos NN que se presenta en el Anexo VI, Apartado A.VI.1). También es evidente que el test I de Moran es similar al coeficiente de autocorrelación temporal: el término del numerador es una medida de la covarianza entre valores de Y en dos localizaciones distintas (i,j) y el denominador expresa la varianza de Y en el punto i. Sin embargo, aunque parecido, el test I no es equivalente al

clásico coeficiente de correlación, fundamentalmente porque no se encuentra centrado en el valor cero. De hecho, la media teórica de la I de Moran es el cociente $-\frac{1}{N-1}$ ¹⁹.

En otras palabras, el valor esperado de I es negativo y función únicamente del tamaño de la muestra (N), aunque esta media tiende a cero a medida que el tamaño de la muestra aumenta. En cuanto a la varianza teórica del coeficiente I, se verá más adelante que depende de determinados supuestos estocásticos. Un coeficiente I de Moran mayor que su valor esperado sería indicativo de autocorrelación espacial positiva, mientras que un valor de I inferior a la media pondría de manifiesto la existencia de autocorrelación espacial negativa. Como puede apreciarse, este estadístico estará muy afectado por aquellos puntos vecinos sensiblemente distintos de la media de la variable en estudio.

Respecto a la distribución del contraste I, según Cliff y Ord (1981) cuando el tamaño muestral es suficientemente amplio, la expresión estandarizada del test I se distribuye como una normal tipificada, $N(0,1)$, como el test de recuento de vínculos de Moran. Por eso, también en este caso, en lugar de considerarse el estadístico inicial I, el proceso inferencial suele utilizar los valores estandarizados (z) de cada uno de ellos, obtenidos, como bien es sabido, a través del cociente entre la diferencia del valor inicial y la media teórica, y la desviación típica teórica, del modo siguiente²⁰:

$$z_I = \frac{I - E[I]}{SD[I]} \quad \text{Eq. III.2.12.}$$

siendo $E[I]$: media teórica del estadístico I.

$SD[I]$: desviación típica del estadístico I.

La interpretación de los valores estadísticamente significativos de la variable tipificada z_I sería la siguiente:

- Valores no significativos del test I estandarizado, z_I , correspondiente a una variable Y, conducirían a aceptar la hipótesis nula de no autocorrelación espacial o inexistencia de patrones de comportamiento de dicha variable sobre el espacio.
- Valores significativos de $z_I > 0$ serían indicativos de autocorrelación espacial positiva, es decir, que es posible encontrar valores parecidos (altos o bajos) de la variable Y, espacialmente agrupados, en mayor medida de como estarían por casualidad.
- Valores significativos de $z_I < 0$ serían indicativos de autocorrelación espacial negativa, es decir, que se produce una no-agrupación de valores similares (altos

¹⁹ Ver demostraciones en Cliff y Ord (1981), pp. 43-44.

²⁰ En Cliff y Ord (1981), se presenta el proceso inferencial de los contrastes I de Moran y c de Geary con mayor amplitud, así como las expresiones de los momentos de primer y segundo orden del test I de Moran necesarias para la obtención de su valor z estandarizado.

o bajos) de la variable Y superior a lo normal en un patrón espacial aleatorio. Se trata de un concepto algo más difícil de captar, que se encuentra representado, de forma perfecta, en la estructura del tablero del ajedrez (ver Anexo VI, Apartado A.VI.1, donde se analiza la autocorrelación espacial en el tablero del ajedrez).

El estadístico I de Moran permite también la obtención de un instrumento de AEDE que, como ya se ha presentado, es muy útil para el análisis de autocorrelación espacial: el **diagrama de dispersión de Moran** (“Moran scatterplot”), propuesto por Anselin (1993B). Efectivamente, del test I (Eq. III.2.10.) podría derivarse la siguiente expresión matricial:

$$I = \frac{N}{S_0} \frac{\sum_{(2)} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^N (y_i - \bar{y})^2} = \frac{N}{S_0} \cdot \frac{\dot{y}' W \dot{y}}{\dot{y}' \dot{y}} \quad \text{Eq. III.2.13.}$$

siendo N: n° observaciones
 S₀: suma de todos los elementos de la matriz de pesos espaciales
 \dot{y} : vector de observaciones en desviaciones a la media
 W \dot{y} : retardo espacial asociado a la variable y en desviaciones a la media

Cuando la matriz de pesos espaciales se encuentra estandarizada por filas (muy habitual en la práctica) de forma que los elementos de cada fila sumen 1 (S₀ = N), esta expresión se simplifica, tal como también se puso de manifiesto en la Eq. III.2.11.:

$$I = \frac{\dot{y}' W \dot{y}}{\dot{y}' \dot{y}} \Rightarrow \dot{y}' W \dot{y} = I \cdot \dot{y}' \dot{y} \quad \text{Eq. III.2.14.}$$

En esta expresión, el estadístico I resulta funcionalmente equivalente al coeficiente (pendiente) de una regresión lineal simple de W \dot{y} sobre \dot{y} (no de \dot{y} sobre W \dot{y} , que sería la forma más natural de especificar un proceso espacial). La interpretación del test I de Moran como pendiente de una línea de regresión es lo que permite visualizar la asociación lineal entre una variable (en desviaciones a la media) y su correspondiente retardo espacial en forma de diagrama de dispersión bivalente de W \dot{y} sobre \dot{y} ²¹. Este diagrama de dispersión es el denominado diagrama de dispersión de Moran, que se expuso en el Apartado III.1.3.2. como herramienta gráfica fundamental para el análisis de dependencia espacial en una serie. Este gráfico es también similar al diagrama de dispersión de retardos espaciales del AEDE geoestadístico, propuesto por Cressie (1991) y que se presenta en el Anexo IV.

El diagrama de dispersión de Moran es, como su nombre indica, un diagrama de dispersión en el que se representa también la línea de regresión, cuya pendiente, en este caso, será el valor del test I de Moran que, por este motivo, puede ser utilizado como indicador del grado de ajuste, así como de la presencia de valores atípicos en la nube de

²¹ Ver en Haining (1994) una interpretación del modo más habitual de regresión del vector y sobre el vector de retardos espaciales Wy.

puntos. Es importante también advertir que la regresión de $W \hat{y}$ sobre \hat{y} se realiza con todas las hipótesis clásicas del análisis de regresión y, por tanto, puede ser sujeto de todos los contrastes del ajuste de modelos. La pendiente de esta regresión es el estimador correcto del test I de Moran, aunque su significación (por ejemplo, utilizando el test de la t de Student) no sería, en este caso, apropiada (ver en el Apartado III.1.3.2. una más extensa presentación de esta técnica).

2. Test c de Geary (1954)

El test c de Geary que es, como el test I de Moran, función de la variable Y y los elementos de la matriz de interacciones espaciales $[\delta_{ij}]$, puede ser también generalizado sustituyendo esta matriz por la más general de pesos espaciales, w_{ij} (Anselin, 1995A), de la manera siguiente:

$$c = \frac{(N-1) \sum_{(2)} w_{ij} (y_i - y_j)^2}{2S_0 \sum_{i=1}^N (y_i - \bar{y})^2} \quad \text{Eq. III.2.15.}$$

El test de Geary considera el cuadrado de las diferencias entre los valores de Y en dos puntos (i,j) , a diferencia del contraste I. El valor esperado teórico del test c de Geary es 1, mientras que la varianza teórica depende, como en el caso de I, de algunos supuestos estocásticos (ver Anexo VI, Apartado A.VI.2). Por tanto, este test, depende de la diferencia (en valor absoluto) entre los valores de una variable en unidades vecinas, más que de las diferencias entre éstos y el valor medio de la variable (como en el caso del test I).

Debe advertirse también que el coeficiente c de Geary se parece en la forma al estadístico d de Durbin y Watson, establecido para contrastar la hipótesis alternativa de existencia de un proceso autorregresivo de primer orden en los residuos de una regresión (Durbin y Watson, 1950, 1951 y 1971)²².

También en este caso (como en el test I), puede afirmarse que el valor estandarizado de c (z_c) se distribuye como una normal tipificada, $N(0,1)$ asintótica. Por eso, también en este caso, el proceso inferencial suele utilizar los valores estandarizados, z_c , pudiéndose obtener a partir de la expresión de los momentos de c del modo siguiente (ver también Anexo VI, Apartado A.VI.2):

$$z_c = \frac{c - E[c]}{SD[c]} \quad \text{Eq. III.2.16.}$$

²² Efectivamente, este estadístico se define como: $d = \frac{\sum_{t=2}^N (e_t - e_{t-1})^2}{\sum_{t=1}^N e_t^2}$, siendo e_t los residuos de una regresión y t el período temporal de referencia..

siendo $E[c]$: media teórica del estadístico c

$SD[c]$: desviación típica del estadístico c .

Por su parte, la interpretación de los valores estadísticamente significativos de la variable tipificada z_c sería la siguiente:

- Valores no significativos del test c estandarizado, z_c , correspondiente a una variable Y , conducirían a aceptar la hipótesis nula de no autocorrelación espacial.
- Valores significativos de $z_c < 0$ serían indicativos de autocorrelación espacial positiva. Efectivamente, la existencia de valores similares de Y (altos o bajos) en unidades vecinas ($w_{ij} = 1$) tiende a anular el numerador del test c , por lo que la diferencia entre el valor de este estadístico y su media (que es igual a 1) será un número negativo.
- Valores significativos de $z_c > 0$ serían indicativos de autocorrelación espacial negativa.

3. Test Γ de Mantel (1967)

Anselin (1995B) menciona también como contraste de dependencia global, el índice general de asociación matricial o índice Γ , originalmente presentado por Mantel (1967), que se expresa como la suma de los productos cruzados de los elementos coincidentes a_{ij} y b_{ij} de dos matrices de similitudes, A y B , del modo siguiente:

$$\Gamma = \sum_i \sum_j a_{ij} b_{ij} \quad \text{Eq. III.2.17.}$$

A su vez, Cliff y Ord (1981) ya habían presentado este estadístico como la base metodológica a partir de la cual se definen las distribuciones de los tests I de Moran y c de Geary. Efectivamente, las medidas de asociación espacial pueden obtenerse, en general, expresando la similitud espacial en una matriz (por ejemplo, una matriz de contigüidades o de pesos espaciales) y la similitud de valores de una variable en otra. Distintas medidas de similitud de valores dan lugar a diferentes índices de asociación espacial. Por ejemplo, para $a_{ij} = w_{ij}$:

- $b_{ij} = y_i y_j$: conduce a una medida parecida al test I de Moran.
- $b_{ij} = (y_i - y_j)^2$: da lugar a una índice parecido al test c de Geary.

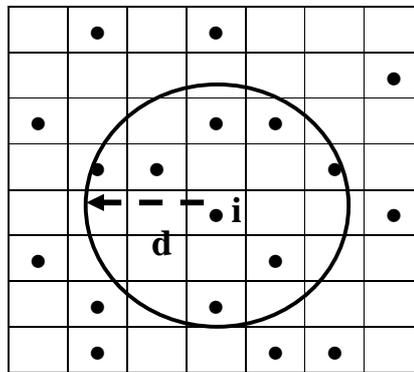
4. Test $G(d)$ de Getis y Ord (1992)

Getis y Ord sugirieron un enfoque ligeramente diferente a los utilizados hasta entonces para medir la autocorrelación espacial, basado en los llamados estadísticos de

distancias o de concentración espacial. El cálculo de estos estadísticos exige la definición, en un espacio, de un conjunto de unidades vecinas a cada punto o región, entendidas éstas como aquellas observaciones que se encuentran a una distancia crítica (d) de dicho punto, de forma que, para cada distancia diferente d , se construya una matriz de ponderaciones espaciales propia, $W(d)$. Con este criterio, estos autores han desarrollado la llamada familia de estadísticos G , que sólo pueden estimarse para variables con valores positivos.

Estas medidas se caracterizan por medir el grado de asociación existente entre los valores de una variable Y en un conjunto dado de puntos (o en el área representada por un único punto, como sucede en la Figura III.2.7) y los valores de dicha variable en otros puntos incluidos en de un radio de distancia d respecto al grupo original.

Figura III.2.7. Ejemplo de sistema de autocorrelacion espacial local



Fuente: Elaboración propia.

El llamado estadístico general $G(d)$ es un contraste global, para toda la muestra, a diferencia de otros tests pertenecientes a la misma familia de indicadores G que, por ser de carácter local, serán presentados en el apartado siguiente. Este estadístico, según el cual dos pares de regiones (i,j) serán vecinas siempre que se encuentren dentro de una distancia d determinada, adopta la forma siguiente:

$$G(d) = \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}(d) y_i y_j}{\sum_{i=1}^N \sum_{j=1}^N y_i y_j} \quad ; \quad \text{para } j \neq i \quad \text{Eq. III.2.18.}$$

donde $w_{ij}(d)$ adopta el valor 1 en los pares de regiones (i,j) situadas dentro de la distancia umbral d y cero, en caso contrario. Su significación estadística se comprueba también a través del valor z_G estandarizado de $G(d)$, obtenido del mismo modo que los anteriormente presentados. En este caso, la hipótesis nula asociada al contraste estandarizado será la

ausencia de autocorrelación, mientras que un valor de z_G significativo positivo (o negativo) indicará la existencia de una tendencia a la concentración de valores similares elevados (o bajos, respectivamente) de la variable Y en el espacio analizado.

Los tres contrastes de autocorrelación espacial global, I , c y $G(d)$, proporcionan una información complementaria y pueden ser utilizados conjuntamente (Moreno y Vayá, 2000). Sin embargo, deben conocerse también las diferencias existentes entre ellos, algunas de las cuales se resumen en la Tabla III.2.3. Por ejemplo, el test I es una medida similar a la covarianza (diferencia de cada y_i respecto del valor medio, \bar{y}) de los valores de Y en cada punto del espacio, mientras que el contraste c de Geary podría ser considerado como un análisis de la varianza, más afectado por la distribución de los datos muestrales que I (Cliff y Ord, 1981). Por su parte, el test $G(d)$ es en realidad una medida de concentración o asociación de una variable geográfica Y , que presenta un gran parecido formal con el test I , sobre todo en el numerador de ambas expresiones matemáticas: mientras que $G(d)$ es proporcional a la suma de los productos $y_i \cdot y_j$, el contraste I mide la correlación de cada valor y_i con sus vecinos (Getis y Ord, 1992).

Una importante limitación, propia de la familia de estadísticos G , consiste en que sólo pueden aplicarse a variables positivas, por lo que el contraste $G(d)$, a diferencia del test I , no podrá ser utilizado para analizar la presencia de autocorrelación espacial, por ejemplo, en los residuos de una regresión. Además, debe tenerse en cuenta que la matriz W , cuando se aplica a la obtención del test $G(d)$, debe ser simétrica por lo que no podrá realizarse, en estos casos, ningún tipo de estandarización por filas en la misma. Estas restricciones, sin embargo, no tienen lugar en los tests I y c , que pueden ser aplicados a valores positivos y negativos de cualquier variable continua, así como a cualquier tipo de matriz de pesos.

Tabla III.2.3. Interpretación de los valores estandarizados de los estadísticos de autocorrelación espacial global

Test	Hipótesis nula (z no significativo)	Hipótesis alternativa (z significativo)	
		$z > 0$	$z < 0$
I de Moran	No autocorrelación espacial	Autocorrelación espacial POSITIVA	Autocorrelación espacial NEGATIVA
c de Geary	No autocorrelación espacial	Autocorrelación espacial NEGATIVA	Autocorrelación espacial POSITIVA
G(d) de Getis y Ord	No autocorrelación espacial	Autocorrelación espacial POSITIVA (valores de Y altos)	Autocorrelación espacial POSITIVA (valores de Y bajos)

Fuente: Elaboración propia.

Por último, tampoco debe olvidarse la importancia que tiene la matriz de pesos espaciales en el proceso de contraste del fenómeno de autocorrelación espacial global, pues

está bien demostrado que los resultados obtenidos por los diferentes estadísticos pueden variar, a veces de forma sensible, en función de la matriz W especificada. Por eso, se recomienda la utilización, en la aplicación de estos contrastes, de diversas expresiones de W . También suele recomendarse calcular el test I de Moran para matrices de contigüidad de órdenes o potencias sucesivas²³ porque permite contrastar si el esquema de autocorrelación espacial detectado entre regiones vecinas es extensible a regiones alejadas en el espacio (Anselin y Smirnov, 1996; Moreno y Vayá, 2000). En el Anexo VI, Apartado A.VI.2, se presentan algunas de las limitaciones de los tests de autocorrelación espacial que han sido demostradas por algunos autores, como Dacey (1965), Cliff y Ord (1981), Anselin (1988A), Chou (1991), Mur (2000), entre otros.

III.2.3.2. Contrastes de autocorrelación espacial local

Los estadísticos de autocorrelación global, centrados en el análisis de dependencia general propia de todas las unidades de un espacio geográfico, no son capaces de detectar la inestabilidad o deriva espacial de ciertas estructuras locales de asociación (“hotspots” o puntos calientes/fríos) o inestabilidades locales que pueden estar, a su vez, presentes o no en una estructura global de dependencia (Getis y Ord, 1992; Openshaw, 1993; Anselin, 1993B, 1995B; Ord y Getis, 1995, 2001; Vayá y Suriñach, 1996; Tiefelsdorf y Boots, 1997; Sokal et al., 1998).

El fenómeno de autocorrelación, dependencia o asociación espacial local puede ser definido como una concentración, en un lugar del espacio global analizado, de valores especialmente altos o bajos de una variable en comparación con el valor medio esperado (o media de la variable considerada). Este fenómeno se produce en procesos espaciales no estacionarios, lo que contrasta con el hecho de que el análisis de la asociación espacial ha estado, durante mucho tiempo, asentado sobre el principio de estacionariedad espacial, hipótesis ésta que ha sido reconocida como bastante irreal, sobre todo cuando se utiliza un gran número de unidades geográficas (Aznar *et al.*, 1996). Efectivamente, en muchas ocasiones las propiedades estadísticas (función de distribución conjunta) de un proceso espacial no dependen sólo de la posición relativa (ángulo) y las distancias respectivas de los puntos en el espacio, sino también de cuestiones como el contexto económico. Es decir, no siempre puede aceptarse un tipo de dependencias similar entre cualquier par de puntos situados a la misma distancia en un mapa, por ejemplo, entre las ciudades de Madrid y Barcelona, por un lado, y Madrid y Ourense, por otro.

Este problema de la dependencia espacial local puede plantearse desde dos puntos de vista (Vayá y Suriñach, 1996):

- Existe la posibilidad de que, en un espacio dado, no se detecte la presencia de autocorrelación espacial global en la distribución de una variable aunque, de hecho, existan pequeños “clusters” espaciales en los que dicha variable experimenta una concentración (o escasez) importante.

²³ Ver en Anselin (1988A), la elaboración y problemática de las matrices de interacciones espaciales para potencias superiores o iguales a dos.

- Existe también la posibilidad de que, habiéndose detectado dependencia a nivel global en una variable, no todas las regiones del espacio considerado contribuyan con igual peso en el indicador global, es decir, que coexistan unas zonas en las que la variable se distribuya de forma aleatoria junto a otras con una importante contribución a la dependencia existente.

Para responder a estas cuestiones, se definen a continuación dos grupos de contrastes de asociación local que indican hasta qué punto una región se encuentra rodeada por otras con valores altos o bajos de una variable determinada: la familia de estadísticos G_i de Getis y Ord (1992, 1995) y los indicadores locales de asociación espacial (LISA), propuestos por Anselin (1995B). Ambos tipos de contrastes parten de la hipótesis nula de ausencia de autocorrelación espacial global, aunque los indicadores LISA serán también capaces de responder a la segunda de las cuestiones planteadas, demostrando la presencia de regiones con una participación en el estadístico global muy superior a la media (“outliers”), como se verá más adelante. Recientemente, Ord y Getis (2001) han presentado el estadístico O_i , que es también capaz de detectar la presencia de autocorrelación espacial local en presencia de autocorrelación global.

1. Familia de estadísticos locales de Getis y Ord

Como ya se ha expuesto, Getis y Ord sugirieron un enfoque diferente para el contraste del fenómeno de autocorrelación espacial, a partir de estadísticos de distancias o concentración. El cálculo de estos estadísticos de dependencia local en un punto (i) exige la definición de una distancia crítica (d) en torno a cada punto. Es decir, si se definen bandas de distancias diferentes, deben también construirse distintas matrices de pesos espaciales, correspondientes a cada distancia, $W(d)$.

En primer lugar, los autores definen un área total subdividida en n regiones, $i = 1, 2, \dots, N$, de forma que cada una de las regiones i se encuentra identificada con un punto, de coordenadas cartesianas conocidas, que lleva asociado un valor (y) de la variable Y , que ha de ser positiva y con un origen natural. El **estadístico $G_i(d)$** permite contrastar la hipótesis alternativa de existencia de autocorrelación espacial en el conjunto de valores de y asociados a los j puntos incluidos en un radio d de un punto inicial dado i .

$$G_i(d) = \frac{\sum_{j=1}^N w_{ij}(d)y_j}{\sum_{j=1}^N y_j} \quad ; \quad \text{para } j \neq i \quad \text{Eq. III.2.19.}$$

donde $[w_{ij}(d)]$ es la matriz simétrica de interacciones espaciales (0-1), tal que:

$w_{ij} = 1$: para todos los vínculos existentes entre aquellos puntos j situados en un radio d del punto i .

$w_{ij} = 0$: para el resto de vínculos, incluida la relación entre el punto i consigo mismo.

En el numerador se consideran únicamente los valores y_j situados en un radio d del punto i (excepto y_i) mientras que, en el denominador, se consideran todos los y_j del sistema, exceptuando también y_i . Obsérvese que, para cada punto del espacio i , se obtiene un valor de $G_i(d)$ que, dependiendo de la distancia considerada d , será también distinto. En cuanto al valor de d , debe ser interpretado como una distancia que incluye un determinado número de unidades espaciales dentro del espacio considerado. Por eso, es lógico pensar que existirá correlación entre los valores de $G_i(d)$ correspondientes a unidades vecinas. Es más, como el denominador de los estadísticos G_i permanecerá constante, el coeficiente de correlación entre dos valores de dicho estadístico en dos unidades próximas i, j (G_i, G_j) será la proporción de unidades vecinas que i y j tienen en común.

El estadístico $G_i(d)$ mide la concentración (o ausencia de concentración) de la suma ponderada de los valores de la variable Y en una región. Es decir, $G_i(d)$ será proporcional a la suma de todos los valores y_j incluidos en un radio d de i , de forma que $G_i(d)$ tendrá un valor alto si en el radio d se concentran valores altos de y_j .

Getis y Ord (1992) proponen también un estadístico G relacionado con éste, $G_i^*(d)$, que mide la asociación espacial en aquellos casos en los que se también se considera la relación $j = i$. Es decir, el concepto de concentración de los valores de Y en torno al punto i tiene en cuenta también el valor de dicha variable en el propio punto i (y_i) para el cual se calcula el estadístico. Además, como por convenio los pesos $w_{ii} = 0$, la diferencia existente entre ambos estadísticos, en términos matemáticos, radica en el número de observaciones incluidas en el denominador, $\sum_j y_j$, del modo siguiente:

$$G_i^*(d) = \frac{\sum_{j=1}^N w_{ij}(d)y_j}{\sum_{j=1}^N y_j} \quad ; \quad \forall j \quad \text{Eq. III.2.20.}$$

Tanto el test $G_i(d)$ como G_i^* , tal como fueron inicialmente formulados, sólo pueden ser aplicados al caso de variables naturales (no enteras) positivas, así como para matrices de interacciones espaciales binarias y simétricas (no estandarizadas por filas). Para superar estas limitaciones, Getis y Ord (1995) presentaron, más tarde, dos nuevas expresiones de estos contrastes, que podrían denominarse (Moreno y Vayá, 2000) **Nueva- G_i** y **Nueva- G_i^*** . Estas medidas se obtienen mediante la estandarización de las antiguas²⁴:

²⁴ Los desarrollos de las siguientes expresiones pueden consultarse en Getis y Ord (1992) y Ord y Getis (1995).

$$Nueva-G_i = \frac{G_i(d) - E[G_i(d)]}{SD[G_i(d)]} = \frac{\sum_{j=1}^N w_{ij}(d)y_j - W_i \bar{y}(i)}{SD(i) \left\{ \left[\frac{(n-1)S_{ii} - W_i^2}{(N-2)} \right] \right\}^{\frac{1}{2}}} ; \quad j \neq i \quad Eq. III.2.21.$$

$$Nueva-G_i^* = \frac{G_i^*(d) - E[G_i^*(d)]}{SD[G_i^*(d)]} = \frac{\sum_{j=1}^N w_{ij}(d)y_j - W_i^* \bar{y}}{SD(y) \left\{ \left[\frac{NS_{ii}^* - W_i^{*2}}{(N-1)} \right] \right\}^{\frac{1}{2}}} ; \quad \forall j \quad Eq. III.2.22.$$

siendo $E[G_i(d)], E[G_i^*(d)]$: media teórica de los estadísticos $G_i(d)$ y $G_i^*(d)$
 $SD[G_i(d)], SD[G_i^*(d)]$: desviación típica de los estadísticos $G_i(d)$ y $G_i^*(d)$

N : n° unidades espaciales

$$W_i = \sum_{j \neq i} w_{ij}(d) ; \quad W_i^* = W_i + w_{ii} , \quad \forall w_{ii} \neq 0^{25}$$

$$S_{ii} = \sum_{j=1}^N w_{ij}^2 ; \quad j \neq i ; \quad S_{ii}^2 = \sum_{j=1}^N w_{ij}^2 ; \quad \forall j$$

$$\bar{y}(i) = \frac{\sum_j y_j}{N-1} ; \quad \bar{y} = \frac{\sum_j y_j}{N} ; \quad SD(i) = \sqrt{\frac{\sum_j (y_j - \bar{y}(i))^2}{N-1}} ; \quad SD(y) = \sqrt{\frac{\sum_j (y_j - \bar{y})^2}{N}}$$

La inferencia de los estadísticos G_i y $G_i^*(d)$ se fundamenta, como en otros casos, en sus correspondientes valores estandarizados, es decir, en las Nueva- G_i y Nueva- G_i^* que se considera se distribuyen como una normal, aproximadamente. Es decir, se dice que una variable posee autocorrelación espacial positiva o negativa cuando los valores de estos estadísticos estandarizados son positivos o negativos, respectivamente, por encima de un determinado nivel de significación.

Debe advertirse que esta interpretación, como la correspondiente al estadístico $G(d)$ global, es diferente del significado del resto de contrastes de autocorrelación espacial global (I de Moran y c de Geary), en los cuales una concentración espacial de valores parecidos de una variable (tanto altos como bajos), ponía de manifiesto la existencia de autocorrelación espacial positiva, mientras que la unión de valores diferentes (al estilo del tablero de ajedrez), implicaba una dependencia negativa. La familia de estadísticos G de Getis y Ord, por su parte, permiten detectar procesos de dependencia espacial positivos (no negativos), diferenciando entre situaciones de agrupamiento de valores altos o bajos.

Recientemente, ambos autores han propuesto también un nuevo estadístico O_i capaz de detectar, sin errores, la existencia de pequeños agujeros de autocorrelación espacial local en presencia de un fenómeno más general y extenso de autocorrelación global. Según Ord y Getis (2000), es fácil confundir un fenómeno de dependencia local con la presencia, en un

²⁵ Como puede observarse, los autores admiten la posibilidad de asimetría en la matriz de pesos espaciales.

determinado lugar, de valores altos o bajos de una variable. Por ejemplo, nunca debería confundirse una zona caliente (“hotspot”) de secciones censales en las que se produce una gran concentración viviendas caras, con lo que sería una manzana o bloque de casas de alto precio en un espacio continuo. La idea de “concentración” de valores que subyace en el fenómeno de autocorrelación espacial local, lleva consigo una noción de extraordinario o atípico (“outlier” en la distribución de una variable geográfica), que no debería confundirse con la simple presencia de valores altos o bajos de la misma en un determinado lugar.

2. Indicadores Locales de Asociación Espacial (LISA), de Anselin

Anselin (1995B) propone un conjunto de indicadores locales de asociación espacial, LISA (“Local Indicators of Spatial Association”)²⁶, capaces de detectar la contribución de cada región a un indicador de dependencia espacial global (por ejemplo, el estadístico I de Moran), así como posibles valores atípicos (“outliers”), con una participación en el estadístico global muy superior o inferior a la media. Este tipo de indicadores permiten la consecución de dos objetivos, como se expuso anteriormente: la determinación de agrupamientos espaciales locales significativos en torno a un punto concreto del espacio (como los estadísticos G_i y G_i^* de Getis y Ord) y la obtención de bolsas de inestabilidad (no estacionariedad) espacial, es decir, la presencia de valores atípicos que también pueden ser visualizados mediante el diagrama de dispersión de Moran.

Los indicadores LISA presentados por este autor son: los estadísticos Gamma, I_i de Moran y c_i de Geary.

- a) **Estadístico local Gamma.** Este estadístico, referido a un punto o región i , está basado en el índice general de asociación matricial o índice Γ (ver *Eq. III.2.17*) y puede ser definido como:

$$\Gamma_i = \sum_j a_{ij} b_{ij} \quad \text{Eq. III.2.23.}$$

La suma de los estadísticos Γ_i dan lugar a la medida global Γ . Es posible que la distribución de los valores individuales de Γ_i pueda llevarse a cabo mediante un enfoque de permutaciones condicionales, como en otros casos (ver en Anexo VI, Apartado A.VI.2), lo que hace posible la interpretación de los valores Γ_i individuales como indicadores de significativos agrupamientos espaciales, de ámbito local. Por otro lado, la segunda interpretación del estadístico LISA como contraste de valores atípicos o puntos de influencia puede también llevarse a cabo mediante la comparación de los valores de Γ_i con el ratio Γ/n .

- b) **Estadístico local de Moran.** Como caso particular del estadístico local Gamma, podría también definirse un indicador de dependencia local basado en el test I de Moran, I_i , del modo siguiente:

²⁶ Ven en Anselin (1995B), una presentación más extensa del concepto de estadísticos LISA.

$$I_i = z_i \sum_{j=1}^{J_i} w_{ij} z_j \quad \text{Eq. III.2.24.}$$

donde z_i, z_j : variable y_i estandarizada

\sum_j : sumatorio que únicamente incluye los valores vecinos a i : $j \in J_i$.

Para lograr una mejor interpretación de este estadístico, la matriz de pesos w_{ij} debería estar estandarizada por filas (aunque no es necesario), siendo, por convenio, cada elemento $w_{ii} = 0$. También, en este caso, es fácil comprobar que la suma de estadísticos locales I_i es, sin duda, el conocido test I de Moran:

$$\sum_i I_i = \sum_i z_i \sum_j w_{ij} z_j = \sum_i \sum_j w_{ij} z_i z_j \quad \text{Eq. III.2.25.}$$

Efectivamente, el test I de Moran (Eq. III.2.10) puede expresarse como:

$$I = \frac{N}{S_0} \frac{\sum_{(2)} w_{ij} z_i z_j}{\sum_{i=1}^n z_i^2} \Rightarrow I = \frac{\sum_i I_i}{S_0 \sum_i \frac{z_i^2}{N}} \quad \text{Eq. III.2.26.}$$

siendo $\Sigma_{(2)} = \Sigma_i \Sigma_j$

De este modo, se cumple que el sumatorio de valores del estadístico local I_i es equivalente al test global I, siendo el factor de proporcionalidad $\gamma = S_0 m_2$:

$$\sum_i I_i = \frac{1}{S_0 m_2} I = \gamma \cdot I \quad \text{Eq. III.2.27.}$$

donde: $m_2 = \sum_i \frac{z_i^2}{N}$, momento de 2º orden de la variable z_i .

En el caso de que la matriz W esté estandarizada por filas, $S_0 = N$, por lo que el factor $\gamma = \sum_i z_i^2$ y, en el caso de trabajar también variables estandarizadas, $m_2 =$

1, por lo que dicho factor $\gamma = S_0$. Por eso, el mismo resultado obtenido en la Eq. III.2.24 se consigue dividiendo todos los indicadores locales por m_2 , que es un término constante en todas las localizaciones, por lo que el test I local de Moran podría también expresarse del modo siguiente:

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \quad \text{Eq. III.2.28.}$$

Como puede verse con detalle en Anselin (1995B), es posible calcular los momentos de I_i , bajo la hipótesis nula de ausencia de asociación espacial, para el supuesto inferencial de aleatoriedad o muestreo aleatorio²⁷. Cualquier contraste de significación de asociación espacial local puede basarse en estos momentos, aunque la distribución exacta de un estadístico de este tipo aún se desconoce. Para facilitar su interpretación, puede utilizarse una distribución normal, pero hay que saber que se trata sólo de una simple aproximación. Alternativamente, es también posible utilizar una aproximación aleatoria condicional (hipótesis de permutación) que, dada la estructura del estadístico local de Moran, consistiría en el cálculo del sumatorio $\sum_j w_{ij} z_j$ para cada permutación (dado que el cociente z_i/m_2 permanece constante en todas las localizaciones).

La interpretación del estadístico local I_i de Moran como un indicador de inestabilidad local se desprende fácilmente de la relación entre estadísticos locales y globales. En concreto, la media de I_i será igual a la del estadístico global I por un factor de proporcionalidad. Las máximas contribuciones de los valores de I_i al estadístico global I pueden ser identificadas a través de criterios sencillos, como el “doble-sigma” o identificando los puntos atípicos en un gráfico de caja. Debe advertirse que este concepto de “extremosidad” únicamente pone de manifiesto la importancia de la observación i en la determinación del estadístico global. Anselin propone también el diagrama de dispersión de Moran, aunque no es un indicador LISA en sentido estricto (no ofrece información alguna acerca de la existencia o no de autocorrelación espacial local), como buen instrumento para la detección de puntos atípicos en una distribución.

- c) **Estadístico local de Geary.** Siguiendo el mismo desarrollo y nomenclatura anteriores, el estadístico local de Geary, para cada observación i , puede definirse de las siguientes dos maneras:

$$c_i = \sum_j w_{ij} (z_i - z_j)^2 \quad \text{Eq. III.2.29.}$$

$$c_i = \frac{1}{m_2} \sum_j w_{ij} (z_i - z_j)^2 \quad \text{Eq. III.2.30.}$$

A partir de la *Eq. III.2.30*, y sin pérdida de generalidad, el sumatorio de los valores de c_i para todas las observaciones da lugar a la siguiente expresión:

²⁷ En Cliff y Ord (1973), se presentan los tres supuestos o hipótesis establecidos en la literatura para la inferencia de los estimadores de autocorrelación espacial: normalidad, aleatoriedad y permutación.

$$\sum_i c_i = N \frac{\sum_i \sum_j w_{ij} (z_i - z_j)^2}{\sum_i z_i^2} \quad \text{Eq. III.2.31.}$$

dado que, cuando se trabaja con variables estandarizadas, $m_2 = \sum_i \frac{z_i^2}{N} = 1$

Como puede observarse, el sumatorio de los valores del estadístico local c_i es equivalente al estadístico global c de Geary, presentado en la Eq. III.2.15, siendo γ el factor de proporcionalidad entre ambos estadísticos:

$$\gamma = \frac{2NS_0}{(N-1)} \quad \text{Eq. III.2.32.}$$

En el caso de que W sea una matriz estandarizada por filas, como $S_0 = N$, este factor se transforma del modo siguiente: $\gamma = \frac{2N^2}{(N-1)}$. La interpretación e inferencia de este estadístico es similar a la de los tests locales Gamma y Moran.

CAPÍTULO IV

Análisis confirmatorio: modelos de regresión espacial

En el Capítulo anterior, se presentaron los efectos de dependencia y heterogeneidad espacial, que son fenómenos propios de los datos espacialmente distribuidos, así como, en concreto, la especificación y contrastes de dependencia o autocorrelación espacial sobre una variable geográfica. Ahora, en el presente Capítulo, se analiza la presencia de dichos efectos de dependencia y heterogeneidad espacial en el contexto de los modelos de regresión lineal, sus consecuencias sobre la especificación (Apartado IV.1), estimación y contraste, así como estrategias para la selección de modelos (Apartado IV.2.). El correcto tratamiento de los efectos espaciales en el proceso de modelización espacial constituye el denominado análisis confirmatorio de datos.

El esquema del capítulo es el siguiente:

CAPÍTULO IV. ANÁLISIS CONFIRMATORIO: MODELOS DE REGRESIÓN ESPACIAL

IV.1. Especificación de los efectos espaciales en modelos de regresión

IV.1.1. Modelos de dependencia espacial o modelos espaciales dinámicos

IV.1.2. Modelos de heterogeneidad espacial

IV.2. Métodos de estimación y contraste en modelos de regresión espacial

IV.2.1. Estimación y contrastes en el modelo básico de regresión lineal.

IV.2.2. Estimación y contrastes en modelos de dependencia espacial

IV.2.3. Estimación y contrastes en modelos de heterogeneidad espacial

IV.1. ESPECIFICACIÓN DE LOS EFECTOS ESPACIALES EN MODELOS DE REGRESIÓN

En el Apartado II.1., se definía la econometría espacial, con palabras de Anselin (2001D), como “la parte de la econometría que se ocupa del tratamiento de la interacción espacial (autocorrelación espacial) y la estructura espacial (heterogeneidad espacial) en los modelos de regresión de corte transversal y de datos de panel”. Es decir, que podría decirse que la principal característica de la econometría espacial consistiría en el modo como los efectos espaciales de dependencia y heterogeneidad espacial son tenidos en cuenta en el contexto del análisis de regresión, para evitar problemas de mala especificación en los modelos. Obviamente, esto implica que las propiedades del espacio geográfico han sido previamente especificadas matemáticamente, por ejemplo, a través de una matriz de pesos espaciales.

Uno de los problemas más importantes planteados en este campo de la modelización espacial, ha sido la enorme variedad de especificaciones aplicadas a los procesos espaciales propios, por ejemplo, de la ciencia regional. Podría dar la impresión de que cada modelo particular necesita su propio esquema metodológico, cuando afortunadamente, gracias a las técnicas econométricas, es posible agrupar estas especificaciones y realizar tipologías de modelos espaciales, como las que se presentan a continuación.

En primer lugar, el efecto de dependencia o autocorrelación espacial puede (o no) estar presente en la especificación de un modelo de regresión lineal de dos formas, que han sido denominadas como sustantiva o residual, ya sea como consecuencia de variables sistemáticas (endógena y/o exógenas) espacialmente autocorrelacionadas o por la existencia de un esquema de dependencia espacial en el término de la perturbación aleatoria, respectivamente (Apartado IV.1.1). Por otro lado, el fenómeno de heterogeneidad espacial debe también especificarse en un modelo cuando la inestabilidad estructural, propia de este efecto, se manifieste como varianza no constante en los residuos de una regresión (heteroscedasticidad) o como variación en los coeficientes del modelo (inestabilidad paramétrica) que, a su vez, puede ser continua o discreta, como se verá en el Apartado IV.1.2. Además, los efectos espaciales de autocorrelación y heterogeneidad pueden ser también especificados en otros esquemas econométricos, como los modelos de datos de panel²⁸, modelos multiecuacionales, modelos de variables truncadas, modelos no lineales, cuya presentación excede los objetivos de este trabajo y queda para futuros desarrollos.

En este Apartado, se exponen las principales especificaciones que han recibido los efectos espaciales en el seno de los modelos de regresión espacial o espacio-temporal, con el objetivo de ofrecer una visión general estructurada en tipologías de este tipo de modelos.

²⁸ Dada la importancia y amplia difusión de este tipo de modelización, recomendamos al lector interesado la lectura de Anselin (1988A) y Moreno y Vayà (2000), donde se presentan las principales especificaciones de los efectos espaciales en modelos de datos de panel: modelos SUR-espaciales, modelos de componentes del error (MCE) y modelos de ecuaciones simultáneas.

En los Apartados siguientes, se ampliará este tema incluyendo los métodos de estimación, contrastación y validación más frecuentemente citados en la literatura, para alguno de estos modelos. Asimismo, en el Anexo I, se presentan algunos ejemplos de modelización econométrica espacial aplicados a la economía y, en general, las ciencias sociales.

IV.1.1. Modelos de dependencia espacial o modelos espaciales dinámicos

Los modelos de dependencia espacial, también denominados modelos espaciales dinámicos, podrían definirse como aquellos modelos de regresión lineal que consideran explícitamente la existencia del efecto espacial de dependencia o autocorrelación. El análisis y corrección de la dependencia espacial en un modelo es fundamental, tanto desde el punto de vista estadístico como económico (Pons y Viladecans, 1998). Por un lado, la presencia de autocorrelación espacial en un modelo, como se demostrará más adelante, invalida la inferencia estadística derivada de la estimación por el tradicional método MCO (mínimos cuadrados ordinarios). A su vez, desde una perspectiva económica, el estudio del efecto de dependencia espacial permite detectar fenómenos como externalidades o efectos “spillover” (desbordamiento) en una determinada unidad espacial.

El efecto de dependencia espacial puede estar presente en un modelo en forma de autocorrelación espacial en alguna variable del mismo (endógena y/o exógenas) o también como existencia de esquemas de dependencia espacial en el término de la perturbación aleatoria, situaciones que han sido denominadas, respectivamente, como dependencia sustantiva (“sustantive dependence”) y dependencia residual (“nuisance dependence”). En cualquiera de ambos esquemas, la inclusión del efecto de autocorrelación espacial en un modelo de regresión requiere de la utilización de una matriz de pesos espaciales, W , capaz de recoger las influencias mútuas presentes entre las unidades espaciales de la muestra.

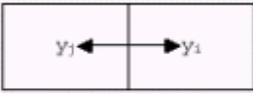
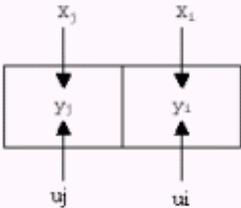
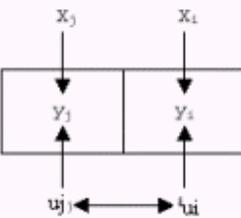
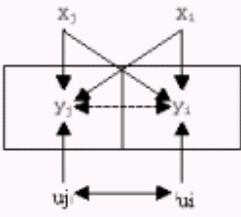
A continuación, se presentan las dos taxonomías de modelos espaciales de regresión lineal más citadas en la literatura y que, como se observará, se complementan, así como una tipología para modelos de externalidades, recientemente propuesta por Anselin (2001E), en la que se amplía el espectro y las posibilidades en este tipo de modelos²⁹. Por tanto, en primer lugar, se expone la tipología sencilla, adoptada por muchos autores en la literatura, que se asienta sobre 4 modelos explicativos de la dependencia espacial en un fenómeno a investigar (modelo de regresión espacial autorregresivo, modelo básico de regresión lineal, modelo del error espacial y modelo del retardo espacial). Este enfoque se complementa con otra tipología más completa que parte de un modelo general para descender a otros casos particulares.

²⁹ Esta exposición intenta completar las tipologías propuestas por Anselin (1988A) y Florax y Folmer (1992) que han sido presentadas, a su vez, por otros muchos autores, con algunas variaciones (ver por ejemplo, Cerejeira, 1998; Moreno y Vayá, 2000). La denominación de los modelos se ha traducido del inglés teniendo en cuenta los trabajos en español de Mur (1999) y, sobre todo, Moreno y Vayá (2000).

IV.1.1.1. Primera taxonomía de Anselin

En la Figura IV.1.1., se presentan distintas formas de tratamiento del fenómeno de dependencia espacial, presente en la variable endógena de un modelo de regresión lineal, que dan lugar a las diversas especificaciones presentadas por Anselin (1988A).

Figura IV.1.1. Tratamiento de la dependencia espacial: primera tipología de modelos

<p>Autocorrelación espacial univariante</p>		$y = \rho W y + u$ $u \approx N(0, \sigma^2 I)$	<p>Modelo autorregresivo de regresión espacial</p>
<p>Relación estructural entre variables</p>		$y = X\beta + u$ $u \approx N(0, \sigma^2 I)$	<p>Modelo básico de regresión espacial (MBRL)</p>
<p>Dependencia espacial residual</p>		$y = X\beta + u$ $u = \lambda W u + s$ $s \approx N(0, \sigma^2 I)$ $y = X\beta + u$ $u = \lambda W s + s$ $s \approx N(0, \sigma^2 I)$	<p>Modelo de regresión con dependencia espacial en la perturbación aleatoria (modelo del error espacial)</p>
<p>Dependencia espacial sustantiva</p>		$y = \rho W y + X\beta + u$ $u \approx N(0, \sigma^2 I)$	<p>Modelo mixto autorregresivo de regresión espacial (modelo del retardo espacial)</p>

Fuente: Elaboración propia, a partir de Baller *et al.* (2000).

➤ **Modelo autorregresivo de regresión espacial de orden 1, SAR(1)**

Este modelo constituye la especificación más sencilla de dependencia espacial, adecuado para expresar situaciones en las que los valores que adopta una variable determinada dependen sistemáticamente de la localización geográfica de la misma. Este

comportamiento de la variable debe ser contrastado estadísticamente mediante las técnicas del AEDE (ver Apartado III.1.). En la Figura IV.1.1., se presenta, en primer lugar, este modelo que podría expresarse de la siguiente manera:

$$\begin{aligned} y &= \rho Wy + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.1.}$$

donde y: vector (N,1) de observaciones de la variable endógena
 W: matriz de pesos espaciales de la variable endógena, siendo “Wy” el retardo espacial de la variable endógena
 ρ: coeficiente autorregresivo espacial (escalar), que recoge la intensidad de las interdependencias entre las observaciones muestrales
 u: perturbación aleatoria ruido blanco.

Esta formulación describe una situación en la que los valores de una variable (y) en un punto geográfico (i) están condicionados por los valores que adopta dicha variable en otro u otros puntos geográficos (j), y viceversa, dando lugar a fenómenos de agrupamiento de valores (altos/bajos) de dicha variable en determinadas zonas del espacio geográfico. Dicho en términos estadísticos, la $\text{Cov}(y_i, y_j) \neq 0$. La simultaneidad en la relación de dependencia entre y_i, y_j , que viene reflejada por una flecha de doble dirección en la Figura IV.1.1., no tiene lugar en el ámbito temporal, donde el fenómeno de autocorrelación tiene un carácter secuencial, no simultáneo.

La especificación de este modelo SAR(1) expresa la ausencia de aleatoriedad espacial en una variable (y) de una manera muy básica, sin considerar ningún tipo de relación causal, en la que otras variables exógenas (X) puedan ser explicativas de esta peculiar distribución de la variable endógena (y), como se verá en otro tipo de modelos

➤ Modelo básico de regresión lineal (MBRL)

Existe la posibilidad de que no se produzca la dependencia espacial en una variable o de que ésta sea recogida correcta y totalmente por un grupo de regresores, a través de una relación estructural, como en el caso del MBRL:

$$\begin{aligned} y &= X\beta + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.2.}$$

siendo X: una matriz (K,N) de K variables exógenas y N observaciones
 β: vector (K,1) de parámetros de las variables exógenas.

El MBRL será la especificación correcta para una variable que presenta autocorrelación espacial, sólo en el caso de que este efecto espacial esté totalmente explicado por los valores de una o más variables explicativas, es decir, por condicionantes internos referidos a dicho lugar i. Por tanto, la inclusión en el modelo de un número (K) de

variables explicativas produce una ausencia de significatividad (a contrastar estadísticamente) en la relación espacial existente entre y_i , y_j , de forma que $Cov(y_i, y_j) = 0$. Esta situación implica también la ausencia de autocorrelación espacial residual, siempre y cuando se trate de un modelo en el que se hayan especificado correctamente los determinantes estructurales de la variable endógena.

Por eso, en la Figura IV.1.1., se representa la estructura explicativa del modelo a través de flechas que, para cada valor de la variable endógena (y), parten de sus correspondientes valores de las variables explicativas (X). En el Anexo I, se incluye una experiencia de Ceccato *et al.* (2001), en la que los autores especifican un MBRL como adecuado para explicar correctamente una variable con clara estructura de dependencia espacial, como la delincuencia en los barrios de una ciudad sueca, a partir de un conjunto de variables explicativas estructurales.

Sin embargo, en la mayor parte de las ocasiones, el MBRL es insuficiente para explicar la estructura espacial de la variable endógena. En estos casos, la estimación del MBRL por el método de mínimos cuadrados ordinarios (MCO) produce un efecto de dependencia espacial, estadísticamente significativo, en la perturbación aleatoria, debido a la mala especificación del mismo. En estos casos, se dice que el modelo tiene problemas de autocorrelación espacial, los cuales deben abordarse a través de diversas especificaciones, siendo las más conocidas las correspondientes a la dependencia espacial residual (modelo del error espacial) y dependencia espacial sustantiva (modelo del retardo espacial).

➤ **Modelo de regresión con dependencia espacial en la perturbación aleatoria o modelo del error espacial**

El modelo del error espacial (“spatial error model”) es, con mucho, la especificación más utilizada en los casos en que el MBRL resulta ineficaz como explicativo de un fenómeno con autocorrelación espacial. La existencia de ciertos factores o variables no explícitamente considerados en el modelo trasladan hacia los términos del error la configuración de agrupación de valores (autocorrelación) presente en la variable endógena. Por eso, en la Figura IV.1.1., las relaciones simultáneas de dependencia espacial existentes entre valores de la variable endógena (y), representadas a través de una flecha de doble dirección, se trasladan de esta variable hacia el término de perturbación aleatoria. Es decir, el efecto de dependencia espacial en la variable endógena (y) es explicado, no sólo por las variables independientes presentes en el modelo, sino por otras que se encuentran ausentes (dependencia espacial residual). En el Anexo I, se presentan bastantes ejemplos de este tipo de modelos (por ejemplo, Anselin, 1988A; Haining, 1994, 1995; Moreno y Vayá, 2000; Baller *et al.*, 2001; Ceccato *et al.*, 2001; Sandberg y Johanson, 2001; Thomas, 2001).

Habitualmente, la distribución de la perturbación aleatoria suele especificarse como un proceso autorregresivo de orden 1 o AR(1), como en el caso siguiente:

$$\begin{aligned}
 y &= X\beta + u \\
 u &= \lambda Wu + \varepsilon \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.3}$$

donde u: perturbación aleatoria distribuida según un proceso AR(1)
 λ : parámetro autorregresivo (escalar) asociado al retardo espacial Wu
 ε : vector de perturbaciones aleatorias, ruido blanco.

Esta expresión puede transformarse sustituyendo la expresión autorregresiva de la perturbación aleatoria en el modelo, de forma que el valor que adopta la variable endógena (y) en un punto (i) depende de los valores de la perturbación aleatoria en todos los puntos del sistema, a través del multiplicador espacial $(1 - \lambda W)^{-1}$:

$$\begin{aligned}
 y &= X\beta + (1 - \lambda W)^{-1} \varepsilon \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.4}$$

Algunos autores (Cliff y Ord, 1981; Sneek y Rietveld, 1997; Mur, 1999; Anselin, 2001) proponen también la especificación de procesos espaciales de medias móviles para el término de la perturbación aleatoria, demostrando que una especificación incorrecta de la perturbación aleatoria como un proceso AR(1), en lugar de MA(1), produce estimadores sesgados³⁰. En este caso, el modelo del error espacial tendría la siguiente forma:

$$\begin{aligned}
 y &= X\beta + u \\
 u &= \theta W\varepsilon + \varepsilon \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.5}$$

siendo θ el parámetro (escalar) de medias móviles asociado al retardo espacial $W\varepsilon$.

➤ **Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial**

El modelo del retardo espacial (“spatial lag model”), resulta también adecuado para aquellos casos en los que el MBRL resulta insuficiente como explicativo del fenómeno de dependencia espacial presente en la variable endógena (y). A diferencia de los casos anteriores, el modelo del retardo espacial incorpora la influencia de las variables omitidas a través de una variable dependiente espacialmente retardada, es decir, a través de los valores que, para cada punto i, adopta la variable endógena en un grupo de localizaciones vecinas, de la manera siguiente:

³⁰ Sneek y Rietveld (1997) atribuyen la impopularidad del modelo MA(1), frente a la especificación AR(1), a la mayor complejidad que presenta la estimación de estos modelos, sólo posible bajo ciertas restricciones. En el Apartado siguiente, se expone, de forma resumida, la nueva taxonomía de modelos de regresión espacial propuesta por Anselin (2001), en la que este autor presenta la especificación de medias móviles como un fenómeno de autocorrelación “local” frente a la especificación autorregresiva, considerada como “global”.

$$\begin{aligned}
 y &= \rho Wy + X\beta + u \\
 u &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.6}$$

Al igual que en el modelo del error espacial, esta expresión puede también adoptar una forma reducida que incluye la especificación del modelo del error espacial (aunque en una forma no anidada), de modo que el valor de la variable endógena (y) en un punto (i) se encuentra afectado no sólo por el valor de las variables exógenas en dicho punto (i), sino también en el resto de localizaciones, a través del multiplicador espacial $(1 - \lambda W)^{-1}$, tal como se expresa en la Figura IV.1.1. con las flechas dobles que parten de los valores de la variable exógena en los puntos (i,j) hacia los valores de y_i, y_j .

$$\begin{aligned}
 y &= (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} u \\
 u &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.7}$$

Existen en la literatura varios contrastes que se han definido para obtener una correcta identificación en situaciones de autocorrelación espacial de la variable dependiente no totalmente explicada por la relación estructural del MBRL. En el caso de revelarse el modelo del retardo espacial como el más adecuado, se estaría poniendo de manifiesto un proceso de difusión espacial en la variable analizada, de forma que unos valores altos/bajos de la variable endógena (y) en un lugar (i), realmente estarían incrementando la probabilidad de ocurrencia de valores altos/bajos de la misma en lugares vecinos (adquiriendo así un cierto carácter de predicción temporal). Éste sería, por ejemplo, el caso expuesto en Baller *et al.* (2001) para la tasa de homicidios cometidos en los estados del Sur de los EEUU (ver Anexo I).

IV.1.1.2. Taxonomía de Florax y Folmer

A partir de los citados modelos del retardo espacial y del error espacial (con perturbaciones aleatorias autorregresivas de primer orden), es posible especificar otros modelos que incluyan, por ejemplo, una o más variables exógenas espacialmente retardadas o combinaciones de los modelos anteriores (ver Figura IV.1.2.), tal como pusieron de manifiesto Florax y Folmer (1992). Todos estos modelos pueden expresarse de forma más general en el llamado **modelo mixto regresivo de regresión espacial, con perturbaciones aleatorias autorregresivas y heteroscedásticas**, al que, de forma más breve, nos referiremos como **modelo general de regresión espacial**:

$$\begin{aligned}
 y &= \rho W_1 y + X\beta_1 + W_2 R\beta_2 + u \\
 u &= \lambda W_3 u + \varepsilon \\
 \varepsilon &\approx N(0, \Omega) ; \Omega_{ii} = h_i(Z\alpha) ; h_i > 0
 \end{aligned}
 \tag{Eq. IV.1.8}$$

donde y: vector (N,1) de observaciones de la variable endógena
 W_1 : matriz de pesos espaciales de la variable endógena, siendo W_1 y el retardo

- espacial de la variable endógena
- ρ : coeficiente autorregresivo espacial (escalar), que recoge la intensidad de las interdependencias entre las observaciones muestrales
- X : matriz (K_1, N) de K_1 variables exógenas y N observaciones
- R : matriz (K_2, N) de K_2 variables exógenas espacialmente retardadas, que pueden o no coincidir con las variables incluidas en X .
- W_2 : matriz de pesos espaciales correspondiente a las variables exógenas espacialmente retardadas, siendo W_2R el retardo espacial de dichas exógenas
- β_1, β_2 : vectores ($K_1, 1$) y ($K_2, 1$) de parámetros de las variables exógenas.
- u : vector ($N, 1$) de perturbaciones aleatorias autorregresivas (de primer orden) y heteroscedásticas, siendo los elementos de la diagonal principal de la matriz de covarianzas (Ω) función de $P+1$ variables exógenas de Z
- α : vector ($P, 1$) asociado a los términos no constantes de la matriz Z
- W_3 : matriz de pesos espaciales de la perturbación aleatoria u
- λ : parámetro autorregresivo (escalar) asociado al retardo espacial W_3u
- ε : vector de perturbaciones aleatorias, ruido blanco.

La especificación del modelo general de regresión espacial, como puede observarse, considera también la existencia del fenómeno de heterogeneidad espacial, en forma de heteroscedasticidad en la perturbación aleatoria, concepto que será ampliado en el Apartado IV.1.2. La consideración o no, en este modelo general, de la posibilidad de existencia de alguna forma de dependencia espacial (por ahora, se excluye el análisis de heterogeneidad espacial), como se aprecia en la Figura IV.1.1., permite derivar del mismo otras especificaciones particulares, a través de la imposición de restricciones en el siguiente vector de parámetros:

$$\Theta = [\rho, \beta_1, \beta_2, \theta, \lambda, \alpha] \quad \text{Eq. IV.1.9.}$$

➤ **Modelo básico de regresión lineal: $\rho = \beta_2 = \theta = \lambda = \alpha = 0$.**

Cuando en un modelo de regresión con datos geográficamente distribuidos no se produce ningún efecto espacial de dependencia (ni heterogeneidad) en la perturbación aleatoria, la especificación más adecuada, como ya se ha visto, coincide con el modelo básico de regresión lineal (MBRL):

$$\begin{aligned} y &= X\beta_1 + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.10.}$$

Algunos autores aconsejan anular el efecto de dependencia espacial, detectado a través de las técnicas del AEDE en una o varias variables del modelo (endógena y/o exógenas), a través de métodos de filtrado espacial, de forma que un modelo con problemas de dependencia espacial pueda ser transformado en el MBRL, mediante el filtrado espacial de las variables afectadas. En concreto, Getis (1995, 2002) propone una transformación de las variables con problemas de dependencia espacial a partir del estadístico $G_i(d)$

presentado en el Apartado III.2.3.2., de forma que se obtenga, por ejemplo, para la variable endógena (y) espacialmente autocorrelacionada, otra variable (y^*) sin tendencia espacial, del modo siguiente:

$$y_i^* = \frac{y_i \left(\frac{W_i}{N-1} \right)}{G_i(d)} \quad \text{Eq. IV.1.11.}$$

donde $W_i = \sum_j w_{ij}(d)$, por lo que el cociente $W_i/N-1$ es la media o esperanza matemática del estadístico G_i . La diferencia $y_i - y_i^*$ sería una nueva variable que contiene los efectos espaciales presentes en y_i . Esta transformación, aplicada a las variables afectadas, daría lugar a nuevas variables (y^* , x_k^*) sin problemas de autocorrelación espacial que podrían formar parte de la especificación de un MBRL (ver los modelos de disparidades regionales y gasto público propuestos por este autor en el Anexo I).

Esta transformación constituye también una solución a situaciones de predicción-extrapolación de datos espaciales en los que se necesite extrapolar un modelo del retardo espacial a un ámbito microterritorial, para el que no se dispone, obviamente, de la variable endógena, ni por tanto, de su correspondiente retardo espacial.

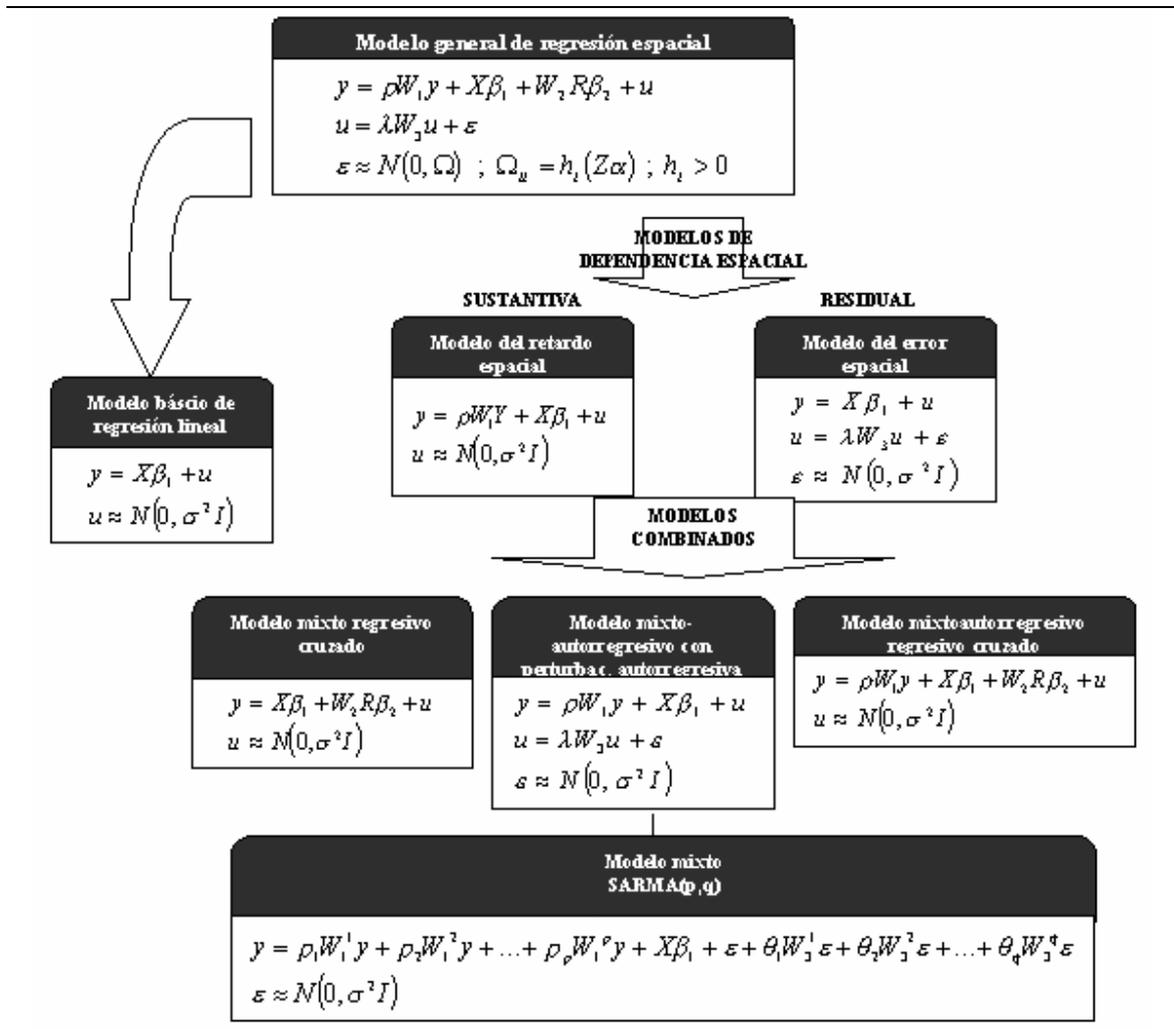
➤ **Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial: $\beta_2 = \theta = \lambda = \alpha = 0$.**

Este conocido modelo es un caso de dependencia espacial sustantiva en el que la variable endógena presenta autocorrelación espacial:

$$\begin{aligned} y &= \rho W_1 y + X\beta_1 + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.12.}$$

Como puede observarse, la inclusión de la variable endógena espacialmente retardada, como explicativa en el modelo, es semejante a la introducción de variables desfasadas en el modelo temporal, aunque con diferencias importantes (ver Kelejian y Robinson, 1995; Anselin y Bera, 1998). Por ejemplo, mientras que en el caso temporal la variable desplazada, y_{t-1} , está incorrelacionada con la perturbación aleatoria, u_t (siempre que, a su vez, no exista autocorrelación temporal en los errores), la variable espacialmente retardada $W_1 y_i$, siempre estará correlacionada con u_i (con independencia del esquema de autocorrelación de la perturbación aleatoria), así como con los valores del término de error en otras localizaciones del espacio. Por eso, a diferencia de lo que sucede en el ámbito temporal, el estimador mínimo-cuadrático no será consistente en este tipo de modelos (Anselin, 1988A), como se expondrá más adelante.

Figura IV.1.2. Segunda taxonomía de modelos de dependencia espacial



Fuente: Elaboración propia.

Esta especificación, como ya se ha expuesto, es la adecuada en casos en los que el valor que adopta una variable en un punto o región depende no sólo de condicionantes internos (valores de otras variables explicativas en el mismo punto o región), sino también del valor de esa misma variable en otras regiones vecinas, incumpléndose así el principio de independencia entre las observaciones muestrales (Moreno y Vayá, 2000). Esto es lo que Paelinck y Klaasen (1979) denominan como principio de alotopía que establece que, en muchas ocasiones, las causas de fenómenos acaecidos en una determinada localización deberían buscarse en otro lugar del espacio. Anselin y Bera (1998) también señalan la necesidad de utilizar esta especificación como elemento de corrección en casos en los que se producen errores de escala, como el presentado en el Apartado III.2.1. para unidades espaciales contiguas.

Debe advertirse también que la inclusión del retardo espacial Wy evita los problemas de identificación y estimación que se producirían al tratar de recoger, con insuficientes grados de libertad, la totalidad de interacciones existentes entre cada región i y sus vecinas.

➤ **Modelo mixto regresivo cruzado de regresión espacial: $\rho = \theta = \lambda = \alpha = 0$**

En este modelo, el efecto de dependencia espacial es también sustantivo, dado que se encuentra presente, en forma de retardo espacial, en una o varias variables exógenas del modelo (no en la dependiente, como en el modelo del retardo espacial):

$$\begin{aligned} y &= X\beta_1 + W_2R\beta_2 + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.13.}$$

Esta especificación sería adecuada, por ejemplo, en una función de producción en la que el nivel de producción de una región i viniera explicada por la disponibilidad del factor trabajo en la propia región (X) y en un conjunto de regiones vecinas a i (W_2R). Las variables exógenas desplazadas pueden ser tratadas de la misma manera que las demás variables exógenas en lo que respecta a los procesos de estimación y contraste, aunque suelen ser fuente de multicolinealidad en el modelo.

➤ **Modelo mixto autorregresivo regresivo cruzado de regresión espacial: $\theta = \lambda = \alpha = 0$.**

Esta especificación es producto de la combinación del modelo del retardo espacial y del modelo mixto regresivo cruzado de regresión espacial, del modo siguiente:

$$\begin{aligned} y &= \rho W_1 y + X\beta_1 + W_2R\beta_2 + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.14.}$$

Siguiendo con el ejemplo mencionado de la función de producción, en este caso, el nivel de producción en una región i vendría explicado, no sólo por la disponibilidad del factor trabajo en dicha región i (X) y en otras relacionadas con la misma (W_2R), sino también por el propio nivel de producción existente en regiones vecinas (variable endógena espacialmente retardada, W_1y). En el Anexo I, Bloomstein (1985), Pace y Barry (1997) y Van der Kruk (2001) presentan modelos de este tipo.

Un caso particular sería el llamado **modelo Durbin espacial**, análogo al propuesto para los modelos de series temporales, en el que los coeficientes de las variables exógenas espacialmente retardadas (β_2) tienen que ser iguales al valor negativo del producto del coeficiente de la variable endógena retardada (ρ) por los coeficientes correspondientes a las variables exógenas (β_1). La expresión matemática de esta restricción es lo que se conoce como hipótesis del factor común ($H_0: \beta_2 = \rho \cdot \beta_1$).

➤ **Modelo de regresión con dependencia espacial en la perturbación aleatoria o modelo del error espacial: $\rho = \beta_2 = \theta = \alpha = 0$.**

En este tipo de modelos, el fenómeno de autocorrelación espacial está únicamente presente en la perturbación aleatoria normalmente, como ya se ha indicado, en una forma autorregresiva de orden 1, SAR(1).

$$\begin{aligned} y &= X\beta_1 + u \\ u &= \lambda W_3 u + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.15.}$$

La consideración explícita de un esquema de dependencia espacial en el término de la perturbación aleatoria resulta adecuada cuando se omiten en un modelo variables que se hallen correlacionadas espacialmente³¹, o también cuando se producen errores de medida. La no esfericidad de las perturbaciones aleatorias descarta el método de estimación mínimo-cuadrático pues da lugar a coeficientes sesgados e ineficientes.

De esta expresión podrían derivarse otras si, por ejemplo, se supone en el término de la perturbación aleatoria la existencia de estructuras autorregresivas de orden superior a uno, por ejemplo, un SAR(2):

$$\begin{aligned} y &= X\beta_1 + u \\ u &= \lambda_1 W_3^1 u + \lambda_2 W_3^2 u + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.16.}$$

donde W_3^1, W_3^2 son las matrices de interacción espacial, correspondientes a la perturbación aleatoria, de orden 1 y 2 respectivamente.

Además, como ya se ha expuesto, también podrían derivarse procesos de medias móviles en la perturbación aleatoria, SMA(q). Por ejemplo, en Cliff y Ord (1981) se sugiere un modelo sencillo de medias móviles en el término de error, SMA(1), que podría ser también generalizable a órdenes superiores:

$$\begin{aligned} y &= X\beta_1 + u \\ u &= \theta W_3 \varepsilon + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.17.}$$

³¹ Obviamente, el término de la perturbación aleatoria también estará correlacionado espacialmente cuando se omita erróneamente un retardo espacial en la variable endógena o en una o varias exógenas, porque la dependencia espacial se trasladaría directamente al término de error. En estos casos de dependencia espacial sustantiva, la solución pasaría por inclusión en el modelo de la variable afectada espacialmente retardada.

donde θ es el coeficiente (escalar) de medias móviles espacial.

Una formulación alternativa del modelo del error espacial sería el llamado modelo de componentes del error espacialmente autorregresivo de orden 1 o **MCE-SAR(1)** de Bolduc *et al.* (1992, 1995), especialmente recomendable para los modelos de flujos de viaje (ver una aplicación en el Anexo I). Este método permite incluir las diversas causas de una mala especificación en un modelo a través de la descomposición de la perturbación aleatoria (u) como suma de tres elementos referidos a la localización del origen (α_i), destino (λ_j) y resto (ε_{ij}), de la forma siguiente:

$$\begin{aligned} y_{ij} &= \beta_0 + N_{ij}\beta_1 + S_i\beta_2 + S_j\beta_3 + u_{ij} \\ u_{ij} &= \alpha_i + \lambda_j + \varepsilon_{ij} \\ \alpha_i &= \rho_1 W\alpha_i + \xi_1 \quad ; \quad \xi_1 \approx N(0, \sigma_{\xi_1}^2) \\ \lambda_j &= \rho_2 W\lambda_j + \xi_2 \quad ; \quad \xi_2 \approx N(0, \sigma_{\xi_2}^2) \\ \varepsilon_{ij} &= \rho_3 W\varepsilon_{ij} + v_{ij} \quad ; \quad v \approx N(0, \sigma_v^2) \end{aligned} \quad \text{Eq. IV.1.18.}$$

donde y_{ij} : flujo entre dos regiones i, j , para $i = 1, \dots, N$ y $j = 1, \dots, T$ regiones del grupo de N regiones de origen y T regiones de destino, respectivamente.
 N_{ij} : variables relativas a redes o vías de comunicación entre i, j (tiempo de viaje, coste, frecuencia de desplazamientos, etc.)
 S_i : variables socioeconómicas relativas a las regiones origen
 S_j : variables socioeconómicas relativas a las regiones destino
 W : matriz de pesos espaciales en la que cada elemento es una función de la distancia existente entre las regiones i, j .

En este modelo, se supone que las perturbaciones aleatorias ξ_1, ξ_2, v_{ij} son esféricas y no están relacionadas unas con otras.

➤ **Modelo mixto autorregresivo de regresión espacial con perturbación aleatoria espacialmente autorregresiva: $\beta_2 = \theta = \alpha = 0$**

Este modelo surge de la combinación de los modelos del retardo espacial y del error espacial, de la manera siguiente:

$$\begin{aligned} y &= \rho W_1 y + X\beta_1 + u \\ u &= \lambda W_3 u + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.19.}$$

La consideración de dos matrices de pesos espaciales, W_1 y W_3 , para los procesos autorregresivos de la variable endógena y la perturbación, respectivamente, implica que ambos procesos pueden tener distinta estructura espacial. Aunque en la práctica puedan no existir importantes diferencias entre ambas matrices (Cerejeira, 1998), éstas no deberían

coincidir totalmente para evitar problemas de identificación en la estimación por el método máximo-verosímil (Anselin, 1988A). O, en caso de coincidir totalmente, los coeficientes ρ y λ únicamente estarían identificados si la matriz X contiene, al menos, una variable exógena (además del término constante) y se imponen las restricciones no lineales correspondientes (Moreno y Vayá, 2000).

➤ **Modelo mixto autorregresivo de regresión espacial, con perturbaciones que incorporan un esquema de dependencia de media móvil.**

Se trata de un modelo en el que el efecto de dependencia espacial se encuentra presente de forma autorregresiva en la variable dependiente y , siguiendo un esquema de medias móviles, en las perturbaciones aleatorias. El caso más sencillo sería el denominado modelo SARMA(1,1), que tendría la forma siguiente:

$$\left. \begin{aligned} y &= \rho W_1 y + X\beta_1 + u \\ u &= \theta W_3 \varepsilon + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \right\} \Rightarrow y = \rho W_1 y + X\beta_1 + \varepsilon + \theta W_3 \varepsilon \quad \text{Eq. IV.1.20.}$$

En este modelo, a diferencia de lo que sucede en el caso de una perturbación espacialmente autorregresiva, los parámetros escalares ρ y θ estarán siempre identificados, aún cuando $W_1 = W_3 = W$.

Si este esquema se extendiera a órdenes superiores, se obtendría el modelo SARMA(p,q) propuesto por Anselin y Bera (1998), que algunos autores denominan como modelos SARMAX, debido a la presencia en los mismos de variables exógenas:

$$\left. \begin{aligned} y &= \rho_1 W_1^1 y + \rho_2 W_1^2 y + \dots + \rho_p W_1^p y + X\beta_1 + \varepsilon + \theta_1 W_3^1 \varepsilon + \theta_2 W_3^2 \varepsilon + \dots + \theta_q W_3^q \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \right\} \quad \text{Eq. IV.1.21.}$$

Las matrices de pesos espaciales utilizadas, W_1 y W_3 , son matrices de interacciones espaciales, que llevan expresado como superíndice la potencia u orden de contigüidad. En estos casos, debe ponerse especial cuidado en evitar el problema de circularidad y redundancia propio de las matrices de potencias superiores a uno.

Otros autores, reservan la denominación SARMA exclusivamente para denominar la estructura de dependencia propia de la perturbación aleatoria que, como en el modelo SARMA(1,1), propuesto por Basu y Reinsel (1994), se distribuye a la vez como un proceso autorregresivo y de medias móviles, ambos de orden 1:

$$\left. \begin{aligned} y &= X\beta + u \\ u &= \lambda Wu + \varepsilon + \theta W\varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \right\} \quad \text{Eq. IV.1.22.}$$

Los autores demuestran que la estimación por MCG, así como el método MV restringido, son mejores que MCO y MV, respectivamente.

IV.1.1.3. Taxonomía de modelos de externalidades espaciales

Anselin (2001E) formula una nueva taxonomía para aquellos modelos de regresión que expresan externalidades (“spillovers”). Las externalidades espaciales constituyen un fenómeno emergente en las ciencias sociales como, por ejemplo, la economía, a través de los modelos de interacción espacial que introducen relaciones de interdependencia entre los actores de un sistema, y también la sociología o la política, con el renacimiento de los estudios de la Escuela de Chicago sobre procesos de vecindad.

La verificación de estas externalidades espaciales, la medida de su fuerza y recorrido exigen la especificación y estimación de modelos de econometría espacial. Sin embargo, aunque el concepto de dependencia espacial es fácil de captar, no resulta tan claro saber exactamente la forma en que ésta se produce y, por tanto, la especificación concreta del modelo (la identificación de los parámetros a estimar).

Para ello, el autor parte de una forma reducida de los modelos de regresión, en la que las externalidades (“spillovers”), se expresan como multiplicadores espaciales. Aunque los modelos espaciales de regresión lineal suelen expresarse en la forma estructural, se propugna, en este caso, la transformación de la misma en su correspondiente forma reducida, caracterizada porque, en la parte derecha de la ecuación, sólo se incluyen variables exógenas, la perturbación aleatoria o transformaciones espaciales de las mismas.

La gran innovación de este enfoque estriba en la diferenciación que se hace, en el contexto de los modelos de regresión lineal, entre dependencia global y local (dependiendo de que los procesos de dependencia espacial sean SAR o SMA, respectivamente)³², que dan lugar a diversas especificaciones, según que los retardos espaciales afecten a la variable endógena (Wy), variables explicativas (WX) y/o la perturbación aleatoria (Wu).

En un contexto práctico, es importante conocer hasta qué punto la influencia ejercida sobre el valor de una variable en una localización geográfica determinada depende de circunstancias referentes a las localizaciones inmediatamente adyacentes a la misma (dependencia local) o a todas las localizaciones del sistema espacial considerado (dependencia global).

³² Debe advertirse que el punto de vista de la econometría espacial difiere del propio de la estadística espacial en su comprensión de los procesos autorregresivos (SAR) y de medias móviles (SMA), considerados en una perspectiva de “simultaneidad” (la primera) frente al concepto de “condicionalidad” propio de la perspectiva geoestadística. En efecto, la estadística espacial habla de procesos CAR, en los que el valor de variable en una localización se encuentra “condicionado” a la presencia de valores (similares o diferentes) de la misma en localizaciones vecinas, siendo este enfoque muy interesante en contextos de predicción temporal. La econometría espacial, por su parte, entiende los procesos SAR como explicativos de una estructura espacial “completa” simultánea.

1. Modelos de dependencia espacial global

La dependencia espacial global tiene lugar en modelos de regresión lineal en cuya especificación se incluye algún proceso autorregresivo espacial (SAR) en la perturbación aleatoria y/o en alguna variable estructural (y , X). El adjetivo “global” se refiere a que, en la correspondiente matriz de covarianzas, intervienen todas las localizaciones del sistema. Como es bien sabido, el fenómeno de autocorrelación espacial en una variable (y) implica que la correspondiente matriz de covarianzas $Cov(y_i, y_j) \neq 0$. Es decir, un impacto producido en dicha variable (y), en una localización geográfica (i), supondrá la transmisión del mismo al resto de localizaciones a través de un multiplicador espacial. Este fenómeno de difusión espacial puede tener lugar en cualquier proceso autorregresivo presente en la perturbación aleatoria, la variable endógena y/o en cualquiera de las exógenas del modelo.

A continuación, se ilustra este concepto a través del conocido modelo de regresión espacial con perturbación aleatoria autorregresiva de orden 1, SAR(1), especificado en su forma estructural de la siguiente manera:

$$\begin{aligned} u &= \lambda Wu + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.23.}$$

Este modelo puede también transformarse fácilmente en su forma reducida:

$$u = [I - \lambda W]^{-1} \varepsilon \quad \text{Eq. IV.1.24.}$$

siendo $[I - \lambda W]^{-1}$ el multiplicador espacial asociado al proceso autorregresivo.

Por tanto, la matriz de varianzas y covarianzas de la perturbación aleatoria con problemas de dependencia espacial, $Cov(uu') = E(uu')$, se obtendría de la siguiente manera:

$$E(uu') = E\{[I - \lambda W]^{-1} \varepsilon \varepsilon' [I - \lambda W]^{-1}\} = \sigma^2 \{[I - \lambda W]^{-1} [I - \lambda W]^{-1}\} \quad \text{Eq. IV.1.25.}$$

La naturaleza “global” de la estructura de dependencia espacial expresada por esta matriz de covarianzas, se observa mejor a través de una transformación matemática de la misma que, gracias a la llamada “expansión de Leontieff”, permite obtener, de forma desarrollada, la expresión de la matriz inversa $[I - \lambda W]^{-1}$ (sabiendo que, en la mayoría de los casos, tanto los elementos de la matriz de pesos espaciales estandarizada por filas como $|\lambda| < 1$):

$$[I - \lambda W]^{-1} = 1 + \lambda W + \lambda^2 W^2 + \dots \quad \text{Eq. IV.1.26.}$$

Por tanto, la expresión completa de la matriz no escalar de varianzas y covarianzas de la perturbación aleatoria se obtiene a partir del producto de la expresión anterior y su

traspuesta (obtenida en la forma habitual, mediante la trasposición de cada una de las matrices), dando lugar a un sumatorio de productos de potencias sucesivas de la matriz W de pesos espaciales, escaladas por potencias del vector autorregresivo λ :

$$Cov(uu') = I + \lambda W + \lambda W' + \lambda^2 (W^2 + WW' + W^2') + \dots \quad Eq. IV.1.27.$$

Es decir, para el caso de una matriz W de contigüidad de primer orden, cada potencia de la misma (W, W², W³, etc.) implica un orden mayor de contigüidad, dando lugar, en torno a cada una de las localizaciones (i), a una expansión creciente en los límites de interrelación. Es más, las potencias también sucesivas de los parámetros autorregresivos (siendo $|\lambda| < 1$) producen una caída progresiva del valor de las covarianzas a medida que aumenta el orden de contigüidad, tal como apunta la citada ley de Tobler (“todo tiene que ver con todo, pero las cosas cercanas tienen más que ver entre sí que las lejanas”).

Como puede observarse, el carácter “global” del modelo SAR en la perturbación aleatoria consiste precisamente en esa relación que se produce, de hecho, entre “todas” las observaciones del sistema. Aunque en la práctica, para bajos valores de λ , la covarianza tiende a cero tras un número relativamente pequeño de potencias, en teoría al menos, la matriz de covarianzas de la perturbación aleatoria autorregresiva de primer orden (u) es una matriz completa no nula.

A partir de esta exposición, se presentan distintos modelos de dependencia espacial global, según que las externalidades globales se produzcan en la perturbación aleatoria, las variables explicativas o en ambos tipos de variables conjuntamente.

➤ **Modelos de externalidades globales no modelizadas**

Se trata de modelos en los que se han omitido una o más variables con dependencia espacial, de forma que este efecto se traslada al término de la perturbación aleatoria, que se distribuye como un proceso autorregresivo (SAR). Éste sería el caso, por ejemplo, del efecto que produce la calidad del aire en un lugar (o en lugares cercanos a uno dado), en un modelo de precios hedónicos de la vivienda, cuando no se dispone de medidas precisas que midan esta variable. Este efecto global de dependencia espacial, omitido del modelo, está presente en la perturbación aleatoria.

Los efectos globales, especificados como un proceso SAR en el término de error, podrían expresarse en forma reducida, a partir de la forma estructural más conocida, del modo siguiente:

$$\left. \begin{array}{l} y = X\beta + u \\ u = \lambda Wu + \varepsilon \end{array} \right\} \Rightarrow y = X\beta + (\lambda Wu + \varepsilon) \Rightarrow \begin{cases} y = X\beta + [I - \lambda W]^{-1} \varepsilon \\ \varepsilon \approx N(0, \sigma^2 I) \end{cases} \quad Eq. IV.1.28.$$

A su vez, a partir de la forma reducida anterior, es posible, a su vez, derivar una nueva forma estructural, que incluya las variables endógena y exógenas espacialmente retardadas, que resulta interesante por coincidir con el llamado modelo Durbin espacial:

$$\begin{aligned} [I - \lambda W](y - X\beta) &= [I - \lambda W][I - \lambda W]^{-1} \varepsilon \\ y &= \lambda Wy + X\beta - \lambda WX\beta + \varepsilon \end{aligned} \quad \text{Eq. IV.1.29.}$$

Como ya se ha indicado, el modelo Durbin espacial contiene K restricciones no lineales, denominadas como restricción del factor común, según la cual:

$$\hat{\lambda} * \hat{\beta} = -\lambda * \hat{\beta} \quad \text{Eq. IV.1.30.}$$

➤ **Modelos de externalidades globales modelizadas**

Cuando las externalidades espaciales globales se consideran explícitamente en el modelo, se producen especificaciones más novedosas. Éste sería el caso de variables exógenas que afectan a la endógena a través de un multiplicador espacial global, de forma que el valor y_i se encuentre afectado, no sólo por el valor de X_i , sino también por el valor de X_j . Por ejemplo, las tasas impositivas en algunos estados o regiones suelen establecerse en función de la renta disponible local, así como de la renta de estados o regiones vecinos, y de los vecinos de los vecinos, etc. (dimensión global). Por ahora, no se considerará la posibilidad de existencia también de externalidades excluidas del modelo.

La introducción en la especificación del modelo de un multiplicador espacial global (proceso autorregresivo de primer orden), afectando a la matriz de regresores, produce la siguiente forma reducida:

$$\begin{aligned} y &= [I - \rho W]^{-1} X\beta + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.31.}$$

De esta expresión puede derivarse la forma estructural, multiplicando en ambos extremos de la igualdad por $[I - \rho W]$:

$$\begin{aligned} [I - \rho W]y &= X\beta + [I - \rho W]u \\ y &= \rho Wy + X\beta + u - \rho Wu \end{aligned} \quad \text{Eq. IV.1.32.}$$

Como puede observarse, este modelo tiene dos variables espacialmente retardadas, la variable dependiente y la perturbación aleatoria, dando lugar a un modelo SARMA, con una única restricción paramétrica según la cual el coeficiente autorregresivo de la variable endógena espacialmente retardada es el número opuesto al coeficiente de medias móviles correspondiente a la perturbación aleatoria.

➤ **Modelos de externalidades globales modelizadas y no modelizadas**

Cuando no existen, a priori, razones teóricas que circunscriban las externalidades globales al término de error o a las variables explicativas, el procedimiento a seguir, en estos casos, sería la inclusión en el modelo de ambos tipos de externalidades. Esto es lo que se produce, por ejemplo, en la forma reducida del modelo del retardo espacial (modelo mixto autorregresivo de regresión espacial), aunque un aspecto interesante de esta especificación, que suele ignorarse, es que conlleva restricciones adicionales. En concreto, la obtención de la forma estructural del modelo del retardo espacial requiere de una condición en el parámetro autorregresivo ρ , así como en la matriz de pesos espaciales: que sean los mismos en la expresión del multiplicador espacial del término del error y los regresores. Efectivamente, la forma reducida del modelo del retardo espacial sólo puede expresarse a través de las citadas restricciones:

$$y = [I - \rho W]^{-1} X\beta + [I - \rho W]^{-1} u \quad \text{Eq. IV.1.33.}$$

para que sea posible, tras las transformaciones pertinentes, la obtención de la mas conocida forma estructural del mismo:

$$y = \rho Wy + X\beta + u \quad \text{Eq. IV.1.34.}$$

Sin embargo, si se prescinde de las restricciones, puede obtenerse la forma reducida de un modelo más general, con diferentes parámetros y matrices de pesos espaciales para el efecto “spillover” (externalidades) asociado a las matrices X y u , de la siguiente manera:

$$y = [I - \rho W_1]^{-1} X\beta + [I - \lambda W_2]^{-1} u \quad \text{Eq. IV.1.35.}$$

Por su parte, la correspondiente forma estructural asociada puede derivarse pre-multiplicando, a ambos lados de la igualdad, por los multiplicadores $[I - \rho W_1][I - \lambda W_2]$:

$$[I - \rho W_1][I - \lambda W_2]y = [I - \lambda W_2]X\beta + [I - \rho W_1]u \quad \text{Eq. IV.1.36.}$$

y, tras algunos cálculos:

$$y = \rho W_1 y + \lambda W_2 y - \rho \lambda W_1 W_2 y + X\beta - \lambda W_2 X\beta + u - \rho W_1 u \quad \text{Eq. IV.1.37.}$$

Como puede observarse, se trata de un modelo SARMA, de orden superior a 1, con varias restricciones en los parámetros. Esta forma estructural general podría simplificarse, por ejemplo, exigiendo una única expresión para la matriz de pesos espaciales:

$$y = (\rho + \lambda)Wy - \rho \lambda W^2 y + X\beta - \lambda WX\beta + u - \rho Wu \quad \text{Eq. IV.1.38.}$$

Otra simplificación de este modelo consistiría en considerar que las matrices de pesos son ortogonales, es decir, que el producto $W_1W_2 = W^2 = 0$, en cuyo caso se obtendría un modelo SARMA(1,1) con un componente regresivo cruzado y restricciones adicionales en los parámetros:

$$y = (\rho + \lambda) Wy + X\beta - \lambda WX\beta + u - \rho Wu \tag{Eq. IV.1.39.}$$

A modo de conclusión, se presenta en la Tabla IV.1.1. un resumen de las principales especificaciones de los modelos de dependencia (externalidades) espacial global, según que se trate de efectos considerados (X) o no (u), o ambas cosas a la vez, explícitamente en un modelo de regresión lineal, tanto en la forma estructural como en la forma reducida (que incluye el multiplicador espacial autorregresivo SAR). En todos estos casos, los efectos espaciales producidos sobre una localización determinada (i) son debidos a circunstancias referidas, no sólo a las localizaciones inmediatamente adyacentes a la misma, sino a todo el sistema espacial considerado (dependencia global).

Tabla IV.1.1. Modelos de externalidades espaciales globales

	Forma estructural	Forma reducida
u	$y = \lambda Wy + X\beta - \lambda WX\beta + \varepsilon$	$y = X\beta + [I - \lambda W]^{-1} \varepsilon$
X	$y = \rho Wy + X\beta + u - \rho Wu$	$y = [I - \rho W]^{-1} X\beta + u$
Ambos	$y = \rho W_1y + \lambda W_2y - \rho \lambda W_1W_2y + X\beta - \lambda W_2X\beta + u - \rho W_1u$ $y = (\rho + \lambda) Wy - \rho \lambda W^2y + X\beta - \lambda WX\beta + \varepsilon - \rho W\varepsilon$ $y = \rho Wy + X\beta + u$	$y = [I - \rho W_1]^{-1} X\beta + [I - \lambda W_2]^{-1} u$ $y = [I - \rho W]^{-1} X\beta + [I - \rho W]^{-1} u$

Fuente: Elaboración propia.

2. Modelos de dependencia espacial local

La dependencia espacial local se produce en procesos espaciales de medias móviles (SMA), tanto en la perturbación aleatoria como en alguna variable estructural. El adjetivo “local” se refiere, en este caso, a que el efecto de difusión del multiplicador espacial derivado de estos procesos es mucho menor que en el caso global.

Este concepto se ilustrará a través de otro conocido modelo de regresión espacial, en este caso, con perturbación aleatoria distribuida como un proceso de medias móviles de orden 1, SMA(1):

$$\begin{aligned}
 u &= \varepsilon + \theta W\varepsilon \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.40.}$$

siendo θ el parámetro espacial de medias móviles. A diferencia del modelo SAR, en este caso, la forma reducida no incluye una matriz inversa, dado que, la expresión anterior es ya, de hecho, una forma reducida. Precisamente, esta ausencia de matrices invertidas da lugar a una matriz de covarianzas con una dimensión “local”:

$$\begin{aligned} Cov(uu') &= E\{(I + \theta W)\varepsilon\varepsilon'(I + \theta W)\} = \sigma^2\{(I + \theta W)(I + \theta W)'\} = \\ &= \sigma^2[I + \theta(W + W') + \theta^2WW'] \end{aligned} \quad Eq. IV.1.41.$$

Como puede observarse, los únicos elementos no nulos de esta matriz de varianzas y covarianzas son aquéllos no nulos también en la matriz W (o, de forma equivalente, W') y WW' . En el caso, por ejemplo, de W definida como una matriz de contigüidades de primer orden, los elementos no nulos serían los pares de localizaciones con relación de vecindad de primer (W, W') y segundo orden (WW'), pero no mayor. Es decir, más allá de estas dos “bandas” de vecindad, la covarianza espacial será cero. Ésta es la razón por la que se afirma que el recorrido del multiplicador espacial $(I + \theta W)$ de un proceso de medias móviles SMA es local, inferior al ámbito global del multiplicador SAR, $(I - \lambda W)^{-1}$.

➤ Modelos de externalidades locales no modelizadas

El efecto de dependencia espacial omitido en un modelo se traslada al término de la perturbación aleatoria que, en el caso local, se distribuye como un proceso espacial de medias móviles (SMA), en el que, como se ha demostrado, no será necesaria la solución de una forma reducida. Continuando con el ejemplo anteriormente expuesto de un modelo de precios hedónicos de la vivienda, en el que la calidad del aire en la práctica totalidad de localizaciones del sistema afectaba a todos y cada uno de los puntos del mismo (efecto global), es posible considerar también otras variables difíciles de cuantificar y, por tanto, no incluidas expresamente en el modelo, con un efecto más localizado en torno a cada punto del sistema, como ruidos, olores, etc., circunscritos a no más de dos órdenes de vecindad de cada observación.

La forma estructural de este tipo de modelos es el modelo del error espacial, en el que la perturbación aleatoria se distribuye como un proceso espacial de medias móviles:

$$\left. \begin{array}{l} y = X\beta + u \\ u = \varepsilon + \theta W\varepsilon \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} y = X\beta + \varepsilon + \theta W\varepsilon \\ \varepsilon \approx N(0, \sigma^2 I) \end{array} \right. \quad Eq. IV.1.42.$$

Esta forma estructural es, a la vez, la forma reducida:

$$y = X\beta + (I + \theta W)\varepsilon \quad Eq. IV.1.43.$$

siendo $[I + \theta W]$ el multiplicador espacial local.

➤ **Modelos de externalidades locales modelizadas**

Cuando las externalidades espaciales locales se consideran explícitamente en el modelo, a través de las variables exógenas, se produce el modelo mixto regresivo cruzado de regresión espacial, adecuado para situaciones en las que el ámbito de influencia de las variables explicativas sobre la endógena, se circunscribe no sólo a una localización dada, sino también a las inmediatamente vecinas a ella. Es decir, se trata de un modelo en el que una serie de variables explicativas se encuentra espacialmente retardada, WX , del modo siguiente:

$$y = X\beta_1 + WX\beta_2 + u \quad \text{Eq. IV.1.44.}$$

En esta expresión, β_2 es un vector $(K-1, 1)$ de regresores espacialmente retardados y no un escalar, como ρ , λ , θ . En este sentido, podría establecerse la restricción de considerar esta relación de dependencia en las variables explicativas como común a todas ellas, a través de un coeficiente común (ρ) que las afectara por igual, de la forma siguiente:

$$y = X\beta + \rho WX + u \quad \text{Eq. IV.1.45.}$$

Como puede observarse, en este caso, la forma reducida coincide con la forma estructural. Además, debe advertirse que no es posible considerar un caso de dependencia (externalidades) espacial local el modelo que incluye la variable endógena espacialmente retardada, ya que la presencia de dicho retardo en el extremo derecho de la ecuación (forma estructural) da lugar a una forma reducida que incluye siempre un multiplicador espacial global, tipo: $[I - \rho W]$.

➤ **Modelos de externalidades locales modelizadas y no modelizadas**

La especificación de externalidades espaciales locales, tanto en la matriz X como en u , es mucho más simple que en el caso global, que daba lugar a procesos SARMA de órdenes elevados. Ahora, el modelo contiene un término espacial regresivo cruzado (WX) así como un proceso de medias móviles (SMA) en la perturbación aleatoria que, en una formulación más general, puede considerar diferentes parámetros y/o matrices de pesos espaciales:

$$\left. \begin{array}{l} y = X\beta_1 + W_1 X\beta_2 + u \\ u = \varepsilon + \theta W_2 \varepsilon \end{array} \right\} \Rightarrow y = X\beta_1 + W_1 X\beta_2 + \varepsilon + \theta W_2 \varepsilon \quad \text{Eq. IV.1.46.}$$

Esta expresión general puede ser también simplificada introduciendo la restricción de considerar, en lugar del vector β_2 (diferente para cada variable exógena), un escalar ρ de efectos comunes a todas ellas (e igual a θ), así como una matriz W común, dando lugar al siguiente modelo en forma reducida:

$$y = X\beta + \rho WX + \varepsilon + \rho W\varepsilon \Rightarrow y = (I + \rho W)(X\beta + \varepsilon) \quad \text{Eq. IV.1.47.}$$

A modo de conclusión, se presenta en la Tabla IV.1.2. un resumen de las principales especificaciones de los modelos de dependencia (externalidades) espacial local, según que se trate de efectos considerados (X) o no (u), o ambas cosas a la vez, explícitamente en un modelo de regresión lineal, tanto en la forma estructural como en la forma reducida (que incluye el multiplicador espacial autorregresivo SMA). En todos estos casos, los efectos espaciales producidos sobre una localización determinada (i) son debidos a circunstancias referidas exclusivamente a las localizaciones inmediatamente adyacentes a la misma (efecto local).

Tabla IV.1.2. Modelos de externalidades espaciales locales

	Forma estructural	Forma reducida
u	$y = X\beta + \varepsilon + \theta W\varepsilon$	$y = X\beta + (I + \theta W)\varepsilon$
X		$y = X\beta_1 + WX\beta_2 + u$ $y = X\beta + \rho WX + u$
Ambos	$y = X\beta_1 + W_1X\beta_2 + \varepsilon + \theta W_2\varepsilon$ $y = X\beta + \rho WX + \varepsilon + \rho W\varepsilon$	$y = X\beta_1 + W_1X\beta_2 + (I + \theta W_2)\varepsilon$ $y = (I + \rho W)(X\beta + \varepsilon)$

Fuente: Anselin (2001E).

Además, el autor deja las puertas abiertas a nuevas especificaciones y análisis más complejos que superan estos esquemas. Por ejemplo, la especificación del ya presentado modelo mixto autorregresivo de regresión espacial con perturbación aleatoria espacialmente autorregresiva no se adapta al esquema anterior de externalidades globales-locales, dando lugar a una forma reducida más compleja. Efectivamente, partiendo de la forma estructural ya presentada:

$$\left. \begin{aligned} y &= \rho W_1 y + X\beta + u \\ u &= \lambda W_2 u + \varepsilon \Rightarrow u = (I - \lambda W)^{-1} \varepsilon \end{aligned} \right\} \Rightarrow \begin{cases} y = \rho W_1 y + X\beta + (I - \lambda W_2)^{-1} \varepsilon \\ \varepsilon \approx N(0, \sigma^2 I) \end{cases} \quad \text{Eq. IV.1.48.}$$

es posible derivar la siguiente forma reducida:

$$y = (I - \rho W_1)^{-1} X\beta + (I - \rho W_1)^{-1} (I - \lambda W_2)^{-1} \varepsilon \quad \text{Eq. IV.1.49.}$$

Aunque este modelo incluye el conocido multiplicador espacial global, propio de los procesos autorregresivos (SAR), la expresión obtenida para la dependencia espacial de la perturbación aleatoria es mucho más compleja y difícil de interpretar en un contexto

práctico que los modelos presentados en las Tablas IV.1.1. y IV.1.2., sobre todo, debido a la interacción de dos matrices de pesos espaciales diferentes, W_1 y W_2 .

Otra línea de investigación abierta sería la correspondiente a modelos híbridos que combinaran externalidades globales, por ejemplo, en la matriz X de regresores, con otras de tipo local, en la matriz de perturbaciones aleatorias, u , y viceversa. Ambos tipos de modelos dan lugar a formas estructurales que incluyen la variable dependiente espacialmente retardada, Wy , así como retardos espaciales, de orden superior a uno, tanto en X como en u , con restricciones paramétricas no lineales.

Así, en primer lugar, se considera un modelo con externalidades espaciales globales en la matriz X y locales en u , que tendría la siguiente forma reducida (suponiendo, para simplificar, que la matriz de pesos espaciales es la misma en todos los casos):

$$y = (I - \rho W)^{-1} X\beta + (I + \theta W)\varepsilon \quad \text{Eq. IV.1.50.}$$

Su correspondiente forma estructural sería la siguiente:

$$\begin{aligned} (I - \rho W)(y - \varepsilon - \theta W\varepsilon) &= X\beta \\ y &= \rho Wy + X\beta + \varepsilon + (\theta - \rho)W\varepsilon - \theta \rho W^2\varepsilon \end{aligned} \quad \text{Eq. IV.1.51.}$$

Por último, sería también posible considerar el caso contrario de externalidades locales en la matriz X y globales en u , con la siguiente forma reducida:

$$y = X\beta_1 + WX\beta_2 + (I - \lambda W)^{-1} \varepsilon \quad \text{Eq. IV.1.52.}$$

siendo su correspondiente forma estructural la siguiente:

$$\begin{aligned} (I - \lambda W)(y - X\beta_1 - WX\beta_2) &= \varepsilon \\ y &= \lambda Wy + X\beta_1 - \lambda WX\beta_1 + WX\beta_2 - \lambda W^2X\beta_2 + \varepsilon \end{aligned} \quad \text{Eq. IV.1.53.}$$

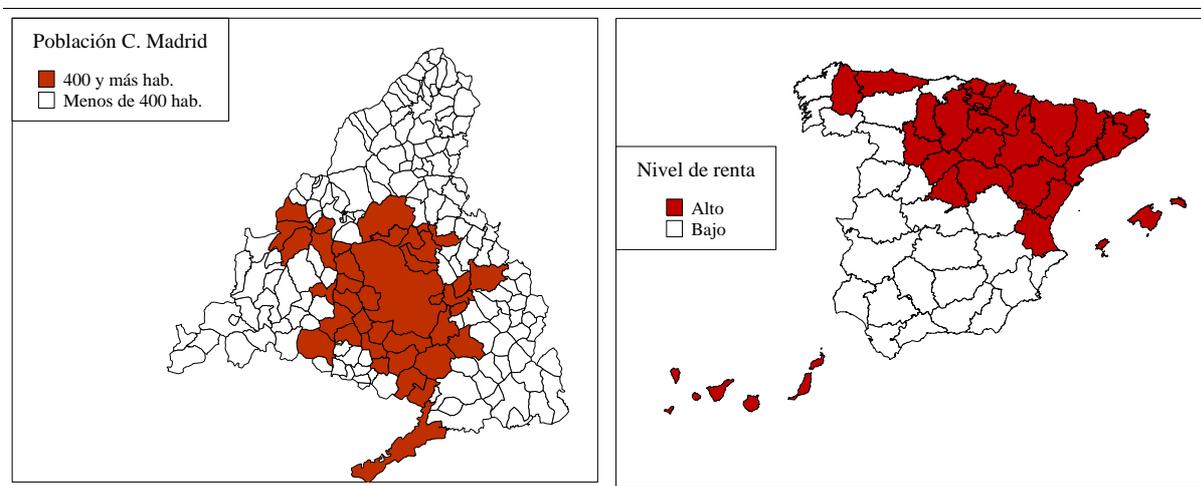
Queda para el futuro, el desarrollo de contrastes y métodos de estimación adecuados que permitan la correcta especificación de las externalidades globales y locales, así como el desarrollo de paquetes informáticos, asequibles a nivel de usuario, que implementen los modelos presentados en esta taxonomía.

IV.1.2. Modelos de heterogeneidad espacial

El efecto de heterogeneidad espacial, que fue presentado en el Apartado III.1.3 como objeto de estudio de los métodos gráficos del AEDE, se expone en este Apartado con una mayor profundidad conceptual, a través de las distintas especificaciones propias de los tipos de heterogeneidad espacial existentes.

La heterogeneidad espacial surge cuando se trabaja con unidades espaciales (países, regiones, municipios, secciones censales) en las que un fenómeno se distribuye de manera distinta sobre el espacio, lo que suele ocurrir con situaciones del tipo centro-periferia, norte-sur, este-oeste, etc., como podrían ser los casos que se presentan en la Figura III.3.1, así como en el análisis de predicción-extrapolación de datos espaciales, que se presenta en el Capítulo V. Por eso, este efecto espacial suele estar directamente relacionado con la localización geográfica, el área o cualquier otra característica de las unidades espaciales muestrales (Anselin, 1988A, Moreno y Vayá, 2000). Según Anselin (2001D)³³, la heterogeneidad espacial puede ser definida como “inestabilidad estructural en forma de varianza no constante de los residuos de una regresión (heteroscedasticidad) o en los coeficientes del modelo, que es posible abordar mediante técnicas de econometría tradicional o con herramientas propias de econometría espacial”.

Figura IV.1.3. Ejemplos de heterogeneidad espacial en situaciones centro-periferia (izquierda) y norte-Sur (derecha)



Fuente: Elaboración propia a partir del SIG MapInfo Professional.

Como el efecto de heterogeneidad espacial puede ser tratado utilizando técnicas econométricas tradicionales, los estudios sobre econometría espacial dedican una menor

³³ La consideración de la heterogeneidad espacial como un segundo efecto espacial, junto con la dependencia, fue inicialmente propuesta por Anselin (1988A).

atención a este tema que a la problemática en torno a la autocorrelación espacial. Sin embargo, como indica Anselin (2001D), hay tres razones por las que se debería analizar este efecto de heterogeneidad a través de técnicas propias de econometría espacial:

- En primer lugar, la estructura que subyace en la inestabilidad espacial es de carácter geográfico, en el sentido de que la localización de las observaciones es fundamental para determinar la forma o especificación de dicha variabilidad. Éste sería, por ejemplo, el caso de la heteroscedasticidad de grupos (“groupwise”), que podría ser modelizada a través de tantos valores de la varianza de la perturbación aleatoria como distintos grupos geográficos compactos puedan derivarse de los datos.
- En segundo lugar, dado que la estructura es espacial, la heterogeneidad suele producirse conjuntamente con el problema de autocorrelación espacial, no siendo ya adecuadas las herramientas de la econometría tradicional, dado que los contrastes habituales de heteroscedasticidad pueden estar sesgados en un contexto espacial.
- Por último, en tercer lugar, en un modelo de regresión de corte transversal, ambos efectos de autocorrelación y heterogeneidad espacial pueden ser, desde una óptica meramente observacional, totalmente equivalentes. Así, por ejemplo, un “cluster” o agrupamiento espacial (observado en localizaciones muy próximas) de los residuos con valores extremos podría ser interpretado como un problema de heterogeneidad espacial (heteroscedasticidad de grupos o “groupwise”), o también como un efecto de autocorrelación espacial. Por eso, deben estructurarse perfectamente ambos efectos espaciales para identificar correctamente los parámetros de un modelo con estos problemas y nunca considerar un aspecto independientemente del otro.

De la citada definición que Anselin (2001D), se deduce que en un modelo de regresión lineal pueden distinguirse diversas especificaciones para el efecto de heterogeneidad espacial, según que se manifieste como heteroscedasticidad o como inestabilidad estructural paramétrica.

IV.1.2.2. Especificación de la heteroscedasticidad espacial

La heteroscedasticidad consiste en la ausencia de estabilidad en la dispersión de un fenómeno, como sucede muchas veces con los residuos de una regresión y puede representarse como:

$$\text{Var}(u_i) = \sigma_i^2 \quad \text{Eq. IV.1.54.}$$

donde σ_i^2 : indica que la varianza de la perturbación aleatoria es diferente para cada observación muestral (i).

Las causas de la existencia de heterogeneidad espacial en un modelo de regresión serían las siguientes:

- Utilización de datos procedentes de unidades espaciales irregulares, es decir, con diferente área o extensión territorial, como es el caso de las divisiones político-administrativas (países, regiones, provincias, municipios, secciones censales,...).
- Tratamiento de unidades geográficas en las que un fenómeno se distribuye de manera desigual en el espacio, sobre todo, cuando se utilizan datos de regiones extremas (centro-periferia, norte-sur, este-oeste) o cuando se trabaja con datos referidos tanto a antiguas áreas metropolitanas como zonas de nuevo asentamiento, provincias urbanas y provincias rurales, secciones censales del centro de una ciudad y secciones del extrarradio, etc. (Anselin, 2001D).
- Causas de tipo sociológico, como la existencia de diversos gustos o actitudes de la población, o político, cuando en la zona analizada se producen diferentes administraciones o políticas regionales que llevan a respuestas diferentes ante un mismo estímulo (Moreno y Vayá, 2000).
- Además, a estas situaciones hay que añadir las causas habituales del problema de heteroscedasticidad en los modelos de regresión lineal: omisión de variables relevantes u otro tipo de especificación errónea del modelo, que producen en el término de la perturbación aleatoria una varianza no constante.
- Por último, cabe señalar que algunas causas que provocan la heterogeneidad espacial pueden también originar la aparición de autocorrelación espacial, (especificaciones erróneas o errores de medida, sobre todo), siendo necesaria la contrastación de ambos efectos conjuntamente (Anselin, 1988A).

Como ya se expuso en el Apartado IV.1.1., el modelo general de regresión espacial considera también la posibilidad de que la perturbación aleatoria presente problemas de heteroscedasticidad. Efectivamente, en la expresión formal del mismo:

$$\begin{aligned}
 y &= \rho W_1 y + X\beta_1 + W_2 R\beta_2 + u \\
 u &= \lambda W_3 u + \varepsilon \\
 \varepsilon &\approx N(0, \Omega) ; \Omega_{ii} = h_i(Z\alpha) ; h_i > 0
 \end{aligned}
 \tag{Eq. IV.1.55}$$

u es un vector $(N,1)$ de perturbaciones aleatorias autorregresivas y heteroscedásticas, siendo los elementos de la diagonal principal de la matriz de covarianzas (Ω) función de $P+1$ variables exógenas de Z y α , un vector $(P,1)$ asociado a los términos no constantes de la matriz Z .

Si se excluye de esta especificación todas las referencias a la dependencia espacial, se obtendrá el **modelo del error heteroscedástico**, que es un caso particular de un modelo de perturbaciones aleatorias no esféricas. En este modelo, la varianza de la perturbación aleatoria no es una constante, sino que varía con cada observación:

$$\begin{aligned}
 y &= X\beta + u \\
 \text{Var}(u_i) &= \sigma_i^2 I \\
 E[uu'] &= \Omega
 \end{aligned}
 \tag{Eq. IV.1.56.}$$

donde u es el vector de perturbaciones aleatorias $(N,1)$, con varianza no constante para cada observación i , y matriz diagonal de varianzas y covarianzas Ω .

Para que este modelo sea identificable, la varianza no constante debe tener alguna estructura que, en el modelo general, queda más o menos sin determinar y que podría especificarse de tres formas, como heteroscedasticidad aditiva, de coeficientes aleatorios y de grupos.

➤ Modelo de heteroscedasticidad aditiva

En muchas ocasiones se recurre a una especificación común denominada “aditiva”, en la que la varianza de la perturbación aleatoria se expresa como una función lineal de un conjunto de Z variables heteroscedásticas, que pueden coincidir o no con las variables explicativas, que son las que, a su vez, producen los problemas de heteroscedasticidad en el modelo, del modo siguiente:

$$\text{Var}(u) = Z \cdot \gamma \tag{Eq. IV.1.57.}$$

donde $\text{Var}(u)$: vector columna de varianzas de la perturbación aleatoria $(N,1)$
 Z : matriz (N,P) de variables heteroscedásticas (en columnas)
 γ : correspondiente vector de coeficientes.

Normalmente, la primera variable es un término constante, de forma que un test sencillo de heteroscedasticidad puede ser formulado como un contraste de significación conjunto del resto de coeficientes. La constante, en sí misma, sería la varianza homoscedástica de la perturbación aleatoria. Éste es el principio sobre el que se fundamenta el contraste de Breusch-Pagan para la heteroscedasticidad.

➤ Modelo de heteroscedasticidad de coeficientes aleatorios

En este modelo, la varianza de la perturbación aleatoria es función, en concreto, del cuadrado de Z variables explicativas del modelo (todas o parte de ellas), con problemas de heteroscedasticidad y, por tanto, causantes de este problema en el modelo:

$$\text{Var}(u) = Z^2 \cdot \gamma \tag{Eq. IV.1.58.}$$

siendo Z el grupo de variables explicativas heteroscedásticas.

➤ **Modelo de heteroscedasticidad de grupos (“groupwise”)**

En otras ocasiones, las Z variables heteroscedásticas, causantes de la no constancia de la varianza de la perturbación aleatoria en un modelo de regresión, se corresponden con variables categóricas, correspondientes a un determinado número de estructuras espaciales o regímenes (países ricos del norte, municipios de la periferia, etc.), que recogen un cambio estructural en la muestra de observaciones. Es decir, en cada estructura o régimen espacial, la varianza del error es distinta, aunque constante dentro de cada una de ellas, por lo que se podría llegar a concluir que la heteroscedasticidad detectada se halla causada por la existencia de diversos regímenes espaciales.

En esta situación, la varianza de la perturbación aleatoria podría ser estimada directamente a partir de los residuos de cada estructura espacial, siempre y cuando exista, en cada una de ellas, un número suficiente de observaciones. Las estructuras espaciales se construyen a partir de una variable indicador categórica que adopta valores enteros diferentes para cada estructura. A diferencia de las otras dos especificaciones, en este modelo no hay término constante.

IV.1.2.2. Modelos de inestabilidad paramétrica

La inestabilidad paramétrica espacial consiste en la falta de estabilidad sobre el espacio (variabilidad espacial) del comportamiento de una variable en estudio. Es decir, en este tipo de situaciones, tanto la forma funcional como los parámetros de una regresión pueden variar según la localización geográfica siendo, por tanto, no homogéneos en toda la muestra de datos. Es probable que este efecto surja, por ejemplo, cuando se utilizan datos de regiones extremas, como pueden ser las regiones ricas del norte europeo y las pobres del sur, para explicar, por ejemplo, un fenómeno económico en el que se produce un esquema norte-sur o centro-periferia. Este problema, según Moreno y Vayá (2000) podría formalizarse de la siguiente manera:

$$y_i = f_i(X_i\beta_i + \varepsilon_i) \quad \text{Eq. IV.1.59.}$$

siendo i : una observación, donde $i = 1, 2, \dots, N$ puntos en el espacio geográfico
 y_i : variable dependiente en la localización i
 X_i : vector $(1,K)$ de K variables explicativas
 β_i : vector $(K,1)$ de parámetros asociados a las variables explicativas
 ε_i : perturbación aleatoria
 $f_i(\cdot)$: función correspondiente a la observación i .

Es decir, de este modelo se deriva que existe una relación funcional concreta para explicar el valor de la variable endógena en cada localización (i), produciéndose el llamado problema de los parámetros incidentales (“incidental parameter problem”). Este problema consiste en la imposibilidad evidente de estimar el grupo de N vectores de parámetros β_i dada una muestra de N observaciones, ya que se carece de suficiente información en dicha matriz de datos con la que poder obtener una estimación para cada observación espacial.

Esta situación suele también denominarse “heterogeneidad espacial extrema” (Anselin, 1988A)³⁴.

Por eso, para poder proceder con un análisis de este tipo, será necesaria una especificación más simple de la variabilidad que sufre en el espacio un determinado fenómeno. Es decir, será necesario conocer de antemano la especificación propia de la variabilidad espacial del fenómeno a explicar y contrastar si dicha especificación es consistente con la información de los datos muestrales (LeSage, 1999). Además, hay que tener en cuenta que este fenómeno de inestabilidad estructural suele presentarse de dos formas, discreta o continua (Anselin, 1999B), según que los parámetros del modelo adopten valores diferentes en cada observación muestral o sólo en un número limitado de regímenes o estructuras espaciales, lo que originará modelos de diferente especificación.

1. Modelos con inestabilidad paramétrica continua

En algunas ocasiones, los parámetros de un modelo de regresión experimentan una deriva espacial continua. Es decir, los coeficientes asociados a las variables explicativas pueden adoptar un valor diferente, o bien para cada observación muestral, como en el modelo de parámetros aleatorios (Hildreth-Houck, 1968), o bien según unas variables de expansión (auxiliares), que pueden ser o no espaciales, como en el modelo de expansión espacial (Casetti, 1972, 1986, 1997A y B) o las regresiones ponderadas geográficamente, de Fotheringham *et al.* (1998) y Brudson *et al.* (1999).

➤ Modelo espacial de parámetros aleatorios, de Hildreth-Houck

El caso de deriva espacial continua de los parámetros, propuesta por Hildreth y Houck (1968), daría lugar a un modelo con la siguiente especificación:

$$y_i = X_i' \beta_i \quad \text{Eq. IV.1.60.}$$

donde X_i' : vector fila (1,K) de variables explicativas

β_i : vector columna (K,1) de coeficientes para cada observación (i).

En el modelo de parámetros aleatorios, en concreto, se desconoce la forma funcional de dicha variación paramétrica, por lo que el vector β_i se estima como la suma de dos elementos:

$$\beta_i = \beta + \mu_i \quad \text{Eq. IV.1.61.}$$

siendo β : constante

μ_i : perturbación aleatoria de distribución normal, con media nula y matriz de

³⁴ En el Capítulo V, se presenta el problema de la predicción-extrapolación de datos espaciales como un caso particular de la heterogeneidad extrema.

covarianzas que puede ser diagonal, en el caso particular de no existir correlación entre los términos de error y los coeficientes de regresión del modelo.

La sustitución del vector de coeficientes en el modelo general da lugar al siguiente modelo:

$$\begin{aligned}
 y_i &= X_i' \beta + u_i \\
 u_i &= X_i' \mu_i \begin{cases} E(u_i) = 0 \\ V(u_i) = X_i' \sum_i X_i \end{cases}
 \end{aligned}
 \tag{Eq. IV.1.62.}$$

Como puede observarse, el modelo de coeficientes aleatorios es, en realidad, un caso particular de la heteroscedasticidad de coeficientes aleatorios, en el que la varianza de las perturbaciones es función de las variables exógenas.

En este tipo de modelos, el efecto de dependencia espacial podría estar también presente, por ejemplo, en forma de proceso espacial autorregresivo, SAR(1), en la perturbación aleatoria (μ_i), definida inicialmente como ruido blanco, de forma que la perturbación general del modelo (u_i) presenta, a la vez, autocorrelación espacial y heteroscedasticidad (ver Anselin, 1988A). Es decir, la dependencia espacial estaría presente en la desviación que experimentan los parámetros (β_i) en torno a la media común (β), del modo siguiente:

$$\begin{aligned}
 \mu_i &= (\beta_i - \beta) = \lambda \sum_{j \neq i} w_{ij} (\beta_j - \beta) + \varepsilon_i \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{Eq. IV.1.63.}$$

O también:

$$\mu_i = \lambda \sum_{j \neq i} w_{ij} \mu_j + \varepsilon_i
 \tag{Eq. IV.1.64.}$$

correspondiéndose el subíndice j con las unidades “vecinas” a i , definidas por los elementos no nulos (w_{ij}) de la matriz W de pesos espaciales. Según esto, la perturbación aleatoria, μ , correspondiente al vector de parámetros cambiantes (β_i), se distribuye como un proceso espacial autorregresivo de orden uno, SAR(1), del modo siguiente:

$$\mu = \lambda W \mu + \varepsilon \quad \Rightarrow \quad \mu = (I - \lambda W)^{-1} \varepsilon
 \tag{Eq. IV.1.65.}$$

siendo la matriz inversa $(I - \lambda W)^{-1}$ que, para su más fácil notación se designará como a_{ij} , el multiplicador espacial asociado al término autorregresivo que expresa tanto la naturaleza conjunta (simultánea) de la dependencia espacial, como la existencia de una relación funcional entre ambas perturbaciones aleatorias que, en términos de Anselin (2001E), es de tipo global.

La expresión de la matriz de varianzas y covarianzas de la perturbación aleatoria general (u) será compleja y puede obtenerse a partir de la especificación de dicha perturbación, con problemas de heteroscedasticidad y autocorrelación espacial:

$$\begin{aligned} u_i &= X' \mu_i \\ u_i &= \sum_{j=1}^N a_{ij} \varepsilon_j X_i \end{aligned} \quad \text{Eq. IV.1.66.}$$

siendo $a_{ij} = (I - \lambda W)^{-1}$.

Por un lado, los términos de la varianza de la perturbación aleatoria general, $\text{Var}(u_i)$, serán función de la varianza constante de la perturbación aleatoria esférica (σ_ε^2), aunque verán agravado su problema de heteroscedasticidad, debido a la presencia del multiplicador espacial autorregresivo (a_{ij}). Por otro lado, los términos de la covarianza de la perturbación general, $\text{Cov}(u_i)$, son no nulos debido a la existencia de autocorrelación espacial entre los errores pertenecientes a unidades espaciales distintas.

Es decir, la estimación de un modelo espacial de parámetros aleatorios debe tener en cuenta correctamente estas circunstancias en la perturbación aleatoria pues, en caso contrario, los estimadores $\hat{\beta}$ podrían no tener buenas propiedades. Como se expone en el Apartado IV.2.3., el método de mínimos cuadrados ordinarios (MCO) no sería correcto, por lo que habría que recurrir a la estimación por máxima-verosimilitud (MV) o por mínimos cuadrados generalizados (MCG) iterativos.

Una especificación alternativa para la dependencia espacial sería la introducción de un retardo espacial de la variable endógena, como en el modelo del retardo espacial o modelo de externalidades globales modelizadas y no modelizadas (ver Apartado IV.1.1.).

$$\begin{aligned} y &= \rho W y + X \beta + u \\ y &= (I - \rho W)^{-1} X \beta + (I - \rho W)^{-1} u \end{aligned} \quad \text{Eq. IV.1.67.}$$

siendo ρ el coeficiente espacial autorregresivo correspondiente a la variable dependiente.

➤ **Modelo de expansión espacial, de Casetti**

En el caso del modelo de expansión espacial, originalmente propuesto por Casetti (1972), la deriva continua en los parámetros se formula en función de un conjunto de variables auxiliares, llamadas variables de expansión, que pueden ser de tipo no espacial (Jones y Casetti, 1992; Moreno *et al.*, 1997). Hace unos años, este paradigma de la expansión se ha extendido hasta llegar a ser un esquema general para el desarrollo de los

modelos (Casetti, 1986, 1997A y B; Casetti y Jones, 1988; Jones y Casetti, 1992)³⁵. Una formulación general, consistiría en considerar cada coeficiente de la regresión, β_j , como función lineal de un conjunto de variables de expansión, z_1, z_2, \dots, z_m , del modo siguiente:

$$\beta_j = \gamma_{0j} + \gamma_{1j}z_1 + \gamma_{2j}z_2 + \dots + \gamma_{mj}z_m \quad \text{Eq. IV.1.68.}$$

La sustitución de los coeficientes expandidos en el modelo general, produce un incremento de m nuevas variables en el mismo, consistentes en el producto de las variables explicativas por cada una de las variables de expansión: x_jz_1, x_jz_2 , etc. Cuando se demuestra que los coeficientes de un modelo dependen directamente del espacio geográfico, éstos podrían expresarse como:

$$\beta_j = \gamma_0 + \gamma_1m_j + \gamma_2p_j \quad \text{Eq. IV.1.69.}$$

siendo m : coordenada geográfica de longitud (tendencia este-oeste)³⁶
 p : coordenada geográfica de latitud (tendencia norte-sur)³⁷

La sustitución de los coeficientes en un modelo que, sin pérdida de generalidad, sólo contiene una variable explicativa daría lugar a la siguiente especificación:

$$\begin{aligned} y &= \beta_0 + \beta_1x + u \\ y &= \beta_0 + (\gamma_0 + \gamma_1m + \gamma_2p)x + u \\ y &= \beta_0 + \gamma_0x + \gamma_1(m \cdot x) + \gamma_2(p \cdot x) + u \end{aligned} \quad \text{Eq. IV.1.70.}$$

Este modelo considera que la perturbación aleatoria (u) es esférica. Sin embargo, esto puede no ser así en la práctica, debido a la dificultad de mantener la hipótesis de existencia de una relación exacta entre los coeficientes del modelo y las variables de expansión. De rechazarse esta hipótesis, sería necesario considerar un término aleatorio en la expansión de los coeficientes que, por ejemplo, en la anterior expresión simplificada, sería:

$$\begin{aligned} \beta_1 &= \gamma_0 + \gamma_1m + \gamma_2p + \mu \\ \mu &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.71.}$$

lo que daría lugar a un modelo diferente:

³⁵ En el Anexo I son interesantes los modelos de expansión espacial propuestos por Kristensen (1998), Sandberg y Johanson (2001).

³⁶ La coordenada de longitud es una medida de localización geográfica, que suele medirse en grados hacia el este/oeste del meridiano cero o de Greenwich. Sobre un mapa es posible dibujar las líneas de longitud como líneas verticales (meridianos), perpendiculares al ecuador y con intersección en ambos polos (no paralelas, dada la forma esférica de la Tierra), con un recorrido de $0,0^\circ$ (meridiano cero) a $+180,0^\circ$ y $-180,0^\circ$. Suele ser denominada coordenada X por determinar, en un eje horizontal, la localización este-oeste de un punto.

³⁷ La coordenada de latitud es una medida de localización geográfica, que suele medirse en grados sobre o bajo el ecuador. En un mapa es posible dibujar las líneas de latitud como líneas horizontales (paralelos) con un recorrido de $0,0^\circ$ (ecuador) a $+90,0^\circ$ (polo norte) y $-90,0^\circ$ (polo sur). Suele ser denominada coordenada Y por determinar, en un eje vertical, la localización norte-sur de un punto.

$$\begin{aligned}
 y &= \beta_0 + (\gamma_0 + \gamma_1 m + \gamma_2 p + \mu)x + u \\
 y &= \beta_0 + \gamma_0 x + \gamma_1 (m \cdot x) + \gamma_2 (p \cdot x) + \omega
 \end{aligned}
 \tag{Eq. IV.1.72.}$$

siendo $\omega = \mu \cdot x + u$, una perturbación aleatoria con problemas de heteroscedasticidad, dado que su correspondiente varianza es función de la variable explicativa (x):

$$\begin{aligned}
 \text{Var}(\omega) &= \sigma_\mu^2 \cdot x^2 + \sigma_u^2 \\
 u &\approx N(0, \sigma_u^2) \\
 \mu &\approx N(0, \sigma_\mu^2)
 \end{aligned}
 \tag{Eq. IV.1.73.}$$

suponiendo también independencia entre las perturbaciones aleatorias del modelo original (u) y del modelo de expansión de los coeficientes (μ), cosa que puede mantenerse en la práctica. Por último, también cabe indicar que las variables de expansión pueden ser cualquier conjunto de variables (espaciales y no espaciales), incluidas las expresiones polinomiales propias de la superficie tendencial, como se verá a continuación.

➤ Modelo de superficie tendencial

Un modelo de superficie tendencial es un modelo de regresión espacial cuyas variables explicativas son los elementos de un polinomio de las coordenadas terrestres de longitud (m) y latitud (p), a través de las cuales es posible conocer la localización exacta de una observación (i) en el espacio geográfico como el par (m_i, p_i). En el caso de que dicha observación (i) no fuera un punto sino un polígono, tal como se define en el Apartado II.1.2., este último podría representarse a través de su correspondiente centroide³⁸.

Por ejemplo, un modelo de superficie tendencial de 2º orden tendría la siguiente especificación:

$$y = \alpha + \beta_1 m + \beta_2 p + \beta_3 m^2 + \beta_4 p^2 + \beta_6 m \cdot p + u
 \tag{Eq. IV.1.74.}$$

Como ya se ha indicado, las variables explicativas (m,p) se corresponden con las coordenadas terrestres. Además, α es el término independiente, β_1 a β_6 son los coeficientes de la regresión en cada término del polinomio, siendo u, la perturbación aleatoria. Este modelo es similar a los métodos de ajuste de tendencias temporales, por lo que, al igual que en estos casos, resulta particularmente útil para eliminar tendencias espaciales de gran escala. Otra aplicación común de este modelo consiste en la posibilidad de obtención de interpolaciones espaciales, dado que la superficie tendencial es función únicamente de una localización de puntos (sus coordenadas), pudiendo conocerse los valores de predicción para cualquier localización para la que se conozcan sus coordenadas.

³⁸ Se entiende por centroide el centro geográfico de un polígono que, en la mayoría de los casos, se encuentra dentro de los límites del mismo, a mitad de camino entre los extremos norte-sur, este-oeste de dicha región.

Un aspecto a tener en cuenta al interpretar los resultados de la estimación de este tipo de modelos es el alto grado de multicolinealidad que suele producirse, debido a la gran relación funcional existente entre los distintos términos del polinomio. Como consecuencia, no debería utilizarse este modelo para otros fines fuera del alisado, filtrado o interpolación de datos (ver en el Anexo I los modelos propuestos por López y Palacios, 2000; López *et al.*, 2001).

➤ **Modelo de regresiones ponderadas geográficamente (RPG), de Fotheringham, Charlton y Brundson**

El modelo RPG pretende también, como en los casos anteriores, recoger adecuadamente el fenómeno de inestabilidad paramétrica continua sobre el espacio geográfico, mediante un modelo estimado por mínimos cuadrados ponderados, siendo los pesos una función de la distancia entre cada punto y el resto. Así, en el modelo siguiente:

$$y_i = X_i \beta_i + u_i \quad \text{Eq. IV.1.75.}$$

el vector de parámetros (β_i) es obtenido a través de un método de estimación de mínimos cuadrados ponderados, de la siguiente forma:

$$\beta_i = [X' W_i X]^{-1} X' W_i y \quad \text{Eq. IV.1.76.}$$

siendo W_i una matriz diagonal de orden (N,N) que, a diferencia del habitual método de mínimos cuadrados ponderados, es diferente para cada observación (i). Fuera de la diagonal principal de W_i , los elementos son nulos, mientras que en la diagonal principal se sitúan las ponderaciones w_{ij} obtenidas como una función de la distancia entre dicha observación (i) y el resto. Por ejemplo, en Fotheringham *et al.* (1997)³⁹, se sugiere la siguiente especificación:

$$w_{ij} = e^{-\alpha \cdot d_{ij}^2} \quad \text{Eq. IV.1.77.}$$

siendo α : parámetro que expresa la caída exponencial de la distancia entre dos puntos
 d_{ij} : distancia entre los puntos i,j.

Así contruidos, los pesos (w_{ij}) de la matriz W_i , correspondientes al punto i, serán mayores para aquellas localizaciones situadas más próximas a i. Esto supone que cada ecuación del modelo mide las relaciones inherentes al modelo alrededor de cada punto i. Asimismo, la posibilidad de obtener la desviación estándar de las N estimaciones de los parámetros del modelo permitirá analizar el alcance de la no estacionariedad en las relaciones entre X e y. Por otro lado, los autores señalan que con las RPG, no sólo es posible obtener diferentes estimadores, sino también distintos contrastes, como la t de

³⁹ Ver en Anexo I un resumen de la aplicación realizada por estos autores.

Student o medidas de bondad del ajuste (R^2), para cada localización geográfica. Estas últimas resultan muy útiles para explorar la posibilidad de añadir nuevas variables explicativas al modelo.

2. Modelos con inestabilidad paramétrica discreta

Por su parte, la inestabilidad paramétrica discreta (“spatial regimes”) es también un caso de inestabilidad paramétrica que evita una estimación global de coeficientes diferentes para toda la muestra de datos (N), mediante la división de la misma en un número limitado (s) de estructuras diferentes (siendo $s < N$), superando así el problema de los parámetros incidentales o la falta de grados de libertad, con el fin de obtener estimaciones eficientes.

Pueden distinguirse dos especificaciones para la inestabilidad paramétrica discreta, el modelo ANOVA espacial (SANOVA), aplicable más bien en un contexto de análisis exploratorio univariante, y el modelo de regresiones cambiantes (“switching regressions”).

➤ Modelo ANOVA espacial (SANOVA)

El análisis espacial de la varianza, ANOVA espacial (Griffith, 1992), consiste en la aplicación del ANOVA tradicional al contexto espacial, con el objetivo de contrastar si la media de una variable determinada difiere de forma significativa entre diferentes grupos de datos (estructuras o regímenes espaciales). Por ejemplo, es evidente que la distribución de variables económicas, como la renta disponible o el PIB, presenta medias diferentes en el grupo de provincias españolas localizadas en las dos submitades, norte y sur, peninsulares.

El SANOVA mide el grado en que estas diferencias en las medias por grupos o zonas del espacio es, de hecho, significativo o puramente casual. En términos propios, el objetivo de este tipo de análisis se centra en encontrar diferencias significativas en el valor medio de una variable cuando ésta es sometida a diferentes “tratamientos” que, en estos casos, son de tipo geográfico, por cuanto engloban diversas sub-zonas.

Este procedimiento parte de la especificación del siguiente modelo de regresión:

$$\begin{aligned} y &= \beta_0 + \beta_1 f + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.1.78.}$$

donde y : variable dependiente (sometida a diferentes tratamientos espaciales)
 f : variable categórica que define los diferentes tratamientos (estructuras)
 β_0, β_1 : coeficientes a estimar
 u : perturbación aleatoria esférica.

En el caso frecuente de que la variable tratamiento categórica (f) sea una variable ficticia (“dummy”) binaria, formada por ceros y unos, el valor estimado para el término constante (β_0) será el valor medio de las regiones que muestren un 0 en la variable ficticia y

el coeficiente estimado para la misma (β_1), reflejará la diferencia entre la media anterior y la derivada del subgrupo de regiones con valor 1 en la ficticia. De este modo, un valor alto significativo en el coeficiente estimado β_1 indicará la existencia de una fuerte discrepancia entre las medias de las dos estructuras espaciales definidas, justificando un tratamiento diferencial de ambos casos como consecuencia de la inestabilidad detectada.

Todos estos resultados están totalmente condicionados al cumplimiento de las hipótesis de homoscedasticidad y no autocorrelación espacial en la perturbación aleatoria que, en caso de no cumplirse, producirá estimaciones incorrectas, a no ser que el modelo sea convenientemente reespecificado. Por ejemplo, en el caso de haberse contrastado en la estimación MCO la existencia de autocorrelación espacial en la variable endógena, el modelo SANOVA anterior podría ser especificado como un modelo del retardo espacial:

$$y = \rho Wy + \beta_0 + \beta_1 f + u$$

$$u \approx N(0, \sigma^2 I)$$

Eq. IV.1.79.

pudiendo producirse alteraciones, más o menos importantes, en el valor y la significación de los estimadores de los parámetros β_0, β_1 .

➤ **Modelo espacial de regresiones cambiantes (“switching regressions”)**

Otra posibilidad de introducir la heterogeneidad en un modelo de regresión espacial y evitar una estimación global, sería una solución de regresiones espaciales cambiantes o “switching regressions”, al estilo de la propuesta por Quandt (1958), que estima tantos valores para los coeficientes de una regresión como estructuras o regímenes espaciales se establezcan en la muestra total de observaciones.

Por ejemplo, si se establecen 2 estructuras, siguiendo el valor de una variable indicador (ficticia) s , tanto el término constante como el resto de parámetros del modelo adoptarán 2 diferentes conjuntos de valores, según la estructura:

$$y_1 = X_1 \beta_1 + u_1 \quad , \quad \text{para } d = 0$$

$$y_2 = X_2 \beta_2 + u_2 \quad , \quad \text{para } d = 1$$

donde y_1 y X_1 son subconjuntos de variables dependientes y explicativas correspondientes a la primera estructura, e y_2 y X_2 son los de la segunda estructura, siendo β_1, β_2 los coeficientes de la regresión, y u_1, u_2 los vectores de las perturbaciones aleatorias. Este modelo puede también expresarse, de una forma conjunta, de la siguiente manera:

$$\begin{bmatrix} y_i \\ y_j \end{bmatrix} = \begin{bmatrix} X_i & 0 \\ 0 & X_j \end{bmatrix} \begin{bmatrix} \beta_i \\ \beta_j \end{bmatrix} + \begin{bmatrix} u_i \\ u_j \end{bmatrix}$$

$$y^* = X^* \beta^* + u^*$$

Eq. IV.1.80.

IV.2. MÉTODOS DE ESTIMACIÓN Y CONTRASTE DE LOS EFECTOS ESPACIALES

Una parte muy importante de la literatura econométrica espacial está centrada en el desarrollo de contrastes sobre los efectos de autocorrelación y heterogeneidad espacial en los modelos espaciales de regresión lineal, así como en la formulación de estimadores eficientes y consistentes para todo este tipo de modelos. Los estadísticos más antiguos para el contraste, en concreto, de autocorrelación espacial son los tests de Moran y Geary, propuestos por Cliff y Ord (1973, 1981). Además, se han añadido también algunas variantes espaciales de los contrastes de Wald, cociente de verosimilitud y multiplicador de Lagrange, englobados en la teoría de la máxima-verosimilitud (Anselin, 1988A).

En cuanto a los métodos de estimación, dada la demostrada invalidez, en presencia de los efectos espaciales, de los estimadores mínimos cuadrados ordinarios, se ha optado preferentemente por el método de máxima-verosimilitud, como procedimiento homólogo a los mínimos cuadrados generalizados en el contexto temporal (aunque, a diferencia de estos últimos, presenta problemas de falta de consistencia).

El análisis confirmatorio espacial puede llevarse a cabo a través de un proceso que tiene por primera fase la especificación del modelo propuesto como un modelo básico de regresión lineal (MBRL) sin efectos espaciales, que se estima por el método de mínimos cuadrados ordinarios (MCO). A continuación, sobre la perturbación aleatoria se aplican una serie de contrastes de autocorrelación y heterogeneidad espacial con el objeto de aceptar o rechazar la hipótesis nula de ausencia de efectos espaciales (aunque también deben contrastarse el resto de hipótesis básicas, comunes al contexto temporal clásico, como la normalidad de la perturbación aleatoria y no multicolinealidad). La aceptación de dicha hipótesis nula supondrá la aceptación del MBRL y sus estimadores MCO, pero el rechazo de la misma requerirá de nuevas especificaciones del modelo (en un proceso de “feedback”) que incluyan convenientemente los efectos espaciales, así como de otros métodos de estimación (máxima-verosimilitud, mínimos cuadrados espaciales en dos etapas, método de los momentos, etc.), que produzcan buenas propiedades en los estimadores.

En este Apartado⁴⁰, se desarrollará con mayor detalle este proceso, propuesto por Anselin (1995A, 1999B), que tiene su punto de partida en la especificación, para un fenómeno espacial determinado (producción, precio de la vivienda, tasa de delitos, etc.), de un modelo espacial de regresión lineal sin efectos espaciales (MBRL), que se estimará por MCO y sobre el que se contrastarán las hipótesis nulas de ausencia de autocorrelación y heterogeneidad espacial en el mismo (Apartado IV.2.1.). El rechazo de dichas hipótesis conducirá a nuevas especificaciones, como los modelos de dependencia espacial (Apartado

⁴⁰ El proceso que se detalla, a continuación, es llamado estrategia modelizadora clásica, propuesta por Anselin, aunque existen también otros procedimientos (Florax y Folmer, 1992), como la metodología Hendry, que ha sido desarrollada por Florax *et al.*, (1998, 2000). Ver también Moreno y Vayá, 2000.

IV.2.2.) y/o modelos de heterogeneidad espacial (Apartado IV.2.3.), que habrán de ser estimados a través de procedimientos diferentes de los mínimos cuadrados ordinarios y sobre los que habrá que contrastar también la hipótesis nula de ausencia de efectos espaciales.

IV.2.1. Estimación y contrastes del modelo básico de regresión lineal

El propósito general de todo análisis de regresión lineal, también en el campo espacial, es encontrar una relación (lineal) entre una variable dependiente y un conjunto de variables explicativas, que ya se ha expresado como:

$$\begin{aligned} y &= X\beta + u \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad \text{Eq. IV.2.1.}$$

La notación es la misma que la utilizada en el Apartado IV.1., representando los coeficientes de regresión desconocidos mediante la letra griega β y la perturbación aleatoria inobservable, con la letra u . Además, se designará con la letra b al coeficiente estimado por la regresión, y el residuo (o diferencia entre los valores observados de la variable endógena y los valores de predicción) será denominado con la letra e .

IV.2.1.1. Estimación por el método de mínimos cuadrados ordinarios

El método de estimación de mínimos cuadrados ordinarios (MCO), en ausencia de efectos espaciales en el modelo, es capaz de encontrar un buen ajuste entre los valores de predicción (Xb) y los valores observados de la variable dependiente (y), siempre y cuando se verifiquen ciertas hipótesis básicas. Es decir, este método puede llegar a producir estimadores (b) lineales insesgados y óptimos (ELIO), para los parámetros teóricos (β), minimizando la suma del cuadrado de los errores de predicción, de la siguiente manera:

$$b = [X'X]^{-1} X'y \quad \text{Eq. IV.2.2.}$$

Cuando existen evidencias de la existencia de heteroscedasticidad en un modelo de datos espaciales (problema que suele ser reconocido como muy específico de los datos de corte transversal), puede realizarse una inferencia robusta a partir de los estimadores MCO, que permanecerán insesgados, aun incluso en presencia de heteroscedasticidad, aunque su varianza adoptaría una forma más compleja:

$$V(b) = \sigma^2 [X'X]^{-1} X'\Sigma X [X'X]^{-1} \quad \text{Eq. IV.2.3.}$$

siendo Σ la matriz de covarianzas de las perturbaciones aleatorias, escalada de forma que su traza es igual al tamaño de la muestra (N). Una consecuencia importante debida a White (1980) muestra que mientras la matriz Σ , en sí misma, no es estimable a no ser que se lleven a cabo otras hipótesis añadidas, la expresión $\sigma^2 X'\Sigma X$ puede ser estimada

consistentemente por el producto $X'SX$, siendo S la matriz diagonal del cuadrado de los residuos de la regresión.

El estimador consistente de la varianza de los coeficientes MCO en presencia de heteroscedasticidad sería el siguiente:

$$Var(b) = \hat{\sigma}_b = \hat{\sigma}_u^2 [X'X]^{-1} X'SX [X'X]^{-1} \tag{Eq. IV.2.4.}$$

Se ha demostrado que este estimador no es muy eficaz en situaciones con muestras finitas, por lo que McKinnon y White (1985) sugieren una versión mejorada, llamada varianza ajustada de White, que consiste en dividir el cuadrado de los residuos de la matriz diagonal S por el factor de corrección $(1 - K_{ii})$, siendo K_{ii} el elemento i -ésimo de la diagonal principal de la matriz idempotente $X[X'X]^{-1}X'$.

Un enfoque alternativo para estimar la matriz robusta de covarianzas de los estimadores MCO, en presencia de heteroscedasticidad, sería el llamado método “Jackknife” (Efron, 1982), que consiste en un enfoque de remuestreo de forma que cada observación es excluida, por turnos, de la base de datos. La distribución empírica de los estimadores MCO obtenida para las M muestras constituye el fundamento para la estimación de la varianza por este método:

$$V(b) = \frac{M-1}{M} \sum_{i=1}^n \left[b(i) - \frac{1}{M} \sum_{j=1}^n b(j) \right] \left[b(i) - \frac{1}{M} \sum_{j=1}^n b(j) \right] \tag{Eq. IV.2.5.}$$

siendo M el número de muestras y $b(i)$ el estimador MCO obtenido en una base de datos en la que la observación i -ésima ha sido excluida⁴¹.

IV.2.1.2. Contrastes de heteroscedasticidad

La heteroscedasticidad se produce cuando la perturbación aleatoria carece de varianza constante en todas las observaciones (homoscedasticidad). Por tanto, se rechaza la hipótesis nula que presupone constancia en la varianza de la perturbación aleatoria en un modelo estimado por MCO. Cuando esto sucede, los estimadores MCO, aunque insesgados, no serán ya los más eficientes. Es más, la inferencia basada en los estadísticos habituales t de Student y F de Snedecor podría producir resultados engañosos y el ajuste de

⁴¹ MacKinnon y White (1985) demostraron que la matriz de covarianzas puede ser calculada directamente como: $V[b] = \frac{N-1}{N} [X'X]^{-1} \left[X'\Sigma^*X - \frac{1}{N} (X'e^*e^{*'}X) \right] [X'X]^{-1}$, donde Σ^* es la matriz diagonal formada

por el cuadrado de los residuos ajustados, $\left[\frac{e_i}{1 - K_{ii}} \right]^2$, y e^* es un vector formado por la raíz cuadrada de los

elementos de la diagonal principal de Σ^* . Para un estudio más profundo sobre los métodos robustos en el análisis de regresión espacial, ver también Anselin, L (1990A).

bondad R^2 sería incorrecto. En el análisis de datos espaciales, es frecuente encontrar este problema sobre todo cuando se utilizan datos procedentes de unidades espaciales irregulares (con áreas diferentes); también se produce cuando hay diferencias regionales sistemáticas en las relaciones del modelo (por ejemplo, estructuras espaciales); o cuando hay una deriva espacial continua en los parámetros de modelo (es decir, expansión espacial), como se vio en el Apartado anterior. La presencia de cualquiera de estos efectos espaciales invalidaría cualquier modelo de regresión tradicional que los ignorase. Por tanto, el indicio de heteroscedasticidad puede apuntar a la necesidad de incorporar más explícitamente los efectos espaciales, en forma de estructuras espaciales o expansión espacial de los parámetros.

Hay muchos contrastes para la heteroscedasticidad en la literatura⁴², pero todos ellos parten de la hipótesis nula de existencia de homoscedasticidad:

$$H_0 : E[u_i^2] = \sigma^2 \quad \text{Eq. IV.2.6.}$$

La hipótesis alternativa consiste en que cada observación tiene una perturbación aleatoria con varianza distinta, σ_i^2 . En muchos casos, esta definición es demasiado general para ser útil. El grado de especificación con la que la hipótesis alternativa de heteroscedasticidad puede expresarse dependerá del conocimiento que tenga el investigador de los factores que pueden causarla. Debe advertirse que si se tiene la sospecha de la existencia de heteroscedasticidad en el modelo y de sus causas (por ejemplo, a partir de consideraciones teóricas), no debería especificarse un MBRL. Por el contrario, debería considerarse explícitamente el modelo del error heteroscedástico o, como ya se ha indicado, también podría introducirse un estimador robusto en presencia de heteroscedasticidad.

Una aproximación común en la especificación de la hipótesis alternativa sería referir la variabilidad de la varianza del error a un cierto número de variables, a través de una forma funcional que incluya una serie de parámetros (parámetros P):

$$\sigma_i^2 = \sigma^2 f\left(\alpha_0 + \sum_p Z_{pi} \alpha_p\right) \quad \text{Eq. IV.2.7.}$$

donde σ^2 : factor simple de escala
 f: forma funcional
 α_0, α_p : parámetros
 Z_{pi} : P variables para la observación i.

Las formas funcionales más comúnmente utilizadas suelen ser las lineales (para la heteroscedasticidad aditiva) y exponenciales (heteroscedasticidad multiplicativa). Las variables Z incluidas en la especificación de la heteroscedasticidad pueden ser cualquiera de las variables. A menudo, suele resultar de utilidad la elección del área de la variable espacial o cualquier otra variable relacionada con su tamaño (población total, renta total).

⁴² Para una revisión más completa, ver cualquier manual de econometría. Por ejemplo, Johnston (1984), Judge *et al.* (1985), Green, W. (1990) o, en español, Pulido y Pérez (2001).

Los contrastes de heteroscedasticidad más citados en la literatura econométrica espacial, son el test del multiplicador de Lagrange, desarrollado por Breusch y Pagan (1979), así como la versión "estudentizada" de este último sugerida por Koenker (1981) y Koenker y Bassett (1982). El que se utilicen uno u otro depende de la distribución de la perturbación aleatoria que, cuando no es normal, produce resultados sesgados en el test de Breusch-Pagan (BP), que se muestra débil en modelos con muestras pequeñas, por lo que se aconseja la aplicación del test de Koenker-Bassett (KB). Ambos tests son asintóticos y siguen una distribución χ^2 con P grados de libertad, siendo P el número de variables Z presentes en la especificación de la heteroscedasticidad⁴³.

Ambos contrastes, BP y KB, requieren de la especificación exacta de las variables que producen la heteroscedasticidad. Por eso, en los casos en que se disponga de poca información previa sobre el tipo de heteroscedasticidad, puede utilizarse el test de White (1980)⁴⁴, capaz de detectar formas de heteroscedasticidad no especificadas expresamente. Este test es también asintótico y sigue una distribución χ^2 .

Debe tenerse en cuenta que, en bastantes situaciones, se producen conjuntamente los problemas de heteroscedasticidad y dependencia espacial. En estos caso, hay que tener en cuenta que los tests de heteroscedasticidad son muy sensibles a la presencia de dependencia espacial⁴⁵. Es decir, puede ser que los tests indiquen la existencia de heteroscedasticidad en el modelo y que éste no sea el verdadero problema, sino la presencia de dependencia espacial (y también puede suceder lo contrario, como se verá más adelante).

IV.2.1.3. Contrastes de autocorrelación espacial

El fenómeno de autocorrelación espacial en un modelo de regresión espacial puede ser de tipo sustantivo o residual, según que sean los valores de la variable dependiente o la perturbación aleatoria, en cada localización, los que se encuentren correlacionados, respectivamente, con las observaciones de la variable dependiente o perturbación aleatoria de otras localizaciones. La expresión matemática general de esta doble situación sería la siguiente:

$$E[y_i y_j] \neq 0 \quad \text{ó} \quad E[u_i u_j] \neq 0 \quad \text{Eq. IV.2.8.}$$

⁴³ El test BP equivale a la mitad de la suma de los cuadrados de la variable endógena en una regresión de $\left(\frac{e_i}{s_{ML}^2} - 1\right)$ sobre una constante y las variables Z. El test KB es la versión "estudentizada" de ésta, en la que el término s_{ML}^2 se sustituye por un estimador más robusto del cuarto momento de dicha distribución.

⁴⁴ El test de White consiste en n veces el coeficiente de determinación R^2 de la regresión auxiliar de los residuos al cuadrado (en la regresión por MCO) sobre los productos cruzados entre las variables explicativas. En algunos casos, los cuadrados o productos cruzados de las variables están ya incluidos como variables explicativas en la regresión original, como en la especificación de los modelos de superficie tendencial. En estos casos, sólo los productos cruzados únicos deberían incluirse, para evitar la multicolinealidad perfecta.

⁴⁵ Ver ejercicios de simulación en Anselin (1990A) y Anselin y Griffith (1988).

para localizaciones vecinas i, j . Esta especificación es demasiado general para permitir la estimación de potenciales interacciones de $[N \times (N - 1)]$, para N observaciones. Por eso, ya se ha indicado que la forma de la dependencia espacial adquiere una estructura a partir de la matriz W de pesos espaciales que reduce el número de parámetros desconocidos a uno sólo (el coeficiente de asociación espacial en un proceso autorregresivo espacial o de medias móviles).

Las consecuencias de ignorar este efecto de autocorrelación espacial en un modelo de regresión, cuando se encuentra de hecho presente, depende de la forma de la hipótesis alternativa. Como en el resto de contrastes, la hipótesis nula pone de manifiesto la ausencia de una mala especificación en un MBRL con errores homoscedásticos e incorrelacionados aunque, como ya se ha indicado, podrían presentarse dos modelos alternativos interesantes según que el fenómeno ignorado de dependencia espacial se encuentre presente en la variable dependiente (modelo del retardo espacial) o en el término de la perturbación aleatoria (modelo del error espacial). Además, como ya se ha expuesto anteriormente, el término del error podría presentar una estructura de dependencia espacial autorregresiva (SAR) y/o de medias móviles (SMA).

La hipótesis nula de ausencia de autocorrelación espacial en la variable dependiente o dependencia sustantiva, $H_0(\rho=0)$, ignora la presencia de este efecto espacial que, de existir, produciría estimadores (b) sesgados, siendo incorrecta toda la inferencia realizada sobre el MBRL. Es decir, se trata de una consecuencia parecida a la que se produce al ignorar una variable explicativa significativa en el modelo de regresión.

Por su parte, la hipótesis nula de ausencia de autocorrelación espacial en el término de la perturbación aleatoria o dependencia residual, $H_0(\lambda=0)$, supone incorrelación en el término de error del modelo. Las consecuencias de ignorar la dependencia espacial en la perturbación aleatoria son las mismas que en el caso de heteroscedasticidad: aunque los estimadores MCO permanecen insesgados, resultan ineficientes al ignorar la correlación existente entre los términos del error. Por tanto, toda la inferencia basada en los estadísticos t de Student y F de Snedecor será confusa y los indicadores de bondad del tipo R^2 , incorrectos.

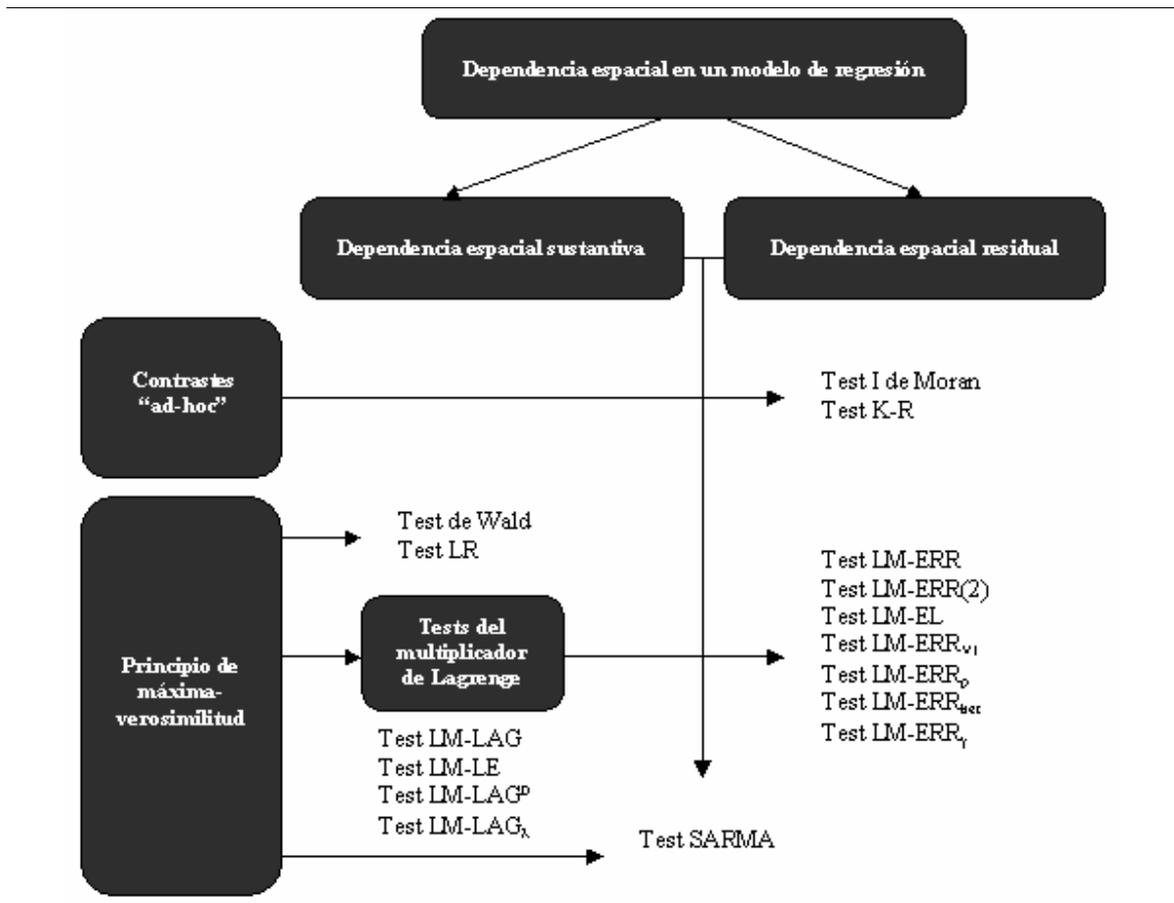
Existe una amplia batería de estadísticos espaciales para contrastar estos dos tipos de estructuras espaciales (dependencia sustantiva y residual) en un modelo (Figura IV.2.1.). Según Moreno y Vayá (2000), podrían distinguirse dos grupos de contrastes de dependencia espacial: los tests “ad-hoc”, que no presentan una hipótesis alternativa definida (tests I de Moran y $K-R$), y los contrastes basados en las propiedades óptimas del estimador máximo-verosímil (MV)⁴⁶, que están rigurosamente estructurados en términos de una

⁴⁶ En términos generales, los tests asintóticos están basados en diferentes medidas de la distancia que existe entre un estimador MV no restringido (bajo la hipótesis alternativa) y el estimador MV restringido, que satisface la hipótesis nula. Si la distancia entre ambos es demasiado grande, la restricción no podrá aceptarse por lo que la hipótesis nula será rechazada. Así, por ejemplo, en el caso del test de Wald, el modelo que se estima es el modelo completo, por lo que se usan las estimaciones no restringidas, mientras que el test de los multiplicadores de Lagrange utiliza el modelo estimado bajo la hipótesis nula, es decir, la especificación más

hipótesis nula y alternativa específicas (tests asintóticos de Wald, razón de verosimilitud y multiplicadores de Lagrange).

A continuación, se presentan los contrastes de dependencia espacial más utilizados en la literatura, algunos de los cuales pertenecen al caso del error espacial y otros, al caso del retardo espacial.

Figura IV.2.1. Métodos de estimación en presencia de dependencia espacial



Fuente: Elaboración propia, a partir de Moreno y Vayá (2000).

- El **test I de Moran**, tal como fue definido matemáticamente en el Apartado III.2.3., puede utilizarse para medir el efecto de autocorrelación espacial en los residuos de una regresión, sin distinguir entre estructuras autorregresivas (λ) y/o de medias móviles (θ). Este contraste se expresa del modo siguiente:

sencilla. Por su parte, el test de la razón de verosimilitud utiliza ambas estimaciones, restringida y no restringida.

$$I = \frac{N}{S_0} \frac{\sum_{(2)} w_{ij} e_i e_j}{\sum_{i=1}^N e_i^2} = \frac{N}{S} \frac{e' W e}{e' e} \quad \text{Eq. IV.2.9.}$$

siendo N: nº de observaciones
 S_0 : suma de los pesos espaciales
 e: vector de residuos MCO (ver Cliff y Ord, 1972 y 1981, para la obtención de los momentos y otras cuestiones técnicas).

Como ya se ha expuesto, la inferencia se basa en un valor (z) estandarizado que se distribuye asintóticamente como una normal. Aunque la interpretación de este estadístico es la misma que en el caso general univariante, las expresiones para sus correspondientes momentos de primer y segundo orden, necesarias para obtener este valor (z) estandarizado, son más complejas, como se verá a continuación:

$$E[I] = \frac{N}{S} \frac{\text{tr}(MW)}{N - K} \quad \text{Eq. IV.2.10.}$$

$$E[I]^2 = \frac{\left(\frac{N}{S}\right)^2 \left[\text{tr}(MWMW') + \text{tr}(MW)^2 + [\text{tr}(MW)]^2 \right]}{(N - K)(N - K + 2)} \quad \text{Eq. IV.2.11.}$$

donde tr: traza de una matriz
 M: matriz idempotente, $M = I - X[X'X]^{-1}X'$.

El test I de Moran de los residuos de la regresión, que se reduce a un simple test de correlación entre vecinos próximos en el espacio, es el contraste de autocorrelación espacial más conocido, aunque es también poco fiable. Efectivamente, como se ha demostrado en un gran número de experimentos de simulación tipo Monte Carlo (Anselin y Rey, 1991), este estadístico puede recoger también un cierto número de errores de mala especificación, como no normalidad y heteroscedasticidad, así como los problemas de dependencia espacial en la variable endógena, no permitiendo discernir entre ambos tipos de dependencia espacial (residual y/o sustantiva) en un MBRL.

- Otro contraste “ad-hoc”, útil para la detección de dependencia espacial residual en un modelo (autorregresiva y/o de medias móviles), es el **test K-R** de especificación robusta desarrollado por Kelejian y Robinson (1992). Este estadístico se obtiene a partir de una regresión auxiliar en la que la variable dependiente está compuesta por los productos cruzados de los residuos MCO de las observaciones que, según la matriz de pesos espaciales, podrían estar afectadas de autocorrelación espacial. Por su parte, la matriz de variables explicativas estaría formada por los productos cruzados de las variables explicativas correspondientes a dichas observaciones, del modo siguiente:

$$C = Z\gamma + A \quad \text{Eq. IV.2.12.}$$

- siendo C: productos cruzados de los residuos MCO (e), siendo el término h-ésimo: $C_h = e_i e_j$
- Z: productos cruzados de variables explicativas del MBRL, siendo el término h-ésimo: $Z_h = X_i X_j$
- γ : vector de coeficientes
- A: perturbación aleatoria de la regresión auxiliar.

Denominando α al vector de residuos resultante, el estadístico K-R tiene la siguiente expresión

$$K - R = \frac{\gamma' Z' Z \gamma}{\alpha' \alpha / h_R} \quad \text{Eq. IV.2.13.}$$

donde h_R es el número total de observaciones del vector auxiliar. Este estadístico se distribuye según una χ^2 con K grados de libertad, siendo K el número de variables explicativas de X.

A diferencia de los tests I de Moran y multiplicador de Lagrange, el test K-R no exige, como hipótesis previa, la normalidad en los términos de la perturbación aleatoria. Por otra parte, es aplicable tanto en regresiones lineales como no lineales y requiere de menos información acerca de la forma exacta de la matriz de pesos espaciales. Sin embargo, se trata de un contraste para grandes muestras que puede no tener mucha fuerza cuando se trabaja con pequeñas bases de datos.

Además, tal como se ha indicado, los contrastes de dependencia espacial “ad-hoc”, I de Moran y K-R, no presuponen ningún esquema de dependencia espacial, ni en la perturbación aleatoria ni en la variable endógena, por lo que no resultan muy útiles a la hora de discriminar entre la existencia de un esquema de autocorrelación espacial residual o sustantiva siendo, en este sentido, inferiores a los contrastes basados en el multiplicador de Lagrange, que se presentan a continuación.

- El **test LM-ERR**, basado en el principio de los multiplicadores de Lagrange, fue originalmente propuesto por Burridge (1980) y se expresa como:

$$LM - ERR = \frac{\left[\frac{e' W e}{s^2} \right]^2}{tr[W' W + W^2]} \quad \text{Eq. IV.2.14.}$$

- donde $s^2 = e' e / N$ el estimador MV de la varianza de la perturbación aleatoria
- e: vector de residuos de la estimación MCO.

Este contraste, que es también un test asintótico, se distribuye como una χ^2 con un grado de libertad y funciona igual tanto para la hipótesis alternativa de errores espacialmente autorregresivos como para medias móviles espaciales: $H_A(\lambda \neq 0, \theta \neq 0)$.

Tal como ponen de manifiesto Moreno y Vayá (2000), existe una relación entre los tests I de Moran y LM-ERR de forma que, cuando la matriz W está estandarizada por filas, puede afirmarse la siguiente relación:

$$LM - ERR = \frac{(NI)^2}{tr[W'W + W^2]} \quad \text{Eq. IV.2.15.}$$

siendo I el contraste I de Moran. Ambos contrastes son asintóticamente equivalentes aunque, a diferencia del test I de Moran, el test LM-ERR no requiere del cómputo de los momentos de primer y segundo orden siendo, por ello, más fácil su implementación.

Anselin (1994B) propone, además, un nuevo contraste LM-ERR(2) que permite contrastar la existencia de procesos autorregresivos o de medias móviles de orden 2, del modo siguiente:

$$LM - ERR(2) = \frac{\left[\frac{e'W_1e}{s^2} \right]^2}{tr[W_1'W_1 + W_1^2]} + \frac{\left[\frac{e'W_2e}{s^2} \right]^2}{tr[W_2'W_2 + W_2^2]} \quad \text{Eq. IV.2.16.}$$

siendo W_1 y W_2 las matrices de pesos espaciales de orden 1 y 2, respectivamente. Este contraste, así definido, sigue una distribución χ^2 con 2 grados de libertad, dado que contrasta la significación de los dos coeficientes autorregresivos/medias móviles de forma simultánea. Como puede observarse, el contraste LM-ERR(2) podría extenderse a órdenes superiores a 2 como la suma de los respectivos tests unidireccionales, agregando los sumandos correspondientes.

- El **test LM-EL** es un contraste de dependencia espacial de la perturbación aleatoria, similar al test LM-ERR, aunque robusto a una mala especificación local de la dependencia espacial como sería el caso de la existencia de una variable endógena espacialmente retardada (Anselin *et al.*, 1996). La hipótesis nula a contrastar sería, en este caso, $H_0(\lambda=0, \theta=0)$, robusto a $\rho \neq 0$, siendo su expresión formal la siguiente:

$$LM - EL = \frac{\left[\frac{e'We}{s^2} - T_1(R\tilde{J}_{\rho-\beta})^{-1} \frac{e'Wy}{s^2} \right]^2}{T_1 - T_1^2(R\tilde{J}_{\rho-\beta})} \quad \text{Eq. IV.2.17.}$$

$$\text{con: } (R\tilde{J}_{\rho-\beta})^{-1} = \left[T_1 + \frac{(WX\beta)'M(WX\beta)}{s^2} \right]^{-1}$$

$$T_1 = tr(W'W + W^2)$$

W: matriz de pesos espaciales

WXβ: retardo espacial de los valores procedentes de la estimación MCO de y sobre X

Este estadístico, que se distribuye como una χ^2 con 1 grado de libertad, ha sido calculado bajo el supuesto de que la matriz de pesos espaciales presente en el término del error coincide con la utilizada para obtener el retardo espacial de la variable endógena (ver en Anselin *et al.*, 1996, una generalización de este supuesto).

- Otro contraste derivado del multiplicador de Lagrange, es el **test LM-LAG** para retardos espaciales de la variable endógena, propuesto por Anselin (1988A), sólo válido bajo la hipótesis de normalidad y asintótico por naturaleza. Como su homólogo LM-ERR, para una perturbación aleatoria con problemas de dependencia espacial, el test de retardos LM se distribuye como una variable χ^2 con 1 grado de libertad. En este caso, la hipótesis nula a contrastar sería $H_0(\rho=0)$. La expresión matemática del mismo es la siguiente

$$LM - LAG = \frac{\left[\frac{e'Wy}{s^2} \right]^2}{R\tilde{J}_{\rho-\beta}} \quad \text{Eq. IV.2.18.}$$

siendo y el vector (N,1) de observaciones de la variable dependiente. Como se ha puesto de manifiesto en un gran número de experimentos de simulación de Monte Carlo, en Anselin y Rey (1991), el uso conjunto de los estadísticos LM-ERR y LM-LAG proporciona la mejor ayuda para elegir uno de los dos modelos alternativos, dependencia estructural o aleatoria, siempre que se cumpla la hipótesis de normalidad. Cada contraste por separado posee la máxima capacidad sobre la otra alternativa. En otras palabras, cuando ambos tests ofrecen valores altos (indicativos de dependencia espacial significativa), el que tenga valor máximo (máxima confianza) será el que muestre la estructura correcta de dependencia espacial.

- En cuanto al contraste de retardos, robusto a la presencia de dependencia local de error espacial, el **test LM-LE**, su expresión matemática es la siguiente:

$$LM - LE = \frac{\left[\frac{e'W_1y}{s^2} - \frac{e'W_1e}{s^2} \right]^2}{R\tilde{J}_{\rho-\beta} - T_1} \quad \text{Eq. IV.2.19.}$$

En este caso, la hipótesis nula a contrastar sería $H_0(\rho=0)$, robusto a $\lambda \neq 0$, $\theta \neq 0$. Esta expresión, que sigue la misma notación que en casos anteriores, está también distribuida como una χ^2 con 1 grado de libertad y ha sido construida bajo el supuesto de coincidencia entre las matrices de pesos espaciales correspondientes a

la variable endógena y la perturbación aleatoria (aunque en Anselin *et al.*, 1996, se propone una generalización de esta restricción).

- Se ha propuesto también un contraste basado en el multiplicador de Lagrange capaz de contrastar conjuntamente la presencia, en un modelo, de dependencia espacial sustantiva y residual, siendo la hipótesis nula $H_0(\rho=\theta=0)$ ó $H_0(\rho=\lambda=0)$. Se trata del **test SARMA** multidireccional, idéntico al LM conjunto para retardo espacial y error espacial autorregresivo, de Anselin (1988A y B), exceptuando la restricción sobre las matrices de pesos espaciales (no siempre puede identificarse un proceso con matrices de pesos idénticas tanto para el retardo como para el error). El estadístico correspondiente es el siguiente:

$$SARMA = \frac{\left[\frac{e'Wy}{s^2} - \frac{e'We}{s^2} \right]^2}{R\tilde{J}_{\rho-\beta} - T_1} + \frac{\left[\frac{e'We}{s^2} \right]^2}{T_1} \quad \text{Eq. IV.2.20.}$$

Este estadístico se distribuye como una χ^2 con 2 grados de libertad.

- El test LM puede también formularse para contrastar la presencia de un retardo espacial en la variable endógena de un modelo en el que la perturbación aleatoria presenta un esquema autorregresivo espacial. Este estadístico ha sido denominado **test LM-LAG λ** y se expresa del modo siguiente:

$$LM_{\rho}^A = \frac{[\hat{u}' B' BW_1 y]^2}{H_{\rho} - H_{\vartheta\rho} \hat{\text{var}}(\hat{\vartheta}) H'_{\vartheta\rho}} \quad \text{Eq. IV.2.21.}$$

donde \hat{u} : vector de residuos de la estimación máximo verosímil del modelo obtenido bajo la hipótesis nula de no dependencia espacial sustantiva, aunque con errores espacialmente autorregresivos.

$$y = X\beta + (I - \lambda W_2)^{-1} u$$

$$\vartheta = [\beta', \lambda, \sigma^2]$$

$$B = I - \lambda W_2$$

$$H_{\rho} = tr W_1^2 + tr [BW_1 B^{-1}] [BW_1 B^{-1}] + \frac{1}{\sigma^2} [BW_1 X\beta]' [BW_1 X\beta]$$

$$H_{\vartheta\rho} = \begin{bmatrix} \frac{1}{\sigma^2} (BX)' BW_1 X\beta \\ tr [W_2 B^{-1}] BW_1 B^{-1} + tr W_2 W_1 B^{-1} \\ 0 \end{bmatrix}_{(3,1)}$$

$\hat{\text{var}}[\hat{\vartheta}]$: matriz de varianzas estimadas del vector de parámetros θ en el modelo estimado bajo la hipótesis nula.

Este estadístico se distribuye como una χ^2 con 1 grado de libertad. Para mayores detalles, ver Anselin et al. (1996) y Anselin y Florax (1995).

El rechazo de la hipótesis nula de ausencia de autocorrelación y heterogeneidad espacial en el MBRL, supone una re-especificación del mismo, a través de modelos que incorporen estos efectos en la forma funcional. Por eso, a continuación, se exponen los métodos de estimación, y los contrastes sobre los efectos espaciales, tanto para los modelos de dependencia espacial (Apartado IV.2.2.) como para los modelos de heterogeneidad espacial (Apartado IV.2.3.).

IV.2.2. Estimación y contrastes de modelos de dependencia espacial

IV.2.2.1. Mínimos cuadrados ordinarios y dependencia espacial

La dependencia espacial en los diversos modelos autorregresivos presenta muchas similitudes con la más familiar dependencia temporal. Por eso, podría resultar lógico que las propiedades de los estimadores mínimos cuadrados ordinarios (MCO) en modelos con variables retardadas y/o autocorrelación temporal pudieran trasladarse directamente al caso espacial.

Sin embargo, esto no es así, debido fundamentalmente a la naturaleza multidireccional de la dependencia en el espacio, en contraste con el sentido bidireccional de la dependencia temporal (del pasado hacia el futuro), tal como se ha presentado en los Capítulos anteriores. Por eso, se analizan previamente los efectos que el fenómeno de autocorrelación espacial, tanto en la variable dependiente como en la perturbación aleatoria, causan sobre las propiedades de los estimadores MCO en modelos de dependencia espacial.

➤ Mínimos cuadrados ordinarios y variables dependientes desplazadas o espacialmente retardadas

En econometría es bien sabido que los estimadores MCO, aunque sesgados, permanecen consistentes en especificaciones con variables dependientes desplazadas, siempre y cuando no exista autocorrelación espacial en la perturbación aleatoria. Por eso, aunque las propiedades de los estimadores se ven afectadas en muestras pequeñas, seguirán siendo asintóticamente insesgados. Sin embargo, en los modelos autorregresivos espaciales los estimadores MCO son, como se verá a continuación, sesgados e inconsistentes, independientemente del comportamiento de la perturbación aleatoria.

Consideremos, para ello, como en Anselin (1988A), el caso más simple del modelo espacial autorregresivo de primer orden, sin autocorrelación espacial en la perturbación aleatoria:

$$y = \rho W y + u = \rho y_D + u \quad \text{Eq. IV.2.22.}$$

- donde $y_{(N,1)}$: vector de observaciones de la variable dependiente.
 ρ : coeficiente de la variable dependiente espacialmente retardada.
 $W_{(N,N)}$: matriz de pesos espaciales simétrica, de orden (N,N), predefinida
 y_D : matriz (W.Y) de observaciones de la variable dependiente espacialmente desplazada.
 $u_{(N,1)}$: vector de términos de la perturbación aleatoria, idéntica e independientemente distribuidos, con media cero $E(u) = 0$ y matriz de covarianzas escalar $E(uu') = \sigma^2$.

Expresado en forma matricial desarrollada, el modelo anterior tendría la siguiente forma:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} = \rho \begin{bmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & 0 & \dots & w_{2N} \\ \dots & \dots & \dots & \dots \\ w_{N1} & w_{N2} & \dots & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_N \end{bmatrix} = \rho \begin{bmatrix} 0 & w_{12}y_2 & \dots & w_{1N}y_N \\ w_{21}y_1 & 0 & \dots & w_{2N}y_N \\ \dots & \dots & \dots & \dots \\ w_{N1}y_1 & w_{N2}y_2 & \dots & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_N \end{bmatrix} \quad Eq. IV.2.23.$$

Aunque este modelo es muy sencillo, recoge totalmente los efectos que produce sobre los estimadores MCO la presencia de una variable dependiente desplazada. Por eso, puede ser utilizado sin pérdida de generalidad.

Así, el estimador MCO del parámetro ρ ($\hat{\rho}$) podría obtenerse como:

$$\hat{\rho} = [y'_D y_D]^{-1} y'_D y \quad Eq. IV.2.24.$$

Sustituyendo el valor del vector y de la Eq.IV.2.22, en la Eq.IV.2.24:

$$\begin{aligned} \hat{\rho} &= [y'_D y_D]^{-1} y'_D (\rho W y + u) = \{y_D = W y\} = \\ &= \rho [y'_D y_D]^{-1} [y'_D y_D] + [y'_D y_D]^{-1} y'_D u = \\ &= \rho + [y'_D y_D]^{-1} y'_D u \end{aligned} \quad Eq. IV.2.25.$$

Una vez obtenido el estimador MCO del parámetro espacial autorregresivo (ρ), se analizan las propiedades del mismo, pudiéndose demostrar que no son buenas, pues se trata de un estimador sesgado e inconsistente, como se verá a continuación.

- **Estimador sesgado:**

$$E(\hat{\rho}) = E\left\{ \rho + [y'_D y_D]^{-1} y'_D u \right\} \neq \rho \quad Eq. IV.2.26.$$

El valor esperado del segundo sumando del término de la derecha no se anula, por lo que los estimadores MCO son sesgados. Efectivamente, como en los modelos de

series temporales, esto es debido a la naturaleza estocástica compleja del término inverso $[y_D' y_D]^{-1}$, cuyos elementos son función de la variable dependiente (y, por tanto, también de la perturbación u). Es más, mientras en los modelos temporales el término $E[y_D' u] = 0$, cuando no existe autocorrelación en la perturbación aleatoria, en los modelos espaciales esto no es así, puesto que el término $E[y_D' u]$ únicamente se anula en el caso particular de que $\rho = 0$:

$$E[y_D' u] = \begin{cases} y = \rho W y + u \\ y = (I - \rho W)^{-1} u \\ y_D = W y \end{cases} = E\left\{W \cdot (I - \rho W)^{-1} u\right\}' u = \{\text{Si } \rho = 0\} = 0 \quad \text{Eq. IV.2.27.}$$

- **Estimador inconsistente:**

Es bien sabido que, cuando el tamaño de la muestra tiende a la población total, se da una coincidencia entre estimador $\hat{\rho}$ y parámetro ρ . En nuestro caso, el estimador $\hat{\rho}$ MCO será consistente si en la Eq.IV.2.25. sucede que:

$$PLim_{N \rightarrow \infty} [y_D' y_D]^{-1} y_D' u = 0 \Leftrightarrow PLim_{N \rightarrow \infty} \left[\frac{y_D' y_D}{N} \right]^{-1} \left[\frac{y_D' u}{N} \right] = 0 \quad \text{Eq. IV.2.28.}$$

Esto sólo se cumple con 2 condiciones:

$$\begin{aligned} 1. \quad PLim_{N \rightarrow \infty} \frac{y_D' y_D}{N} &= Q \\ 2. \quad PLim_{N \rightarrow \infty} \frac{y_D' u}{N} &= 0 \end{aligned} \quad \text{Eq. IV.2.29.}$$

siendo Q una matriz finita y no singular.

Mientras que la primera condición puede llegar a cumplirse a través de restricciones apropiadas para el valor de ρ y en la estructura de la matriz de pesos espaciales (W), la segunda condición es imposible de obtener en la situación espacial, salvo en el caso de que $\rho = 0$, por lo que el estimador MCO es inconsistente. Efectivamente:

$$PLim_{N \rightarrow \infty} \frac{y_D' u}{N} = \left\{ y_D = W(I - \rho W)^{-1} u \right\} = PLim_{N \rightarrow \infty} \frac{1}{N} \left[u' W (I - \rho W)^{-1} u \right] \quad \text{Eq. IV.2.30.}$$

En consecuencia, los estimadores MCO de los parámetros en modelos espaciales son sesgados e inconsistentes, independientemente de las propiedades del término de error.

➤ **Mínimos cuadrados ordinarios y autocorrelación espacial en la perturbación aleatoria**

En este caso, se tendrán en cuenta los efectos que la autocorrelación espacial de la perturbación aleatoria tienen sobre los estimadores MCO en el mismo modelo anterior (Eq.IV.2.22.), cuya especificación se correspondería entonces con el modelo mixto-autorregresivo con perturbación aleatoria autorregresiva, según la taxonomía de Florax y Folmer (1992). Estos efectos son mayores que los que produce la autocorrelación temporal en el MBRL, pues los estimadores, aunque **sesgados** son, en este caso, también **ineficientes**, dada la estructura no escalar de la matriz de varianzas de la perturbación aleatoria: $E(uu') = \sigma^2\Omega$, siendo los elementos de la matriz Ω (Ω_{ij}) función de la variable dependiente. Estas malas propiedades de los estimadores se demuestran mediante la comparación de las expresiones de las matrices de varianzas y covarianzas del parámetro autorregresivo (ρ) en situación de ausencia/presencia de autocorrelación espacial en la perturbación aleatoria.

1. Ausencia de autocorrelación espacial en la perturbación aleatoria:

$$Cov(\hat{\rho}) = E[(\hat{\rho} - \rho) \cdot (\hat{\rho} - \rho)] = E\left\{ \left[y_D' y_D \right]^{-1} y_D' u \right\} \cdot \left[u' y_D \left[y_D' y_D \right]^{-1} \right\} = \Omega \left[y_D' y_D \right]^{-1} \quad \text{Eq. IV.2.31.}$$

2. Presencia de autocorrelación espacial en la perturbación aleatoria:

$$\begin{aligned} Cov(\hat{\rho}) &= E[(\hat{\rho} - \rho)(\hat{\rho} - \rho)] = E\left\{ \left[y_D' y_D \right]^{-1} y_D' u \right\} \cdot \left[u' y_D \left[y_D' y_D \right]^{-1} \right\} = \\ &= \sigma^2 \left[y_D' y_D \right]^{-1} y_D' \Omega y_D \left[y_D' y_D \right]^{-1} \end{aligned} \quad \text{Eq. IV.2.32.}$$

Como puede observarse, esta última expresión es superior anterior, por lo que queda demostrada la ineficiencia de los estimadores MCO, en modelos con perturbación aleatoria espacialmente dependiente. Los modelos de dependencia espacial podrían ser considerados, desde este punto de vista, como casos particulares del modelo de regresión lineal con matriz de covarianzas de la perturbación aleatoria no escalar.

IV.2.2.2. Modelo del retardo espacial

El modelo del retardo espacial o modelo mixto autorregresivo de regresión espacial incluye la variable dependiente espacialmente retardada, Wy , como una variable explicativa más, del modo siguiente:

$$y = \rho Wy + X\beta + u \quad \text{Eq. IV.2.33.}$$

donde y : vector de observaciones de la variable dependiente, de orden (N,1)
 Wy : vector de retardos espaciales de la variable dependiente, de orden (N,1)
 ρ : coeficiente espacial autorregresivo
 X : matriz de observaciones de las variables explicativas, de orden (N,K)

- β : vector de parámetros correspondientes a las variables explicativas, (K,1)
- u : vector de orden (N,1) de perturbaciones aleatorias normalmente distribuidas, con media cero y varianzas constantes (homoscedásticas), σ^2 .

La presencia de un retardo espacial es parecida a la inclusión de variables endógenas en los sistemas de ecuaciones simultáneas, en el contexto temporal clásico. Por eso, este modelo suele ser llamado también **modelo simultáneo espacial autorregresivo** (Anselin, 1995A) Si el parámetro autorregresivo fuera conocido, el modelo podría simplificarse, adoptando la forma de un MBRL cuya variable endógena es la endógena del modelo inicial espacialmente filtrada ($y - \rho Wy$), de un modo similar al filtro espacial de Getis, anteriormente presentado (Getis, 1990, 1995), del siguiente modo:

$$y - \rho Wy = X\beta + u \qquad \text{Eq. IV.2.34.}$$

Sin embargo, lo habitual es que el coeficiente espacial autorregresivo (ρ) sea desconocido, por lo que tenga que ser conjuntamente estimado con los demás parámetros de regresión. La transformación de la variable endógena a través de su filtrado espacial, permitiría contralar el efecto de autocorrelación espacial en el modelo, de forma que sea posible conocer el grado de significatividad de las demás variables. Esta interpretación de la especificación mixta regresiva espacial autorregresiva podría también llevarse a cabo de otro modo, entendiendo la explícita inclusión del retardo espacial (Wy) como un añadido al resto de variables explicativas que permita conocer el grado de dependencia espacial existente en el modelo, a la par que se controla el efecto del resto de variables.

La consecuencia principal de la inclusión del retardo espacial (Wy) en la especificación de un modelo es que la pérdida de consistencia de los estimadores MCO, como sucede también en los sistemas de ecuaciones simultáneas. En vez de utilizar MCO, la estimación debería basarse en el método de máxima-verosimilitud (MV) que es, con mucho, el más utilizado, aunque puede estimarse válidamente este modelo mediante métodos de variables instrumentales (VI), como los mínimos cuadrados espaciales en dos etapas (MC2E) y el método “bootstrap”. De todos ellos, se harán, a continuación, algunas observaciones.

a) Estimación por el método de máxima-verosimilitud (MV)

La estimación MV del modelo de retardos espaciales se basa en la hipótesis de normalidad de la perturbación aleatoria. A partir de ella, puede obtenerse la función de verosimilitud como una función no lineal de los parámetros que debe maximizarse. La función de verosimilitud resultante tiene la siguiente forma:

$$L = \sum_i \ln(1 - \rho \omega_i) - \frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{(y - \rho Wy - X\beta)'(y - \rho Wy - X\beta)}{2\sigma^2} \qquad \text{Eq. IV.2.35.}$$

siendo ω_i los valores propios de la matriz de pesos y el resto de la notación como en el resto del capítulo. Para los detalles completos de la derivación y aplicación, ver Ord (1975), Anselin (1980, 1988A) y Anselin y Hudak (1992).

Esta expresión de optimización no lineal podría solucionarse, en el terreno aplicado de otro modo, expresando los estimadores de los parámetros de la regresión (β) y la varianza de la perturbación aleatoria (σ^2) en función del coeficiente autorregresivo (ρ). La sustitución de estas expresiones en la función de verosimilitud da lugar a la función concentrada de verosimilitud, con un único parámetro, el coeficiente ρ autorregresivo, que tiene la siguiente forma:

$$L_C = -\frac{N}{2} \ln \frac{(e_0 - \rho e_L)'(e_0 - \rho e_L)}{N} + \sum_i \ln(1 - \rho \omega_i) \quad \text{Eq. IV.2.36.}$$

siendo e_0 y e_L los residuos de la regresión MCO de la variable y sobre X y la variable Wy sobre X , respectivamente. La simple búsqueda de valores de ρ dará lugar rápidamente al estimador MV. Los otros parámetros pueden obtenerse a partir de la estimación MCO de la variable $(y - \rho Wy)$ sobre X .

El método MV dispone de un menor número de **contrastes sobre los efectos espaciales** que el método MCO, siendo los más utilizados el test del multiplicador de Lagrange (LM) y el test del cociente de verosimilitud (LR) que, por ser asintóticos, pueden dar a conclusiones inconsistentes en el caso de muestras pequeñas⁴⁷.

El contraste de heteroscedasticidad en el modelo del retardo espacial puede realizarse a través del test de Breusch-Pagan, basado en los residuos de la estimación MV, que es idéntico a la formulación utilizada en el MBRL⁴⁸. Aunque de forma estricta este contraste es incorrecto, dado que ignora la presencia de dependencia espacial en el modelo, en la práctica suele utilizarse habitualmente sin muchos problemas.

En cuanto a los contrastes de dependencia espacial, el test LR del cociente de verosimilitud sobre el coeficiente autorregresivo (ρ) se calcula como el doble de la diferencia entre el logaritmo de verosimilitud del modelo de retardos espaciales y el logaritmo de verosimilitud en el MBRL, con el mismo conjunto de variables explicativas (para $\rho = 0$). Este estadístico se distribuye como una variable χ^2 con 1 grado de libertad.

A pesar de que tanto el test LR, como el contraste de Wald (test t asintótico) y LM-LAG son todos equivalentes asintóticamente, tienden a ofrecer resultados distintos en muestras finitas. En la mayoría de las circunstancias, el orden de dichos estadísticos en

⁴⁷ Ver Anselin (1988, Capítulo 6) para una discusión extensa sobre el multiplicador de Lagrange de Wald y el test del cociente de verosimilitud (LR) en los modelos espaciales.

⁴⁸ El test BP equivale a la mitad de la suma de los cuadrados de la variable endógena en una regresión de

$\left(\frac{e_i}{s_{ML}^2} - 1 \right)$ sobre una constante y las variables heteroscedásticas Z .

términos de su magnitud es la siguiente: $W \geq LR \geq LM$. Este orden implica que será más fácil considerar significativo el coeficiente autorregresivo a partir de los resultados del test de Wald, que del test LM. En algunos casos, estos 3 contrastes pueden alcanzar resultados contradictorios; es decir, un test de Wald significativo ($p = 0,01$), un test LR marginalmente significativo ($p = 0,05$) y un no significativo LM ($p = 0,15$). En estos casos, deberían interpretarse los resultados con precaución. Habitualmente, los resultados se ajustan al orden anterior y, de no ser así, debería sospecharse de un error potencial en la significación, como por ejemplo, no normalidad en las perturbaciones aleatorias, no linealidad en la relación entre variables dependientes y explicativas o incorrecta elección de las variables explicativas y/o la matriz de pesos espaciales.

Si el modelo de retardos espaciales especificado la especificación correcta, entonces no debería quedar en los residuos de la regresión ningún tipo de dependencia espacial. El test del multiplicador de Lagrange para la autocorrelación espacial de los errores LM-ERR propuesto por Anselin (1988A), contrasta esta cuestión. Un resultado significativo en este test sería indicativo de una mala especificación de la matriz de pesos espaciales o de la existencia de algún tipo de dependencia espacial no bien eliminada del modelo. En estos casos, debería probarse con un modelo autorregresivo de mayor orden (en lugar del modelo de retardos espaciales habitual), o también con nuevas especificaciones de la matriz de pesos espaciales, o bien con una especificación del modelo completamente distinta (por ejemplo, un modelo del error espacial). Por otro lado, si la matriz de pesos utilizada en el contraste no es la misma que la empleada para el retardo espacial, un valor significativo del estadístico pondría de manifiesto la conveniencia de un modelo mixto-autorregresivo con perturbación aleatoria autorregresiva.

b) Estimación por el método de variables instrumentales (VI)

La estimación de modelos con una variable dependiente espacialmente retardada puede también llevarse a cabo mediante métodos de variables instrumentales (VI), que constituyen una alternativa más robusta a la estimación máximo-verosímil, pues no se exige la hipótesis de normalidad en la distribución de la perturbación aleatoria. Suelen aplicarse, en concreto, el procedimiento de mínimos cuadrados espaciales bietápicos y métodos tipo "bootstrap".

El primero de ellos es una aplicación directa, al modelo del retardo espacial, del método de estimación VI llamado **mínimos cuadrados bietápicos (MC2E)**. El segundo procedimiento utiliza la estimación con variables instrumentales para construir un procedimiento tipo "bootstrap"⁴⁹. Dado que se trata de métodos robustos, no se ofrece ningún tipo de contraste en ambos casos.

⁴⁹ Los métodos estadísticos clásicos se apoyan en modelos matemáticos de naturaleza estocástica, de tal forma que los resultados que de ellos se derivan requieren, en muchas ocasiones, complejos desarrollos analíticos, lo que ha supuesto un obstáculo para su utilización comprensiva en muchas áreas científicas. Además, dichos desarrollos se basan en hipótesis que algunas veces no son soportadas por los datos o se obtienen resultados asintóticos que no son válidos cuando el tamaño muestral no es suficientemente elevado. Efron (1979) introduce la metodología "bootstrap" para estimar las distribuciones de algunos estadísticos cuando el tamaño muestral es pequeño o las expresiones de dichas distribuciones son analíticamente intratables. Los métodos

Como ya se ha expuesto en repetidas ocasiones, el modelo de retardos espaciales o modelo mixto-autorregresivo de regresión espacial incluye una variable espacialmente retardada (Wy) como una variable explicativa más. Este retardo espacial y la matriz X de regresores, a efectos de notación, se llamará Z y será una matriz de orden $[N(K+1)]$. El estimador de las variables instrumentales (VI) o estimador MC2E aprovechará esta característica para la construcción de un instrumento adecuado para el retardo espacial que, aunque consistente, no será necesariamente muy eficiente, pero constituirá la base para el procedimiento "bootstrap".

El principio de la estimación de variables instrumentales se basa en la existencia de un conjunto de instrumentos, Q , que están fuertemente correlacionados con las variables originales, Z , aunque asintóticamente incorrelacionados con la perturbación aleatoria. Una vez identificados estos instrumentos, se utilizarán en la construcción de una variable "proxi" de las variables endógenas a partir de sus valores estimados en una regresión entre los instrumentos y las variables exógenas. Esta "proxi" se utilizará, a continuación, en un modelo de regresión estimado por MCO. Matemáticamente, este proceso de MC2E da lugar al siguiente estimador:

$$\theta_{IV} = [(Z'Q)(Q'Q)^{-1}(Q'Z)]^{-1}(Z'Q)(Q'Q)^{-1}Q'y \quad \text{Eq. IV.2.37.}$$

donde $\theta_{IV} = \begin{bmatrix} \hat{\rho} \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{bmatrix}_{(K+1,1)}$

Q : matriz de instrumentos (N,P), que incluye la matriz X de variables exógenas.

Puede demostrarse que este estimador es consistente y asintóticamente eficiente. Aunque, en el caso de contar con muestras pequeñas, sus propiedades dependen muy estrechamente de la elección de los instrumentos y no siempre tienen tratamiento matemático⁵⁰. Un problema potencial de la estimación VI es que el estimador del parámetro autorregresivo (ρ) no tiene por qué encontrarse situado necesariamente dentro del intervalo de valores aceptables mayores que la unidad, en valor absoluto lo que, de producirse, suele poner de manifiesto la existencia de problemas de especificación en el modelo.

En el esquema clásico (temporal) de los modelos de ecuaciones simultáneas, los instrumentos son las variables exógenas "excluidas". Sin embargo, en los modelos de retardos espaciales no existe una noción totalmente equivalente a ésta y, por eso, se han ofrecido múltiples sugerencias (ver en Anselin, 1988A, una visión general del tema). Hace

"bootstrap" (como el "jackknife" anteriormente presentado) se basan en la reproducción de los datos originales mediante un remuestreo, de forma que si las observaciones tuvieran, por ejemplo, una estructura de dependencia, ésta debería estar reflejada en los nuevos datos.

⁵⁰ Ver Bowden y Turkington (1984), y ver también Anselin (1980, 1984, 1988A) para un tratamiento más detallado de su aplicación a modelos de retardos espaciales.

unos años, Kelejian y Robinson (1992) demostraron que las series de variables exógenas espacialmente retardadas constituyen el conjunto de variables instrumentales más adecuado, siempre y cuando se trabaje con matrices de contigüidades de uno o más órdenes. En la práctica, estas series pueden estar truncadas de forma que sólo se incluyan las variables explicativas con un retardo espacial de primer orden, lo que ocasionaría que la matriz Q contuviera tanto la matriz X como WX, siendo excluidos de esta última tanto el término constante como cualesquiera otras variables que pudieran dar lugar a una perfecta multicolinealidad.

El **estimador "bootstrap"** es un estimador robusto que aprovecha la aleatoriedad presente en distintas muestras de datos, creadas de forma artificial, como base para la inferencia estadística, lo que produce diversos estimadores de los parámetros, medidas de sesgo y varianza, así como la construcción de niveles de pseudo-significación e intervalos de confianza⁵¹. En el análisis de regresión, hay dos procedimientos importantes que dan lugar a estimadores "bootstrap": uno se basa en los residuos, mientras que el otro parte de las observaciones representadas en un espacio multidimensional. Como se demuestra en Anselin (1988A, 1990A), sólo es aceptable el primero de ellos en modelos del retardo espacial. El "bootstrap" se basa en un procedimiento propio de las ecuaciones simultáneas, sugerido por Freedman y Peters (1984A y B)⁵². En un primer paso, se lleva a cabo la estimación VI, que proporciona un estimador del vector de perturbaciones (u) en forma de residuos:

$$e = y - \rho_{VI}Wy - X\beta_{VI} \quad \text{Eq. IV.2.38.}$$

siendo ρ y β los estimadores VI obtenidos anteriormente. En la segunda etapa, se generan los términos de pseudo-errores mediante muestreo aleatorio (con reemplazamiento) a partir del vector de residuos (e). Tal como demuestra Anselin (1988A, 1990A), se puede obtener un vector de pseudo-observaciones de la variable dependiente, para cada conjunto de (e_r) de los N residuos remuestrados, del modo siguiente:

$$y_r = (I - \rho_{VI}W)^{-1}(X\beta_{VI} + e_r) \quad \text{Eq. IV.2.39.}$$

siendo X el vector de variables exógenas fijas. Mediante el método VI se obtiene un estimador de ρ y β en cada conjunto de datos remuestrados, utilizando Wy_r como retardo espacial. Este procedimiento se repite un gran número de veces, por ejemplo R, generando una distribución de frecuencias empírica para los estimadores ρ y β . El estimador "bootstrap" será la media de esta distribución de frecuencias empírica.

⁵¹ Para más detalles, consultar Efron (1982), Efron y Tibshirani (1986) y Lèger *et al.* (1992).

⁵² FREEDMAN, D. y S. PETERS (1984a), "Bootstrapping a regression equation: some empirical results". *Journal of the American Statistical Association*, 79; pp. 97-106.
FREEDMAN, D. y S. PETERS (1984b), "Bootstrapping an econometric model: some empirical results". *Journal of Business and Economic Statistics*, 2; pp. 150-58.

IV.2.2.3. Modelo del error espacial

El modelo del error espacial es un caso particular de modelos con perturbación aleatoria no esférica, que no satisface los supuestos de homoscedasticidad (varianza constante) y no autocorrelación. La dependencia espacial en la perturbación aleatoria puede adoptar formas diversas. La especificación de este modelo es el MBRL con perturbación autorregresiva espacial:

$$\begin{aligned} y &= X\beta + u \\ u &= \lambda Wu + \varepsilon \end{aligned} \quad \text{Eq. IV.2.40.}$$

- X: matriz de observaciones de las variables explicativas (N,K)
 β : vector de coeficientes de la regresión (K,1)
u: vector de términos de la perturbación aleatoria (N,1)
 λ : coeficiente autorregresivo
Wu: retardo espacial de la perturbación aleatoria
 ε : perturbación aleatoria esférica, con media nula y matriz de varianzas y covarianzas $\sigma^2 I$ (homoscedasticidad y no autocorrelación).

A consecuencia de la dependencia espacial, la perturbación aleatoria no tiene ya una matriz de varianzas y covarianzas diagonal, por lo que adoptará la siguiente forma:

$$E(uu') = \Omega = \sigma^2 [(I - \lambda W)(I - \lambda W)]^{-1} \quad \text{Eq. IV.2.41.}$$

Si el coeficiente λ fuera conocido, los coeficientes de la regresión podrían estimarse a través de un modelo MCO con variables espacialmente filtradas ($y - \lambda Wy$), ($X - \lambda WX$), del modo siguiente:

$$(y - \lambda Wy) = (X - \lambda WX)\beta + u \quad \text{Eq. IV.2.42.}$$

siendo Wy, WX las variables dependiente y explicativas desplazadas. En este caso, la perturbación aleatoria (u) es esférica. Este método es denominado **mínimos cuadrados generalizados** (MCG). Sin embargo, el parámetro λ no suele ser conocido y debe ser estimado conjuntamente con el resto de coeficientes de la regresión. En el caso de autocorrelación temporal, se han desarrollado algunos métodos, como mínimos cuadrados generalizados factibles o estimados (MCGF o MCGE), como el conocido estimador Cochrane-Orcutt. Pero, debido a la simultaneidad propia de la naturaleza espacial de la dependencia, en este caso, estos procedimientos no son aplicables al caso espacial, por lo que debe llevarse a cabo un procedimiento completo de **MV**⁵³.

Como ya se puso de manifiesto anteriormente, las consecuencias de ignorar la dependencia espacial en la perturbación aleatoria (dependencia espacial residual) no son tan

⁵³ Ver una exposición técnica más extensa sobre la capacidad relativa de los diversos estimadores sugeridos en la literatura, en Anselin (1988, capítulo 8).

importantes como las de ignorar la dependencia en el retardo espacial (dependencia espacial sustantiva). El principal problema consiste en que los estimadores MCO resultan ineficientes, aunque aún permanecen insesgados. Los estimadores MV son consistentes y, en la mayoría de las ocasiones (incluyendo el caso espacial), también producen estimadores insesgados para los parámetros β de la regresión (aunque no para el llamado parámetro λ de “ruido”). Como en el caso del modelo de retardo espacial, la estimación MV del modelo del error espacial se fundamenta en la hipótesis de normalidad de las perturbaciones aleatorias. Establecido este supuesto, podrá obtenerse una función de verosimilitud, que consiste en una función de los parámetros no lineal a maximizar⁵⁴.

Como se puso de manifiesto, los estimadores de los parámetros (β) y la varianza de las perturbaciones aleatorias (σ^2) pueden expresarse en función de un valor del parámetro autorregresivo λ (en el método MCG). Una vez sustituidas estas expresiones en la función de verosimilitud, podría encontrarse una función de verosimilitud concentrada que sea sólo función del parámetro autorregresivo λ , similar la obtenida para el modelo del retardo espacial⁵⁵. Podría encontrarse un estimador MV del parámetro (λ) a través de un proceso simple de búsqueda en el intervalo aceptable $[1/\omega_{\min}$ a $1/\omega_{\max}]$, también como en el caso del retardo espacial..

En cuanto a **los contrastes de los efectos espaciales**, se han formulado, también para este modelo, el test de heteroscedasticidad de Breush-Pagan (del multiplicador de Lagrange) y el contraste del cociente de verosimilitud (LR) para el coeficiente espacial autorregresivo (λ). Además, existe también un test LR y otro de Wald para la hipótesis del factor común, que es un contraste sobre la consistencia interna de la especificación del error espacial, como se verá más adelante.

En cuanto a la autocorrelación espacial, el contraste del cociente de verosimilitud (LR) para el coeficiente espacial autorregresivo (λ), se calcula como el doble de la diferencia entre el logaritmo de verosimilitud del modelo de error espacial y el logaritmo de verosimilitud del MBRL, con el mismo conjunto de variables explicativas (siendo nulo el parámetro λ). Este contraste se distribuye como una χ^2 con 1 grado de libertad. El modelo de error espacial podría transformarse dando lugar al llamado modelo mixto autorregresivo regresivo cruzado de regresión espacial o **modelo Durbin espacial** del factor común⁵⁶ que,

⁵⁴ La verosimilitud adopta la forma propia de una perturbación aleatoria no esférica:

$$L = \sum_i \ln(1 - \lambda \omega_i) - N/2 \ln(2\pi) - N/2 \ln \sigma^2 - (y - X\beta)'(I - \lambda W)'(I - \lambda W)(y - X\beta)/2\sigma^2,$$

siendo ω_i los valores propios de la matriz de pesos y el resto de la notación como en el texto principal. Para una mayor consideración del proceso de obtención y aplicación, ver Ord (1975), Anselin (1980), Anselin (1988A, cap. 8) y Anselin y Hudak (1992).

⁵⁵ La función de verosimilitud concentrada adopta la forma: $L_c = -N/2 \ln[e'e/N] + \sum_i \ln[1 - \lambda \omega_i]$, donde $e'e$ es la suma de los cuadrados de los residuos de la regresión de la variable espacialmente filtrada y las variables explicativas: $(y - \lambda WY)$, $(X - \lambda WX)$. Para mayor detalle, ver Anselin (1980), Anselin (1988A) y Anselin y Hudak (1992).

⁵⁶ Este modelo se denomina como modelo Durbin por analogía con la especificación similar para los modelos de series temporales (para detalles técnicos, ver Burrige, 1981; Bivand, 1984; Anselin, 1988A. La propuesta de Durbin, en modelos de series temporales, consiste en trabajar con el modelo transformado de forma que,

en la nueva taxonomía de Anselin (2001E), sería también un modelo de externalidades globales no modelizadas (presentes en la perturbación aleatoria):

$$y = \lambda Wy + X\beta - \lambda WX\beta + u \tag{Eq. IV.2.43}$$

con la misma notación que anteriormente. Sin restricciones, este modelo sería:

$$Y = \lambda WY + X\beta - WX\gamma + u \tag{Eq. IV.2.44}$$

Lo que hace especial a este modelo del factor común es la restricción no lineal presente en los coeficientes del retardo espacial (λ), las variables explicativas (β) y las variables explicativas espacialmente retardadas (γ). Para ser consistentes con la formulación del error espacial, los coeficientes de los términos WX tendrían que ser iguales al valor negativo del producto del coeficiente del término Wy por los coeficientes de los términos de la matriz X. La expresión matemática de esto mismo es lo que se conoce como hipótesis del factor común:

$$H_0: \lambda \cdot \beta = \gamma \tag{Eq. IV.2.45}$$

Hay dos formas de contrastar esta hipótesis. Una de ellas, se lleva a cabo mediante el contraste de Wald sobre el conjunto de restricciones no lineales derivadas del factor común. Este test se distribuye asintóticamente como una χ^2 con tantos grados de libertad como coeficientes tenga la regresión, sin incluir el término independiente. El test se calcula a partir de una estimación MV auxiliar del modelo del retardo espacial sin restricciones. Otro modo de contrastar esta hipótesis consistiría en obtener un test del cociente de verosimilitud (LR) a partir de la maximización de la función de verosimilitud del modelo del error espacial (es decir, el modelo con restricciones) y la verosimilitud del modelo del error espacial sin restricciones. El test LR también se distribuye asintóticamente como una χ^2 con los mismos grados de libertad con el contraste de Wald.

En alguna ocasión, las variables explicativas espacialmente retardadas WX dan lugar a una multicolinealidad perfecta (por ejemplo, cuando se incluyen variables ficticias), en cuyo caso no se podría calcular el contraste sobre la hipótesis del factor común.

por ejemplo, en un modelo con tres variables explicativas en el que existe autocorrelación espacial, el modelo original se transforma del modo siguiente:

$$\begin{aligned} y_i - \rho y_{i-1} &= \beta_1(1 - \rho) + \beta_2(x_{2i} - \rho x_{2i-1}) + \beta_3(x_{3i} - \rho x_{3i-1}) + \beta_4(x_{4i} - \rho x_{4i-1}) + (u_i - \rho u_{i-1}) = \\ &= \beta_1(1 - \rho) + \rho y_{i-1} + \beta_2 x_{2i} - \rho \beta_2 x_{2i-1} + \dots \end{aligned}$$

En términos generales, este modelo se expresaría como:

$$y_i = \beta_1(1 - \rho) + \rho y_{i-1} + \beta_2 x_{2i} - \rho \beta_2 x_{2i-1} + \dots + \beta_K x_{Ki} - \rho \beta_K x_{Ki-1} + u_i$$

Este modelo debe ser estimado por MCO, aceptando como estimador de ρ el coeficiente de la variable y_{i-1} de forma que, con este valor de ρ , se transforme el modelo original, estimando por MCO los parámetros del nuevo modelo definido y calculando a partir de aquí los parámetros ($\hat{\beta}$) del modo siguiente:

$$y_i = \beta_1(1 - \hat{\rho}) + \hat{\rho} y_{i-1} + \beta_2 x_{2i} - \hat{\rho} \beta_2 x_{2i-1} + \dots + \beta_K x_{Ki} - \hat{\rho} \beta_K x_{Ki-1} + u_i$$

IV.2.3. Estimación y contrastes de modelos de heterogeneidad espacial

Como ya se ha indicado, en un modelo de regresión lineal pueden distinguirse diversas especificaciones para el efecto de heterogeneidad espacial, según que se manifieste como heteroscedasticidad (modelo del error heteroscedástico) o como inestabilidad paramétrica que puede ser, a su vez, continua (modelo de parámetros aleatorios, modelo de superficie tendencial) o discreta (modelo espacial de regresiones cambiantes). Cada una de estas formulaciones requiere de métodos propios de estimación y contraste que se exponen a continuación.

IV.2.3.1. Modelo del error heteroscedástico

El modelo del error heteroscedástico es un caso particular del llamado modelo de perturbaciones aleatorias no esféricas. La varianza de la perturbación aleatoria no es ya una constante, sino que varía con cada observación:

$$\begin{aligned} y &= X\beta + u \\ \text{Var}(u_i) &= \sigma_i^2 \\ E[uu'] &= \Omega \end{aligned} \quad \text{Eq. IV.2.1.}$$

donde y: vector de observaciones de la variable dependiente (N,1)
 X: matriz de observaciones de las variables explicativas (N,K)
 u: vector de perturbaciones aleatorias (N,1), con varianza σ_i^2 , para cada observación i, y matriz diagonal de varianzas y covarianzas Ω .

La estimación de este modelo puede llevarse a cabo a través del método de mínimos cuadrados generalizados factibles (MCGF) y máxima-verosimilitud (MV), que se exponen a continuación.

El principio que está detrás de la estimación **MCGF** del modelo del error heteroscedástico es la obtención de estimadores consistentes para los elementos de la varianza de la perturbación aleatoria (Ω). Estos estimadores podrán ser utilizados en la expresión de los estimadores MCGF:

$$b_{MCGF} = [X \Omega^{-1} X]^{-1} X \Omega^{-1} Y \quad \text{Eq. IV.2.2.}$$

Como se ha expuesto en el Apartado IV.1.2., para que este modelo sea identificable, la varianza no constante del término del error debe tener alguna estructura, como la heteroscedasticidad aditiva (en la que la varianza de la perturbación aleatoria se expresa como una función lineal de un conjunto de variables explicativas), de coeficientes aleatorios y de grupos.

Tanto para el caso de heteroscedasticidad aditiva como de coeficientes aleatorios, los estimadores MCGF pueden obtenerse a través de un procedimiento en tres etapas, sugerido por Amemiya (1977, 1985), que parte de los residuos de una regresión MCO (e). En una primera etapa, se realiza una regresión entre el cuadrado de los citados residuos (e^2), y las variables Z heteroscedásticas, que da lugar a un primer conjunto de estimadores γ_1 :

$$\gamma_1 = [Z'Z]^{-1} Z'e^2 \quad \text{Eq. IV.2.3.}$$

En una segunda etapa, los estimadores de la varianza de la perturbación aleatoria, obtenidos a partir de los estimadores γ_1 , se utilizarán para obtener unos estimadores MCGF más eficientes en una regresión del cuadrado de los residuos y las Z variables:

$$\gamma_2 = [Z'D^{-2}Z]^{-1} Z'D^{-2}e^2 \quad \text{Eq. IV.2.4.}$$

donde D es una matriz diagonal de elementos $Z \cdot \gamma_1$. En una tercera etapa, los estimadores γ_2 se utilizarán para construir un estimador consistente de la matriz Ω , adoptando como elementos de la diagonal principal de Ω el producto de $Z \cdot \gamma_2$. Esta matriz Ω será luego utilizada para obtener los estimadores MCGF definitivos.

En el caso de la heteroscedasticidad de grupos (“groupwise”) la estimación MCGF es mucho más sencilla. Los residuos de la regresión MCO se agrupan según las estructuras definidas por una variable indicador y , para cada grupo (g) la varianza de la perturbación aleatoria (σ_g^2) es estimada como $e'_g e_g / N_g$, donde e_g es el vector de residuos del grupo g , y N_g , el número de observaciones del grupo. Los estimadores de σ_g^2 se sustituyen entonces en los propios elementos de la diagonal principal de Ω para obtener los estimadores MCGF.

Cuando los términos de la perturbación aleatoria se distribuyen normalmente, el procedimiento iterativo de MCGF puede demostrarse que es equivalente a la estimación **MV**⁵⁷.

El contraste de **heteroscedasticidad** en este modelo, puede realizarse a través del test de Wald (W) y el cociente de verosimilitud (LR). En el caso de heteroscedasticidad aditiva y coeficientes aleatorios, en los que el término constante puede identificarse con la varianza homoscedástica, el test de Wald consistiría en un contraste de significación conjunta del resto de coeficientes de la especificación heteroscedástica, mientras que en el caso de heteroscedaticidad de grupos (“groupwise”), este test sería un contraste sobre la igualdad de las varianzas en cada grupo o estructura considerados. Por su parte, el test del cociente de verosimilitud (LR) se calcula como el doble de la diferencia entre el logaritmo

⁵⁷ La función matemática del logaritmo de verosimilitud para el modelo del error heterocedástico tiene la forma: $L = -N/2 [\ln 2\pi] - 0,5 \ln |\Omega| - 0,5 u' \Omega^{-1} u$, donde $|\Omega|$ es el determinante de la matriz de covarianzas de la perturbación no escalar y el resto de la notación, como anteriormente.

de verosimilitud en el modelo de error heterocedástico y el MBRL, con el mismo número de variables.

En cuanto a la dependencia espacial, este efecto puede ser estimado a través de los contrastes LM-ERR (dependencia residual) y LM-LAG⁵⁸ (dependencia sustantiva) que, como se ha indicado, son contrastes asintóticos cuya distribución sigue una χ^2 con 1 grado de libertad.

IV.2.3.2. Modelos de inestabilidad paramétrica continua

Los modelos de inestabilidad paramétrica continua más frecuentemente utilizados son el modelo de superficie tendencial y el modelo de parámetros aleatorios de Casetti, cuyos procedimientos de estimación y contraste sobre los efectos espaciales se exponen a continuación.

1. Modelo de superficie tendencial

Se trata de un modelo de regresión espacial cuyas variables explicativas son los elementos de un polinomio de las coordenadas terrestres (longitud/latitud) de las unidades espaciales (l, m). Este modelo, expresado únicamente como función de las coordenadas, puede ser estimado por MCO. Pero, en este tipo de especificación, podrían implementarse también otras formulaciones, como el modelo del retardo espacial, error espacial o error heterocedástico, debiendo emplearse, en estos casos, los métodos de estimación adecuados anteriormente expuestos.

2. Modelo de parámetros aleatorios

La variedad de la heterogeneidad espacial de los parámetros se expresa en forma de variación continua. La especificación del modelo de regresión espacial puede dar lugar a 2 tipos de problemas que merecen alguna atención.

- Cuando se utilizan muchas variables, será inevitable que se produzca un cierto grado de multicolinealidad (por ejemplo, cuando hay superficie tendencial), lo que complicaría la interpretación de los resultados de la regresión. Una de las soluciones aportadas por Casetti y Jones (1988) evita este problema reemplazando el conjunto original de variables de expansión por sus

⁵⁸ Matemáticamente, el test LM-LAG se expresa como: $LM_{LAG} = \frac{[e' \Omega^{-1} W Y]^2}{[D + Tr(W' W + W^2)]}$, siendo $D = [W X b]' \Omega^{-1} [W X b] - [W X b]' \Omega^{-1} X (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} [W X b]$, donde “e” son los residuos de la estimación MCGF, Ω es la matriz de estimadores de las varianzas de las perturbaciones aleatorias, $W X b$ son los valores de predicción espacialmente retardados y “Tr” es el operador traza de una matriz. Este test es una extensión directa del contraste del retardo espacial expuesto anteriormente.

componentes principales obtenidos a través del método llamado de expansión ortogonal.

- Otro problema consiste en la heteroscedasticidad potencial que podría producirse cuando no se especifica correctamente la expansión (Anselin, 1988 y 1992), lo que podría corregirse a través de una estimación por MCO robustos o mediante los métodos de estimación del modelo del error heteroscedástico, que no requeriría modificaciones en la expansión.

Existe un contraste especial sobre la estabilidad de los coeficientes de regresión, en forma de contraste sobre la significatividad conjunta de los coeficientes de expansión, siendo la hipótesis nula la siguiente:

$$H_0: \gamma_{1K} = \gamma_{2K} = \dots = \gamma_{MK} = 0 \quad \text{Eq. IV.2.5.}$$

para todas las variables de expansión. En la regresión por MCO, este contraste consiste en un **estadístico F**, con $[M.(K-1), N-M(K-1)]$ grados de libertad, siendo M el número de variables de expansión y K el número de variables explicativas del modelo inicial. En todos los demás métodos de estimación, este test es un test asintótico de Wald, distribuido como una χ^2 con $[N.(K-1)]$ grados de libertad. Además del contraste conjunto, también se incluye un test sobre la significación de la expansión para cada coeficiente del modelo inicial. Una vez más, este test será un estadístico F para la estimación MCO, con $(M.N-M)$ grados de libertad, y un estadístico asintótico de Wald para el resto de métodos (χ^2 con M grados de libertad).

IV.2.3.3. Modelo espacial de regresiones cambiantes

En muchos casos, la asunción de una relación fija entre las variables explicativas y la variable dependiente, que se mantenga a lo largo de toda la base de datos, no es sostenible, porque pueda producirse heterogeneidad en forma de distintos términos independientes y/o pendientes en la ecuación de regresión para subconjuntos de datos⁵⁹. Este fenómeno es, a menudo, denominado en la literatura econométrica como inestabilidad estructural o cambio estructural y puede expresarse en forma de regresiones cambiantes (“switching regressions”), planteados por primera vez por Quandt (1958). Cuando los diferentes subconjuntos de datos pertenecen a regiones o grupos de unidades espaciales, la especificación de regresiones cambiantes es denominada por Anselin (1988A) como estructuras o regímenes espaciales (“spatial regimes”).

Estos modelos pueden resolverse estimando conjuntamente los coeficientes de todas las estructuras, a partir de una matriz aumentada de observaciones de las variables independientes, de dimensión (N,MK) , siendo M el número de estructuras, mediante la transformación de cada una de las variables explicativas en tantas variables nuevas como

⁵⁹ Las diferentes ordenadas en el origen o términos independientes pueden ser fácilmente considerados con la inclusión de variables ficticias en la especificación de la regresión.

estructuras existan. Las nuevas variables tendrán valor cero para todas las observaciones no incluidas en la estructura correspondiente.

Existe un contraste añadido, el **test de Chow espacial** (Anselin, 1990) sobre la estabilidad de los coeficientes de la regresión a través de las estructuras identificadas, cuya hipótesis nula es la estabilidad de los parámetros en todas las estructuras. Por ejemplo, para el caso de 2 estructuras:

$$H_0 : \alpha_1 = \alpha_2 \text{ y } \beta_1 = \beta_2$$

Este contraste se realiza para todos los coeficientes conjuntamente, así como para cada coeficiente por separado. Se trata de un test basado en el conocido test de Chow (1960) en modelos temporales. El estadístico de Chow espacial se distribuye como una F con (K,N-MK) grados de libertad. Este contraste se basa en un estadístico de Wald (W) asintótico, distribuido como una χ^2 con (M-1*K) grados de libertad (ver, por ejemplo, White, 1980, y MacKinnon y White, 1985).

En la Tabla IV.2.1., se resumen los métodos de estimación y contrastes presentados en este Apartado.

Tabla IV.2.1. Métodos de estimación y contrastes de efectos espaciales

Modelo	Método de estimación	Contrastes de dependencia		Contrastes de heteroscedasticidad
		Retardo espacial	Residuos	
MBRL	MCO	LM-LAG	I de Moran	Breusch-Pagan (BP)
		LM-LE	Kelejian y Robinson (K-R) LM-ERR LM-EL	Koenker-Bassett (KB)
			SARMA LM-LAG λ ,	
Modelo del retardo espacial	MV	W de Wald	LM-ERR	Breusch-Pagan (BP)
	VI-MC2E	LR		
	VI-Bootstrap	LM-LAG		
Modelo del error espacial	MCG	W de Wald	LM-ERR	Breusch-Pagan (BP)
	MV	LR		
Modelo del error heteroscedástico	MCO-robustos	LM-LAG	LM-ERR	W de Wald
	MCGF			LR
	MV			
Modelo de superficie tendencial	MCO (otros, según especificación)			

Modelo	Método de estimación	Contrastes de dependencia		Contrastes de heteroscedasticidad
		Retardo espacial	Residuos	
Modelo de parámetros aleatorios	MCO MCO-robustos			Test F de inestabilidad W de Wald
Modelo de regresiones cambiantes	MCO (otros, según especificación)			Chow espacial

Fuente: Elaboración propia.

CAPÍTULO V

Predicción-extrapolación espacial de datos microterritoriales

La predicción espacial es la parte de la econometría espacial que consiste en la estimación, para determinadas localizaciones espaciales, de valores no existentes. Este proceso puede llevarse a cabo a través de la interpolación espacial, aplicada en la geoestadística, así como de la extrapolación de datos microterritoriales a partir de relaciones establecidas en un ámbito espacial agregado. En el presente Capítulo, se amplía este concepto, que se encuentra iluminado por el esquema teórico del problema de la unidad espacial modificable (MAUP) y las inferencias ecológicas, siendo también necesario un correcto tratamiento de los efectos espaciales (Apartados V.1. y V.2.). Además, se propone un proceso de realización de un ejercicio de predicción-extrapolación, mediante la estimación de la renta familiar disponible de los municipios de la Comunidad de Madrid en 2000 (Apartado V.3.).

El esquema del capítulo es el siguiente:

**CAPÍTULO V. PREDICCIÓN-EXTRAPOLACIÓN ESPACIAL DE DATOS
MICROTERRITORIALES**

- V.1. Principios básicos de la predicción-extrapolación espacial
 - V.1.1. Concepto de predicción-extrapolación espacial
 - V.1.2. El problema de la unidad espacial modificable (MAUP)
 - V.1.3. Inferencias ecológicas (IE)
 - V.1.4. Efectos espaciales y el sesgo de la agregación
 - V.1.5. Problemas y vías de solución
- V.2. Modelos de predicción-extrapolación de datos microterritoriales
- V.3. Proceso de realización de un ejercicio de predicción-extrapolación de la renta familiar disponible municipal de la Comunidad de Madrid, en 2000
 - V.3.1. Fundamentos teóricos.
 - V.3.2. Selección y tratamiento de la información disponible
 - V.3.3. Análisis exploratorio
 - V.3.4. Análisis confirmatorio
 - V.3.5. Predicción-extrapolación de datos microterritoriales
 - V.3.6. Principales resultados

V.1. PRINCIPIOS BÁSICOS DE LA PREDICCIÓN-EXTRAPOLACIÓN ESPACIAL

V.1.1. Concepto de predicción-extrapolación espacial

La econometría espacial consta, según Anselin (2001B), de los siguientes elementos constitutivos:

- Especificación de los efectos espaciales de dependencia y heterogeneidad espacial
- Contraste de la presencia de dichos efectos en las variables del modelo
- Estimación del modelo
- Predicción espacial de datos

La predicción espacial es, por tanto, considerada como la parte de la econometría espacial que tiene como misión estimar, para determinadas localizaciones, valores no existentes. Se trata de una actividad que podría ser denominada como la “alquimia de la estadística”, es decir, la creación de datos en donde no los hay (Anselin, 2000B). Aunque este autor considera la predicción espacial como equivalente a la interpolación espacial, aplicada desde hace años por la geoestadística, este concepto tendría que ampliarse para incluir también otros métodos inferenciales destinados a obtener datos microterritoriales a partir de información espacial agregada, lo que podría ser denominado como extrapolación espacial, dado que el objeto a estimar se encuentra fuera del ámbito territorial de la estimación, a diferencia de lo que sucede con las técnicas de interpolación espacial.

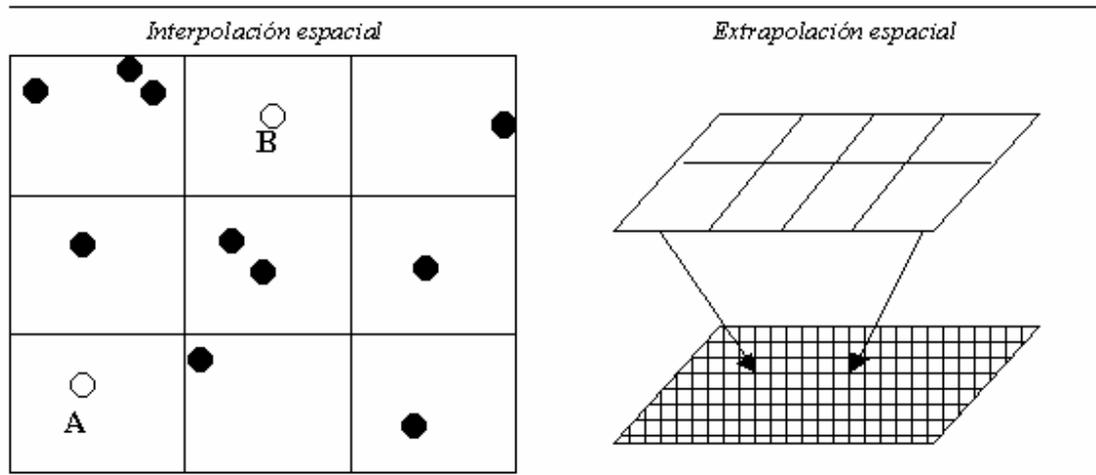
La interpolación espacial encuentra interesantes aplicaciones en el campo de la economía medioambiental y de los recursos, que suele trabajar con datos procedentes de monitorización o fotografías vía satélite correspondientes a localizaciones fijas de puntos de la atmósfera o la corteza terrestre. La necesidad de utilizar estos datos conjuntamente con información sociodemográfica, disponible para áreas (no puntos) geográficas administrativas más o menos extensas (municipios, secciones), supone la “creación”, para la primera variable, de los valores intermedios inexistentes que permitan asignar un valor a cada una de las áreas administrativas consideradas, tal como sucede con los puntos A y B en la Figura V.1.1⁶⁰. En cambio, el ejercicio de extrapolación espacial consiste en obtener, a partir de datos y relaciones procedentes de un ámbito espacial agregado, información estadística microterritorial para ámbitos inferiores o desagregados, como en el ejemplo de la citada Figura.

Aunque el tema de la predicción-extrapolación de datos espaciales pueda parecer novedoso, como ya se indicó en el Apartado II.2.4., se trata de una cuestión relacionada con los fenómenos denominados por los investigadores sociales como inferencias ecológicas (King, 1997), problema de la unidad espacial modificable (Arbia, 1989) o inferencias de

⁶⁰ Existen, además, otras interesantes aplicaciones más relacionadas con la economía como, por ejemplo, la aplicación del método del “krikeaje” a la valoración inmobiliaria para, entre otras posibilidades, realizar inferencias sobre el precio de la vivienda en una ciudad (ver Chica, 1994; Cano *et al.*, 1997).

corte transversal (Achen y Shively, 1995), también conocidos por físicos y geólogos como problema del cambio de soporte (Chilès y Delfiner, 1999) o estimaciones indirectas⁶¹, según estadísticos y econométricos (Vicéns y Chasco, 1998, entre otros). Todas estas cuestiones, no exactamente identificables, tienen en común el tratamiento y/o inferencia de relaciones estadísticas entre variables procedentes de distintos ámbitos geográficos.

Figura V.1.1. Interpolación y extrapolación espacial de datos



Fuente: Elaboración propia.

Más adelante, se expondrá que la predicción-extrapolación de datos espaciales es aún una cuestión no totalmente resuelta desde el punto de vista metodológico debido a las dificultades que plantea. Como se presentó en el Apartado II.2.2., en líneas generales, los problemas se producen cuando las observaciones de una variable, procedentes de un nivel espacial agregado (por ejemplo, provincias), no son capaces de explicar correctamente el comportamiento de dicha variable en unidades espaciales desagregadas, más cercanas al individuo (como municipios, barrios, secciones, etc.), lo que suele producirse cuando existe heterogeneidad espacial en la distribución microterritorial de dicha variable. Es más, tal como demostró Arbia (1989), algunos estadísticos, como los coeficientes de correlación lineal o autocorrelación espacial, pueden cambiar de valor (y hasta de signo) según el ámbito espacial desde el que son calculados.

⁶¹ Este concepto no es equivalente al conocido como estimaciones indirectas en pequeñas áreas que, según los estadísticos (Rao, 2000), tienen como objetivo la obtención de una información estadística completa en unidades territoriales muy pequeñas (por ejemplo, municipios de escasa población). Efectivamente, el tamaño de las muestras en este tipo de zonas raras veces es lo suficientemente amplio (o es nulo) como para tener una información correcta de las mismas, por lo que debe recurrirse a métodos indirectos de estimación. Algunos autores (Rojo y Sanz, 2001) han utilizado este procedimiento, que se fundamenta en métodos bayesianos como el muestreo de Gibbs, para realizar desagregaciones subsectoriales de una magnitud sectorial agregada, aunque apuntan también la capacidad de esta metodología para desagregar magnitudes correspondientes a zonas geográficas agregadas entre áreas más pequeñas.

Estas dificultades plantean un reto que necesita ser abordado de la mejor manera posible debido a la creciente demanda de datos espaciales microterritoriales existente en diversos campos del saber (King, 1997). En concreto, desde las ciencias sociales se están planteando cuestiones fundamentales que sólo tienen respuesta mediante ejercicios de predicción-extrapolación de datos más o menos individuales a partir de información agregada. A modo de ejemplo, se presentan, a continuación, algunos casos:

- La **ciencia política** necesita conocer el sentido del voto de ciertos colectivos más o menos minoritarios (población femenina, de raza negra, religión católica, etc.) que, cuando no se dispone (a través, por ejemplo, de encuestas personales), sólo puede obtenerse mediante un ejercicio de extrapolación de datos censales o individuales a partir de información disponible a nivel de distrito electoral (Ogburn y Goltra, 1919; King, 1997; Cho, 2001).

En otras ocasiones, son los partidos políticos los interesados en conocer las políticas concretas que deben aplicarse en diferentes ámbitos más o menos microterritoriales (municipios, distritos, barrios), lo que resulta de vital importancia para lograr la máxima sintonía entre candidatos y electorado. Esta información sólo puede conocerse a través de encuestas fiables o, en ausencia de éstas, mediante inferencias ecológicas a partir de resultados electorales procedentes de ámbitos agregados.

- Los **historiadores** desean muchas veces conocer las preferencias políticas de determinados grupos sociales en tiempos pasados, en los que no se realizaban encuestas entre la población con capacidad de voto. También en estos casos, a partir de los resultados electorales, podrían inferirse datos más o menos individuales. Por ejemplo, se han realizado estudios para detectar los colectivos concretos (clases sociales, grupos políticos, religiosos, etc.) causantes del triunfo del partido nazi en las elecciones generales alemanas de 1930, a partir de resultados electorales y variables sociodemográficas disponibles para los 1.200 distritos electorales de aquel tiempo (Childers, 1983).
- Además, los **sociólogos y políticos** suelen inferir a ámbitos geográficos desagregados, la relación existente entre variables como la tasa de delincuencia y ciertos indicadores socioeconómicos causantes de la misma (tasa de paro, población de raza negra, divorcios, hogares monoparentales, etc.). Muchas veces, no es posible contar, con la suficiente extensión, de datos microterritoriales sobre actos delictivos, por lo que no queda más remedio que realizar ejercicios de predicción-extrapolación de los mismos a partir de información procedente, por ejemplo, de censos y padrones (Sampson, 1987).
- En el campo de la **sanidad**, resulta necesario muchas veces conocer el riesgo que ciertos colectivos padecen de contraer determinado tipo de enfermedades graves, como el cáncer, debido a la presencia en el ambiente de elementos radioactivos o contaminantes. En muchos de estos casos, podría resultar éticamente inaceptable o económicamente prohibitivo realizar experimentos aleatorios a gran escala entre la población afectada, por lo que no queda más remedio que acudir a procedimientos

de predicción-extrapolación que, a partir de datos territoriales agregados (regiones, provincias) permitan obtener información de ámbito microterritorial (Lubin, 1994).

- En cuanto al **campo económico**, el problema general de las inferencias ecológicas, es decir, la extrapolación de relaciones procedentes de ámbitos geográficos agregados a unidades microterritoriales, tiene mucho que ver con los estudios destinados a clarificar las relaciones entre la macroeconomía y el comportamiento microeconómico de consumidores y empresas (Klein, 1946), no siempre directas ni claras (Theil, 1954). Además, resulta de especial interés, tanto para las empresas como las instituciones públicas conocer el comportamiento microterritorial de algunos fenómenos económicos relevantes, como la producción interior (total y sectorial), las cuentas de los hogares (renta, consumo, ahorro), sólo disponibles para ámbitos geográficos agregados, nacional, regional y, en algunos casos, provincial (Vicéns y Chasco, 1998).
- Por último, el **geomarketing** incluye también la predicción-extrapolación como una técnica fundamental, por ejemplo, para conocer el consumidor-tipo de una empresa o producto comercial. Efectivamente, a partir de los datos de ventas, muchas veces sólo disponibles en las empresas por municipios, podrían extrapolarse estos datos a ámbitos territoriales inferiores (distritos, barrios, secciones) a partir de relaciones entre esta variable y otros indicadores sociodemográficos relevantes, como la tasa de paro, índice de infancia, tasa de hogares monoparentales, etc.

Estos casos son sólo algunos ejemplos de las muchas circunstancias que podrían demandar ejercicios de predicción-extrapolación de datos espaciales, lo que pone de manifiesto la importancia del tema, así como la necesidad de superar dos grandes escollos muy relacionados entre sí, la falacia ecológica y el MAUP (problema de la unidad espacial modificable). La falacia ecológica llama la atención sobre la posible falsedad que pudieran entrañar ciertas estimaciones o correlaciones, realizadas en ámbitos geográficos agregados, cuando son inferidas a unidades microterritoriales (municipios, secciones) o, incluso, indivisibles (personas, empresas, sindicatos, etc.). Por su parte, el MAUP tiene que ver con las diferentes formas y criterios con que puede dividirse un espacio geográfico, así como con las desviaciones significativas experimentadas por algunos estadísticos cuando son calculados en diversos ámbitos geográficos, con mayor o menor agregación espacial.

Dada la gran relación existente entre la predicción-extrapolación de datos espaciales y las cuestiones conocidas como problema de la unidad espacial modificable e inferencias ecológicas, se expone a continuación un resumen de los conceptos más importantes puestos de manifiesto por los dos temas. En realidad, estos dos enfoques, que están destinados a conocer el comportamiento más o menos individual a partir de relaciones (económicas, políticas, sociales, etc.) establecidas en un ámbito territorial agregado, estarían dentro de lo que la econometría clásica denomina como análisis estructural, por lo que no se trata tanto de un análisis de predicción, que es lo que se propone aquí. Sin embargo, los conceptos teóricos desarrollados tanto por el problema de la unidad espacial modificable como por las inferencias ecológicas son fundamentales para abordar, lo más correctamente posible, el problema de la predicción-extrapolación de datos espaciales. Por eso, constituyen el punto de partida ineludible en este Capítulo.

V.1.2. El problema de la unidad espacial modificable (MAUP)

Los datos espaciales con los que trata la ciencia regional se caracterizan por ser agregados de información individual, como sucede, por ejemplo, con la población de un país, que es la suma de los habitantes del mismo, o la renta per cápita provincial, que es la media de la renta de los hogares localizados en dicha provincia. Además, estos datos individuales suelen agruparse en unidades espaciales superiores, la mayor parte de las veces de carácter político-administrativo (como municipios, provincias, comunidades autónomas, etc.), diferentes en tamaño y forma. Estas arbitrariedades en el modo de agregación y las irregularidades de forma y tamaño, propias de los datos geográficos, fueron denominadas por Openshaw y Taylor (1979) como Problema de la Unidad Espacial Modificable o MAUP (“Modifyable Areal Unit Problem”).

El MAUP suele producirse en situaciones en las que un proceso debe expresarse en un ámbito espacial diferente de la escala concreta en que fue medido, lo que se presenta, a menudo, en estudios que combinan fenómenos físicos y económicos (Anselin, 2001A). Por ejemplo, en el análisis de los factores socioeconómicos que subyacen al proceso de deforestación tropical (Wood y Skole, 1999), la variable dependiente, la deforestación, se calcula a partir de la interpretación de imágenes vía satélite que dan lugar a datos correspondientes a una cuadrícula regular o “pixels”. Se trata de un enrejado cuadrangular que no se corresponde, ni en el tamaño ni en las fronteras geográficas, con las unidades administrativas de las que proceden las variables estadísticas de censos o encuestas. Como normalmente las unidades administrativas son, en estos casos, las de mayor tamaño, son también las que determinan el mínimo ámbito geográfico de análisis, lo que conlleva la agregación (o alisado) de los valores “pixel” y la pérdida de variabilidad espacial en cada unidad geográfica finalmente utilizada. Y ya se sabe que, cuanto mayor es la variabilidad (es decir, cuanto menor es la escala espacial en la que se trabaja), menos precisa será la información agregada como estimador de la variable dependiente. Es más, la variabilidad interna no siempre permanece constante entre unidades, lo que origina problemas de heteroscedasticidad. A su vez, el MAUP puede dar lugar a inferencias ecológicas cuando la información individual procedente de los “pixel” debe agregarse a ámbitos territoriales superiores, como las secciones censales (máximo nivel de disponibilidad de los datos procedentes de censos y padrones) para utilizarse conjuntamente como variables de un modelo de regresión. Es decir, el MAUP y las inferencias ecológicas suelen ser temas profundamente relacionados.

La denominación “unidad espacial modificable” se remonta a los trabajos de Jule y Kendall (1950), que distinguen entre datos estadísticos divisibles (en unidades inferiores) e indivisibles. Así, por ejemplo, la última unidad de análisis de la variable de renta sería el individuo, para el que, según los citados autores, la renta adoptaría un “valor numérico único y no modificable”, dado la imposibilidad de dividir dicho agente económico en unidades menores. Asimismo, otras variables económicas como la producción o los precios adoptan un valor único no modificable en una empresa individual o un bien particular, respectivamente.

En la mayoría de los casos, la información estadística procede de ámbitos territoriales superiores a la unidad espacial propia de un fenómeno socioeconómico, como sucede, por ejemplo, con los datos de renta bruta disponible que el INE proporciona para el ámbito provincial como suma de la renta de los hogares correspondientes a cada provincia (Anselin, 2001A). Es decir, las variables suelen estar referidas a unidades espaciales divisibles y, por tanto, “modificables”. Según Openshaw y Taylor (1979), en un mismo conjunto de datos es posible obtener diferentes resultados estadísticos según el ámbito territorial considerado (municipio, provincia, región,...) o el criterio de agregación seguido (provincias administrativas, comarcas históricas, partidos judiciales, áreas comerciales, diócesis de la Iglesia Católica, etc.).

En esta línea, Arbia (1989) presenta el MAUP como un doble problema, de escala (tamaño) y de agregación (forma) de datos espaciales:

- Cuando existe, en un análisis, algún tipo de imposición previa (exógena) para la adopción de una determinada jerarquía⁶² espacial, como suele ser la división político-administrativa de un país (en España: municipio-provincia-comunidad autónoma), se produce un problema de escala.
- Por el contrario, cuando el propósito de un análisis es precisamente la obtención de la jerarquía espacial óptima para un caso dado (división político-administrativa, histórica, judicial, económica, religiosa, etc.), se trata de un problema de agregación espacial.

V.1.2.1. Problema de la escala

El problema de la escala o nivel de resolución de los datos espaciales consiste, según Arbia (1989), en la variación que puede experimentar el análisis estadístico de una variable espacial según la escala o ámbito territorial de referencia. Así, por ejemplo, la agregación de datos microterritoriales (individuos, secciones, municipios) a una escala geográfica superior (provincias, regiones) tiene, al menos, las siguientes consecuencias sobre los estadísticos básicos de una variable medida en el ámbito agregado:

- a) Disminución de la varianza.
- b) Disminución del grado de autocorrelación espacial.
- c) Aumento del coeficiente de correlación lineal entre dos variables.
- d) Cambios en la estacionariedad espacial.

Efectivamente, las relaciones entre dos variables pueden variar, incluso de signo, cuando se consideran ámbitos geográficos diferentes, como sucede en la Tabla V.1.1 entre el impuesto de actividades económicas por habitante (variable “proxi” del PIB per capita) y las variables tasa de paro y teléfonos por habitante. Incluso, dentro del ámbito municipal se producen diferencias en los coeficientes de correlación lineal, según el tamaño poblacional

⁶² La división territorial adopta frecuentemente jerarquías anidadas en las que cada unidad espacial de ámbito inferior se encuentra contenida en una (y sólo una) unidad espacial de ámbito superior.

considerado, siendo más parecidas las relaciones del grupo de grandes municipios respecto de los ámbitos superiores regional y provincial. Se observa que puede no existir una estabilidad en las relaciones económicas cuando se consideran distintos niveles de agregación, produciéndose, como en la tasa de paro, diferencias de signo que contradicen seriamente la teoría económica.

Tabla V.1.1. Coeficientes de correlación del impuesto IAE por habitante

	Ámbito							
	Regional	Provincial	Municipal				Más cien mil hab.	
			Total municip.	Menos cien hab.	Menos mil hab.	Más mil hab.		Más diez mil hab.
Tasa de paro	-0,73	-0,58	0,02	0,01	0,02	-0,09	-0,23	-0,40
Teléfonos por hab.	0,77	0,66	0,07	0,03	0,03	0,25	0,55	0,76

Fuente: Elaboración propia.

Yule y Kendall (1950) demostraron que, en la práctica, el coeficiente de correlación entre dos variables espaciales tiende a aumentar con la escala o ámbito geográfico, tal como se pone de manifiesto en las variables de la Tabla V.1.1., según se consideren en el ámbito municipal, provincial o regional. Es más, tal como reza el título de un conocido artículo, “A million of so correlation coefficients” (Openshaw y Taylor, 1979), sería posible obtener cualquier valor del intervalo [-1,1] para el coeficiente de correlación entre dos variables, tan solo agrupando las unidades espaciales en ámbitos territoriales de diferente tamaño. Esta conclusión resulta del todo preocupante por cuanto no sería posible responder, por ejemplo, si existe alguna correlación real entre un indicador de actividad económica per capita y la tasa de paro en España.

Ambos autores explicaron este hecho con un ejemplo, en el que se consideran dos variables, X e Y, compuestas, cada una de ellas, por la suma de dos elementos, uno sistemático y otro aleatorio:

$$\begin{aligned} X &= Z + E \\ Y &= W + F \end{aligned} \quad \text{Eq. V.1.1.}$$

siendo Z, W: componentes sistemáticos de X, Y, respectivamente

E, F: componentes aleatorios de X, Y, respectivamente.

Asimismo, la **varianza** de dichas variables estará también compuesta de la suma de dos elementos, del modo siguiente:

$$\begin{aligned} \text{Var}(X) &= \text{Var}(Z) + \text{Var}(E) \\ \text{Var}(Y) &= \text{Var}(W) + \text{Var}(F) \end{aligned} \quad \text{Eq. V.1.2.}$$

Según estos autores, la agregación de unidades espaciales produce un alisado o compensación en los valores de X, Y, por lo que la variabilidad de sus correspondientes elementos aleatorios (E, F) disminuirá, lo que provocará también un decremento en la varianza de las variables X, Y. De esta manera, el coeficiente de **correlación** lineal de las variables X, Y experimentará incrementos con sucesivas agregaciones de datos espaciales, ocasionados éstos por la disminución en el denominador de la citada expresión:

$$r_{x,y} = \frac{Cov(X,Y)}{Var(X) \cdot Var(Y)} \quad Eq. V.1.3.$$

Cliff y Ord (1981) también estudian este tema y llegan a la conclusión de que el grado de correlación entre dos variables es función del tamaño de la escala geográfica considerada. Además, estos autores analizan la influencia que tiene el ámbito geográfico, más o menos desagregado, sobre la **dependencia espacial** en una variable, concluyendo que cuanto más pequeñas sean las áreas geográficas (menor agregación), mayor será el grado de dependencia espacial en una variable dada. La causa de esta relación se encontraría, de nuevo, en el mayor grado de dispersión (e interacción) existente en las unidades espaciales microterritoriales respecto de áreas más agregadas, en las que los valores de una variable tienden a compensarse en torno a un valor medio.

Otra cuestión relacionada con el problema de la escala es la posible ausencia de **estacionariedad espacial** en una variable, al pasar de un ámbito geográfico desagregado a otro más agregado, incluso aunque dicha variable fuera estacionaria en la escala inferior. Este efecto se ha demostrado que suele producirse cuando se trabaja con áreas geográficas irregulares, de diferente tamaño y forma, que es el caso más frecuente en las ciencias sociales. En estas situaciones, los resultados de cualquier análisis estadístico están muy condicionados por el nivel de resolución o escala territorial considerada. Tagashira y Okabe (2002) demuestran que la perturbación aleatoria de un modelo de regresión estimado con datos agregados, presenta heteroscedasticidad (o ausencia de estacionariedad en varianza), incluso en el caso de ser homoscedástica en el modelo con datos desagregados. Efectivamente, sea el siguiente modelo de regresión para un ámbito espacialmente desagregado:

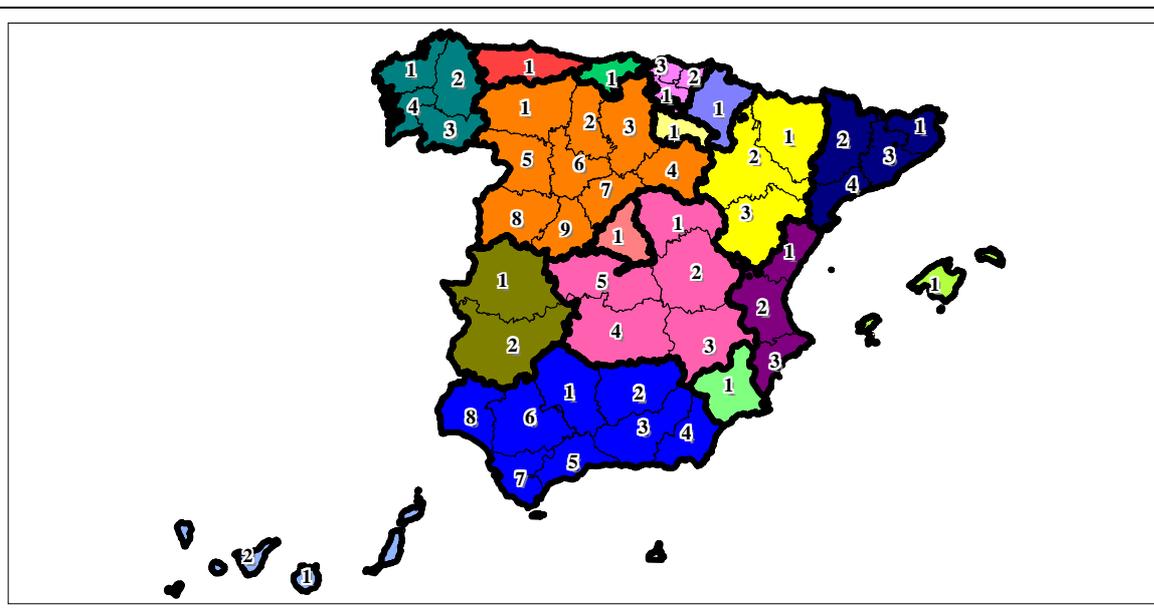
$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad Eq. V.1.4.$$

donde y_i : valor de la variable endógena en la unidad espacial desagregada i
 $i = 1, \dots, N$ unidades territoriales desagregadas.
 x : variable exógena
 β_0, β_1 : parámetros a estimar
 u : perturbación aleatoria esférica.

Las N unidades espaciales microterritoriales pueden ser agregadas en M regiones, correspondientes a una escala superior. Es decir, se considera que el sistema espacial se encuentra estructurado a modo de jerarquía anidada que, como ya se ha expuesto, suele producirse en la organización político-administrativa de los estados. Así, por ejemplo,

España presenta una estructura similar entre comunidades autónomas y provincias, como se aprecia en la Figura V.1.2., de forma que las 50 provincias ($N=50$) se encuentran agregadas en 17 comunidades autónomas ($M=17$).

Figura V.1.2. Jerarquía anidada de unidades espaciales en España



Fuente: Elaboración propia.

El modelo anterior adoptaría, en el nivel agregado, la forma siguiente:

$$y_k = \beta_0 + \beta_1 x_k + u_k \quad \text{Eq. V.1.5.}$$

donde k : unidad espacial agregada, $k = 1, \dots, M$ unidades territoriales agregadas.

$$y_k = \frac{\sum_{i=1}^{N_k} y_i}{N_k} ; \quad x_k = \frac{\sum_{i=1}^{N_k} x_i}{N_k} ; \quad u_k = \frac{\sum_{i=1}^{N_k} u_i}{N_k}$$

N_k : número de unidades espaciales microterritoriales contenidas en la unidad regional agregada k ; por ejemplo, en la Figura V.1.2., la comunidad de Castilla y León contiene 9 provincias ($N_k=9$).

La perturbación aleatoria de este segundo modelo con datos agregados, u_k , ya no es esférica pues aunque su valor medio sigue siendo nulo y carece de autocorrelación espacial, sí presenta problemas de heteroscedasticidad, como se demuestra a continuación:

$$\begin{aligned}
 E[u_k] &= \frac{1}{N_k} \sum_{i=1}^{N_k} E[u_i] = 0 \\
 \text{Var}[u_k] &= \frac{1}{N_k^2} \sum_{i=1}^{N_k} \text{Var}[u_i] = \frac{1}{N_k} \sigma^2 \\
 \text{Cov}[u_k, u_h] &= \frac{1}{N_k^2 \cdot N_h^2} \sum_{i=1}^{N_k} \sum_{j=1}^{N_h} \text{Cov}[u_i, u_j] = 0
 \end{aligned}
 \tag{Eq. V.1.6.}$$

Por ese motivo, los autores aconsejan estimar el modelo agregado por mínimos cuadrados ponderados, utilizando como peso el número de unidades microterritoriales, N_k , existente en cada agregado espacial. La solución de la ponderación por el tamaño de las unidades territoriales agregadas se remonta a los trabajos de Robinson (1950) y, aunque no es totalmente satisfactoria, está muy aceptada por los investigadores sociales (Arbia, 1989; Fotheringham *et al.*, 1997)⁶³. En concreto, Arbia (1989) propone una solución particular para aquellas variables con problemas de autocorrelación espacial que consiste en considerar, no sólo el tamaño de las unidades agregadas, sino también la forma de las mismas (las interacciones espaciales), así como la función de autocorrelación espacial existente entre las observaciones en el nivel microterritorial. Este autor propone reducir al estado estacionario los procesos espaciales agregados mediante transformaciones adecuadas, como sucede en el campo temporal.

Openshaw y Taylor (1979), tras múltiples esfuerzos por llegar a alguna formulación general sobre el MAUP, reconocen que no es posible aún, es decir, que cada caso de unidad espacial modificable debe de ser tratado de forma individual y específica. Esta conclusión sigue siendo válida todavía hoy, aunque con algunos matices, como se verá a continuación.

V.1.2.2. Problema de la agregación

El problema de la agregación espacial es la segunda manifestación del MAUP, que aflora en estudios destinados a la obtención de una jerarquía espacial determinada (político-administrativa, histórica, económica, etc.). En este caso, la variabilidad de las medidas estadísticas no es debido tanto a diferencias en el tamaño de las áreas geográficas (mayor o menor agregación), cuanto a su forma. Es decir, la cuestión ahora estriba en el criterio de agrupación de las unidades microterritoriales, que en el caso de los municipios españoles podría ser en provincias civiles, comarcas históricas, partidos judiciales, áreas comerciales, diócesis o provincias eclesiásticas, etc., dando lugar a diferentes tipos de agregación municipal. Es decir, en esta situación no se producirá, en una variable dada, una pérdida de información, como sucede cuando se produce un aumento de escala, sino una alteración de la misma. Dado que este tema tiene menos incidencia en el proceso de extrapolación espacial, no será abordado en profundidad.

⁶³ Ésta es también una de las soluciones propuestas para la estimación de renta familiar disponible municipal, por parte de los autores de un estudio patrocinado por el Instituto de Estadística de Andalucía (1999) y que se expondrá con mayor detalle en el Apartado siguiente.

V.1.3. Inferencias ecológicas (IE)

El tema de las inferencias ecológicas está muy relacionado y, a su vez, se deriva del MAUP, por lo que tiene también grandes repercusiones en el proceso de la predicción-extrapolación de datos microterritoriales. Efectivamente, el problema de la escala puede dar lugar a inferencias espúrias, que es lo que se ha dado en llamar “falacia ecológica”, cuando se infieren relaciones entre variables, obtenidas en un ámbito territorial superior, a ámbitos microterritoriales. Como demuestra el MAUP (Robinson, 1950), las relaciones estadísticas entre variables obtenidas en un nivel agregado nunca pueden identificarse directamente con las relaciones existentes entre las mismas variables en ámbitos geográficos inferiores.

Se entiende por inferencia ecológica (IE) aquel proceso por el que, a partir de datos agregados (es decir, ecológicos), se estiman datos microterritoriales o individuales no disponibles (King, 1997). Esta cuestión surge por primera vez en 1919, cuando un grupo de investigadores se propone conocer el sentido del voto femenino en los EEUU, en un momento en el que se había aprobado el derecho al voto de la mujer en todo el país (Ogburn y Goltra, 1919). El voto secreto hacía imposible diferenciar, en un mismo distrito electoral, el voto femenino del masculino. Hoy día, pese a los esfuerzos del AEDE por descubrir estructuras espaciales a partir de datos individuales, la IE es aún el único procedimiento de estimación para este tipo de información estadística cuando no existen datos directos.

Las IE plantean un problema, aún sin resolver definitivamente: la imposibilidad de verificación de la acuracidad de los valores de predicción. Éste es también el principal problema en la predicción-extrapolación de datos espaciales que busca conocer los valores que adopta una variable en unidades microterritoriales, a partir de la estimación de dicha variable en un ámbito territorial superior. Obviamente, si los datos microterritoriales estuvieran disponibles, no sería necesario realizar esta estimación (Anselin, 1999A).

El problema de la IE constituye actualmente una materia de debate controvertida, e incluso amarga, en la literatura de la estadística aplicada y la ciencia política, sobre todo, a raíz de la publicación de King (1997), en la que se asegura haber encontrado una solución para esta cuestión. A partir de entonces, desde diversos frentes, se han sucedido abundantes críticas (a favor y en contra), así como respuestas formuladas por este profesor de ciencia política a sus detractores (por ejemplo, Burden y Kimball, 1998; Cho, 1998, 2001; Rivers, 1998; Freedman *et al.*, 1999; King, 1999, 2000; King *et al.*, 1999; Anselin, 2000B; Anselin y Cho, 2002A y B; Sui, 2000; Davies, 2001; Johnston y Pattie, 2001; McCue, 2001), no existiendo aún un completo acuerdo sobre este tema.

V.1.3.1. Concepto de IE

El problema de las IE, a diferencia del MAUP, consiste en los diferentes resultados obtenidos por algunos estadísticos cuando son calculados para una muestra de individuos, en lugar de unidades espaciales agregadas (secciones, municipios, etc.). Este problema fue

puesto de manifiesto, por primera vez, por Robinson (1950) que distinguía entre dos tipos de correlaciones, individual y ecológica:

- La **correlación individual** es la que se produce entre objetos estadísticos indivisibles lo que, en términos del MAUP serían las unidades no modificables.
- En la **correlación ecológica**, el objeto estadístico es un grupo de personas.

Al igual que en el MAUP, toda correlación individual (puntos) inferida a partir de datos procedentes de áreas (polígonos⁶⁴), como los municipios o secciones censales, podría estar fuertemente sesgada, tanto en la cuantía como en el signo. Sin embargo, la necesidad de descubrir el comportamiento (desconocido) de unidades individuales (residentes, votantes, consumidores, empresas, etc.) es lo que motiva la realización de este tipo de inferencias, que encuentran su lugar más genuino en el campo de las ciencias sociales⁶⁵, especialmente la ciencia política, geografía y economía, cada una de las cuales ha abordado el tema desde su peculiar punto de vista, como se verá a continuación.

1. Las IE en la ciencia política

El ejemplo clásico de este tipo de inferencia procede de la ciencia política, donde se han realizado estudios sobre el comportamiento en las elecciones de algunos colectivos, a partir de información estadística sobre los porcentajes de voto en los distritos electorales. King (1997) presenta esta situación a través de los resultados obtenidos por los partidos demócrata y republicano en las elecciones presidenciales del Estado de Ohio en 1990. Se plantea una inferencia ecológica que, partiendo de relaciones establecidas entre los datos existentes a nivel de distrito y sección electoral, sea posible conocer el comportamiento individual de los colectivos de raza blanca y negra (incluyéndose aquí todas las personas de diferente raza que, en este estado, son en su mayoría de raza negra). Para ello, se cuenta con los porcentajes de voto emitidos por cada colectivo, así como los obtenidos por cada partido.

El problema de las IE, tal como se presenta en la Tabla V.1.2., estriba en la estimación de los valores correspondientes a los signos de interrogación expuestos en el centro, a partir de la información situada en los márgenes de la tabla de tabulación cruzada que, precisamente porque suele presentarse en ese lugar de la misma, suele conocerse con el nombre de “los márgenes”. En la mayor parte de la literatura científica sobre este tema, tanto sociólogos, como políticos, geógrafos y algunos estadísticos presentan el problema de las IE a través de una tabla similar a ésta. Por ejemplo, el signo de interrogación de la celda superior izquierda de la tabla correspondiente al distrito 42, representa el (desconocido) número de negros que votaron en ese distrito por el candidato demócrata (también de raza

⁶⁴ En el Apartado II.2.1., se definen los “puntos” como aquellos objetos espaciales determinados por las coordenadas terrestres de latitud y longitud, pudiendo corresponderse con individuos, empresas, etc., mientras que los “polígonos” son figuras planas conectadas por distintas líneas u objetos cerrados que cubren un área determinada, como es el caso de países, provincias o secciones censales.

⁶⁵ Este tema tiene también importantes aplicaciones en fenómenos físicos y geológicos (ver Chilès y Delfiner, 1999).

negra). Evidentemente, podrían estimarse muchos valores para esta casilla sin contradecir los márgenes de sus correspondientes fila y columna (en este caso, cualquier número entre 0 y 19.896), para lo que podría recurrirse a algún tipo de operación lógica, como el llamado “método de los límites” o cualquier otro tipo de ajuste.

Tabla V.1.2. Planteamiento del problema de la IE en las elecciones presidenciales del Estado de Ohio (EEUU)

Distrito 42 del Estado de Ohio					
Raza de los electores:	Decisión de voto			Total votos	% votantes
	Demócratas	Republicanos	Abstención		
Negra	?	?	?	55.054	68%
Blanca	?	?	?	25.706	32%
Total votos	19.896	10.936	49.928	80.760	
% votos	25%	14%	62%		

Sección P del distrito 42 del Estado de Ohio					
Raza de los electores:	Decisión de voto			Total votos	% votantes
	Demócratas	Republicanos	Abstención		
Negra	?	?	?	221	31%
Blanca	?	?	?	484	69%
Total votos	130	92	483	705	
% votos	18%	13%	69%		

Fuente: Elaboración propia a partir de King (1997).

Además de los datos a nivel de distrito, también se cuenta con información para el ámbito inferior de la sección, como se expone también en la Tabla para la sección P, una de las 131 secciones del distrito 42. Obviamente, la suma de los datos de los márgenes, a nivel de sección, para cada concepto, deberá ser igual al valor correspondiente a nivel de distrito, es decir, si al número de votantes de raza negra del partido demócrata, en la sección P, se le suman los votantes negros de este mismo partido en las otras 130 secciones, se obtendrá el total de votantes demócratas de raza negra del distrito 42.

Ante situaciones como la descrita, muchos investigadores se han concentrado en el análisis de las posibles regularidades existentes entre los datos censales de los márgenes, a nivel de sección, para ayudar así a reducir el recorrido de posibles resultados en el ámbito de los distritos electorales. Por ejemplo, si el candidato demócrata recibiera la mayor parte de los votos de los distritos con altas tasas de población americana de raza negra, podría ser razonable, desde un punto de vista intuitivo, suponer que los negros están votando de forma desproporcionada por los demócratas (por lo que el valor de la celda superior izquierda de la primera tabla sería probablemente muy alto, cercano a 19.896). Este supuesto es, a menudo razonable, aunque Robinson (1950) demostró que podría ser totalmente falso, ya que las relaciones establecidas entre unidades individuales (votantes) pueden ser de signo contrario a las obtenidas de sus correspondientes agregados (secciones, distritos), como sucedería, por ejemplo, si la población de raza blanca residente en zonas de gran mayoría negra concentraran más su voto sobre el partido demócrata que los blancos residentes en barrios de mayoría blanca.

Dado que uno de los problemas de las IE, así como sucedía con el MAUP, es la imposibilidad de contrastar directamente los resultados obtenidos, debería procurarse información complementaria procedente, por ejemplo, de encuestas realizadas en otro momento o lugar. En cualquier caso, King (1997) advierte del peligro de aceptar, sin más, cualquier resultado que sea técnicamente posible o lógicamente viable.

A modo de resumen, podría concretarse que las IE tienen el peligro de ser falsas debido a dos problemas:

- a) El **sesgo de la agregación**, que consiste en la inevitable pérdida de información que se produce en todo proceso de agrupamiento de datos individuales en unidades superiores, ya que, al no ser el mismo tipo de pérdida en cada caso, dará lugar a un sesgo determinado en cualquier tipo de inferencia que, desde este ámbito, pretenda estimar datos individuales.
- b) La **heterogeneidad espacial** presente en los datos agregados procedentes de información individual es un problema derivado del anterior sesgo, que no debe negarse y al que se debe responder adecuadamente.

2. Las IE en la geografía

Los geógrafos se han concentrado más en el análisis de las diferencias existentes en las relaciones obtenidas desde diversos ámbitos geográficos. Éste sería el caso del conocido experimento de Robinson (1950), en el que se obtuvieron los coeficientes de correlación lineal entre la tasa de analfabetismo y dos variables relacionadas con la misma (tasa de natalidad y tasa de población de raza negra), en tres ámbitos: estados, secciones censales e individuos. Los resultados obtenidos pusieron de manifiesto diferencias similares a las apreciadas en el MAUP, tanto respecto a la cuantía del coeficiente de correlación, mayor cuando mayor es la escala territorial considerada (y menor el número de observaciones), e incluso de signo contrario, en el ámbito individual.

Duncan *et al.* (1961), siguiendo el esquema de Robinson (1950), proponen una demostración matemática que explique las anteriores diferencias en los coeficientes de correlación lineal a partir del esquema de un ANOVA anidado, en el que las unidades espaciales agregadas actuarían como efectos discriminantes de los datos individuales. Así, por ejemplo, dadas dos variables (X,Y) a relacionar, en desviaciones a la media, pueden definirse las diferentes sumas cuadráticas del modo siguiente:

- Suma total de los cuadrados de las variables X, Y:

$$\begin{aligned} SX_T &= \sum_{k=1}^M \sum_{i=1}^{N_k} (x_{ik} - \bar{x}_{..})^2 \\ SY_T &= \sum_{k=1}^M \sum_{i=1}^{N_k} (y_{ik} - \bar{y}_{..})^2 \end{aligned} \quad \text{Eq. V.1.7.}$$

siendo x_{ik}, y_{ik} : valores de las variables X, Y para el individuo i perteneciente a la unidad espacial k

i : unidad individual: $i = 1, \dots, N$

k : unidad espacial agregada: $k = 1, \dots, M$

N_k : número de individuos (i) incluidos en la unidad espacial agregada k :

$$\sum_{k=1}^M N_k = N$$

$\bar{x}_{..}, \bar{y}_{..}$: media de las variables X, Y para el total de individuos, tal que:

$$\bar{x}_{..} = \frac{\sum_{k=1}^M \sum_{i=1}^{N_k} x_{ik}}{N} \quad ; \quad \bar{y}_{..} = \frac{\sum_{k=1}^M \sum_{i=1}^{N_k} y_{ik}}{N}$$

- Suma intra-áreas de los cuadrados de las variables X, Y:

$$\begin{aligned} SX_I &= \sum_{k=1}^M \sum_{i=1}^{N_k} (x_{ik} - \bar{x}_{.k})^2 \\ SY_I &= \sum_{k=1}^M \sum_{i=1}^{N_k} (y_{ik} - \bar{y}_{.k})^2 \end{aligned} \quad \text{Eq. V.1.8.}$$

$\bar{x}_{.k}, \bar{y}_{.k}$: media de las variables X, Y para cada unidades agregada k :

$$\bar{x}_{.k} = \frac{\sum_{i=1}^{N_k} x_{ik}}{N_k} \quad ; \quad \bar{y}_{.k} = \frac{\sum_{i=1}^{N_k} y_{ik}}{N_k}$$

- Suma entre-áreas de los cuadrados de las variables X, Y:

$$SX_E = \sum_{k=1}^M N_k (\bar{x}_{.k} - \bar{x}_{..})^2 \quad \text{Eq. V.1.9.}$$

$$SY_E = \sum_{k=1}^M N_k (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2$$

De forma similar, podrían definirse diversos coeficientes de correlación lineal, total (T), intra-áreas (I) y entre-áreas o ecológico (E):

- Coeficiente de correlación lineal total:

$$r_T = \frac{\sum_{k=1}^M \sum_{i=1}^{N_k} (x_{ik} - \bar{x}_{\cdot\cdot})(y_{ik} - \bar{y}_{\cdot\cdot})}{\sqrt{SX_T} \cdot \sqrt{SY_T}} \quad \text{Eq. V.1.10.}$$

- Coeficiente de correlación lineal intra-áreas:

$$r_I = \frac{\sum_{k=1}^M \sum_{i=1}^{N_k} (x_{ik} - \bar{x}_{\cdot k})(y_{ik} - \bar{y}_{\cdot k})}{\sqrt{SX_I} \cdot \sqrt{SY_I}} \quad \text{Eq. V.1.11.}$$

- Coeficiente de correlación lineal entre-áreas o ecológico:

$$r_E = \frac{\sum_{k=1}^M N_k (\bar{x}_{\cdot k} - \bar{x}_{\cdot\cdot})(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})}{\sqrt{SX_E} \cdot \sqrt{SY_E}} \quad \text{Eq. V.1.12.}$$

De las expresiones anteriores, es posible derivar la siguiente relación entre los tres coeficientes de correlación lineal:

$$r_E = r_T \cdot k_1 - r_I \cdot k_2 \quad \text{Eq. V.1.13.}$$

siendo: $k_1 = \frac{SX_T}{SX_E} \cdot \frac{SY_T}{SY_E}$

$$k_2 = k_1 \cdot \sqrt{1 - \frac{SX_E}{SX_T}} \cdot \sqrt{1 - \frac{SY_E}{SY_T}}$$

De esta expresión, pueden extraerse algunas conclusiones:

- La correlación ecológica o entre-áreas (r_E) es la diferencia ponderada entre la correlación total (r_T) y la media de las correlaciones individuales intra-áreas (r_I). Es decir, la correlación ecológica será siempre superior a la correlación individual, siempre y cuando se cumpla la siguiente condición:

$$r_I \leq k_3 \cdot r_T \quad \text{Eq. V.I.14.}$$

$$\text{para } k_3 = \frac{k_1 - 1}{k_2}.$$

- b) La correlación ecológica será siempre superior, en valor absoluto, a la correlación individual total siempre que la correlación individual intra-áreas no sea, a su vez, superior a la correlación individual total, lo que, según Robinson (1950), suele suceder en la práctica.
- c) Cuanto menor sea la correlación intra-áreas, mayor será la correlación ecológica. Esta situación se agudiza especialmente cuando las variables se encuentran afectadas de autocorrelación espacial que consiste, precisamente, en el agrupamiento de valores similares de una variable (menor varianza) en una misma zona (en el caso de dependencia espacial positiva, que es el más común en la realidad).

3. Las IE en la economía

La teoría económica lleva años debatiendo una cuestión estrechamente ligada a las IE: la posible inconsistencia entre las relaciones micro y macroeconómicas. Ya hace años, el profesor Klein (1946) ponía de manifiesto las diferencias existentes entre las relaciones económicas “comportamentales”, establecidas en un nivel individual (como las funciones de producción de una única empresa o la función de consumo de un individuo), y dichas relaciones consideradas en una escala agregada, denominadas relaciones “técnicas”. Según este autor, la función de producción correspondiente a un conjunto agregado de empresas, a diferencia de la función relativa a una única empresa, no depende ya del principio del máximo beneficio/mínimo coste, sino únicamente de factores tecnológicos. Del mismo modo, la estimación del consumo privado, en función de la renta disponible, para una agregación de consumidores, dará lugar a una propensión marginal al consumo muy diferente de la que se obtendría para un único consumidor.

Los economistas se han centrado más en analizar los efectos que tiene esta cuestión, sobre los estimadores de una regresión, así como sobre el coeficiente de determinación, R^2 , llegando a las siguientes conclusiones:

- d) Cualquiera que sea el método de agregación escogido, los estimadores de una IE serán siempre sesgados.
- e) La varianza de los estimadores obtenidos en la IE, siempre será mayor que la correspondiente a la regresión con datos individuales.
- f) El valor del coeficiente de determinación en la IE se incrementa sensiblemente con la agregación de datos individuales, siendo este aumento proporcional a la siguiente expresión:

$$\frac{(M - 2) \cdot SX_T}{(N - 2) \cdot SX_E} \quad \text{Eq. V.1.15.}$$

siendo M: número de unidades agregadas
 N: número de unidades individuales
 SX_T: suma de los cuadrados de los valores individuales de la variable independiente
 SX_E: suma de los cuadrados de los valores de la variable en las unidades agregadas.

Como puede observarse, cuanto mayor es la dispersión de la variable explicativa, en el ámbito individual (SX_T), o menor la dispersión de la misma en el ámbito agregado o ecológico (SX_E), mayor incremento experimentará el coeficiente de determinación, R², en la regresión ecológica. Una vez más, se pone de manifiesto el posible error que podría cometerse al inferir relaciones extraídas de un ámbito agregado al ámbito individual.

V.1.3.2. Modelos de IE

A continuación, se exponen tres modelos, de los varios que se han propuesto en la literatura (ver King, 1997), fundamentales para abordar el problema de la agregación de datos espaciales, que está en el centro de toda IE: el modelo de Goodman, origen y fuente de otras muchas especificaciones, el modelo de Duncan y el modelo de King, recientemente presentado como “la” solución a este problema.

1. Modelo básico de Goodman

La regresión de Goodman (1953) es el modelo de IE más utilizado en la literatura, a partir del cual se han formulado otras muchas especificaciones. Este autor se propone superar la visión pesimista de Robinson (1950), sobre la problemática de las IE, dada la necesidad evidente de tantos investigadores sociales de conocer el comportamiento más o menos individual (en situaciones para las que no se dispone de información directa). Para ello, bajo unos supuestos restrictivos, Goodman especifica el siguiente modelo determinista que es, en realidad, una identidad contable aplicada al caso, expuesto en la Tabla V.1.2., de la intención de voto en unas elecciones:

$$T_i = \beta_i^n X_i + \beta_i^b (1 - X_i) \quad \text{Eq. V.1.16.}$$

siendo T_i: proporción de votantes sobre el total de población en edad de votar, en la sección electoral i
 X_i: proporción de población en edad de votar de raza negra, en la sección i
 (1-X_i): proporción de población en edad de votar de raza blanca, en la sección i
 β_iⁿ: parámetro correspondiente a la proporción (desconocida) de votantes de raza

negra en la sección i
 β_i^b : parámetro correspondiente a la proporción (desconocida) de votantes de raza blanca en la sección i

Este modelo relaciona la tasa de votantes en las secciones electorales (T_i), con la composición de la población residente en las mismas según dos subgrupos mutuamente excluyentes y exhaustivos, como son la proporción de votantes de raza blanca ($1-X_i$) y no-blanca (considerada, en términos generales, como raza negra), X_i .

Se trata, pues, de una identidad que se cumple exactamente para cada una de las N unidades microterritoriales (secciones) del espacio total (el Estado de Ohio, en el citado ejemplo de King, 1997), y que produce un sistema de N ecuaciones con $2N$ incógnitas, una por sección. Tal como indican Anselin y Cho (2002A), desde un punto de vista clásico (no bayesiano), la estimación de los parámetros de este modelo constituye un caso de problema de parámetros incidentales (ver Apartado IV.1.2.), que es imposible de resolver mediante la incorporación al modelo de nuevas observaciones muestrales, dado que, cada una de ellas, daría lugar a dos nuevos parámetros a estimar. King (1997) describe esta situación como un problema de indeterminación, que puede ser resuelto de diferentes formas.

Goodman resuelve este problema de indeterminación, o de parámetros incidentales, imponiendo diversas restricciones a su modelo: la hipótesis de constancia de los parámetros y la restricción del valor de los parámetros en el intervalo $[0,1]$.

- a) La hipótesis de **constancia radical de los parámetros** implica que los estimadores de los parámetros (β_i^n, β_i^b) sólo serán correctos si son constantes para todas las secciones (i) del espacio:

$$\beta_i^n = B^n \quad ; \quad \beta_i^b = B^b \quad ; \quad \forall i \quad \text{Eq. V.1.17.}$$

Es decir, la estimación del anterior modelo por el método de MCO, producirá unos coeficientes (\hat{B}^n, \hat{B}^b), constantes para todas las secciones (i) de un distrito (k), que serán los mejores estimadores de los parámetros (B^n, B^b) correspondientes a ese nivel espacial superior, que es el distrito electoral. Por tanto, en el modelo de Goodman hay que distinguir los dos modelos siguientes (uno para cada ámbito espacial de sección y distrito):

$$\begin{aligned} \cdot \text{ Sección electoral (i):} \quad & T_i = \beta_i^n X_i + \beta_i^b (1 - X_i) \\ \cdot \text{ Distrito electoral (k):} \quad & T_{ki} = B^n X_k + B^b (1 - X_k) \end{aligned} \quad \text{Eq. V.1.18.}$$

El problema de este modelo se produce cuando se incumple la citada restricción de constancia de los parámetros (B^n, B^b) en todas las secciones, porque dependen de

los valores de la variable exógena⁶⁶ (X_i), la estimación de este modelo por MCO dará lugar a estimadores sesgados (cuya media no coincidirá con la media de los parámetros teóricos). Efectivamente, si de la inferencia ecológica se derivase una relación negativa entre la tasa de votos válidos (T) y la proporción de población de raza negra en edad de votar (X), habría que concluir que la proporción de votantes de raza negra (B^n) sería siempre inferior a la tasa de votantes blancos (B^b). Este resultado se vería, por ejemplo, seriamente falseado en aquellas secciones (i) de mayoría negra en las que los blancos pudieran comportarse de forma extrema: o bien votaran desproporcionadamente como respuesta a una “amenaza negra”, o bien experimentarían una alta abstención al carecer, en ese lugar, de candidatos a su gusto.

b) Aunque el modelo de Goodman no considera explícitamente la aplicación del **método de los límites**, existe también en el mismo una restricción en el valor de los parámetros (B^n, B^b) que, por tratarse de proporciones, debe situarse en el intervalo $[0,1]$. Así, los casos extremos vendrían dados por la siguientes situaciones:

- $B^n = B^b = 1$. Esta situación se produce en casos denominados de “unanimidad”, es decir, cuando en una unidad electoral todas las personas, blancas y negras, en edad de votar ejercen este derecho en unas elecciones, de forma unánime. En este caso, la tasa de votantes, T_i , será también igual a 1.
- $B^n = B^b = 0$. Ésta sería la situación contraria, en la que se produce una total abstención, por parte de negros y blancos, en unas elecciones, por lo que $T_i = 0$.

La dificultad se encuentra en aquellas situaciones en las que cabe esperar un cierto grado de voto/abstención, por parte de ambos colectivos (negros/blancos), en las que resulta muy difícil obtener el valor exacto de los parámetros, (B^n, B^b). En la práctica, la estimación de este modelo puede producir estimadores imposibles que no se ajusten al citado intervalo (ver en King, 1997, una explicación más amplia sobre esta cuestión).

El problema de la indeterminación o parámetros incidentales es también conocido, en la literatura econométrica espacial, como heterogeneidad espacial extrema (Anselin, 1988A, 2000B) y la única solución al mismo pasaría por la imposición de una estructura geográfica a la naturaleza de la variación de los coeficientes en las observaciones muestrales. Esta “solución” es obviamente parcial dado que los parámetros a estimar ya no serán incidentales (diferentes para cada observación muestral) y tendrán que ajustarse a unas restricciones que les impondrán una cierta constancia (para determinados grupos de observaciones). A continuación, se presentan dos ejemplos, el modelo de Duncan y el modelo de King, que resuelven esta situación mediante especificaciones de heterogeneidad espacial, ya presentadas en el Apartado IV.1.2., como el modelo de expansión espacial y el modelo de coeficientes aleatorios, respectivamente.

⁶⁶ Efectivamente, por ejemplo, la proporción de votos del colectivo de raza blanca (β_i^b), en una sección i cuya población residente es mayoritariamente de raza blanca, pudiera no coincidir con la proporción correspondiente a este colectivo en secciones con mayoría negra.

2. Modelo de Duncan

Duncan *et al.* (1961) plantean como solución al problema de la indeterminación, cuando los parámetros individuales (β_i^n, β_i^b) se encuentran correlacionados con la variable independiente (X), introduciendo dicha dependencia en la especificación del modelo de Goodman, que quedaría transformado en un modelo de expansión de parámetros, similar al propuesto por Casseti (1972). Efectivamente, los autores consideran que los parámetros del modelo de Goodman experimentan una deriva continua que es función lineal de la variable exógena (X), del modo siguiente:

$$\begin{aligned}\beta_i^n &= \gamma_{0n} + \gamma_{1n} X_i \\ \beta_i^b &= \gamma_{0b} + \gamma_{1b} X_i\end{aligned}\quad \text{Eq. V.1.19.}$$

siendo γ_{0n}, γ_{0b} : término independiente correspondiente a los parámetros de votantes negros y blancos, respectivamente.

γ_{1n}, γ_{1b} : pendiente correspondiente a los parámetros de votantes negros y blancos, respectivamente.

A partir de este modelo, propuesto para el ámbito inferior de las secciones, habría que obtener los verdaderos parámetros (o proporción de votantes blancos y negros), en el ámbito agregado de los distritos (B^n, B^b), que serían comunes a todo el espacio muestral, como medias ponderadas de los anteriores parámetros individuales (β_i^n, β_i^b), del modo siguiente:

$$\begin{aligned}B^n &= \frac{\sum_{i=1}^N N_i^n \beta_i^n}{N^n} = \gamma_{0n} + \gamma_{1n} \sum_{i=1}^N \frac{N_i^n X_i}{N^n} \\ B^b &= \frac{\sum_{i=1}^N N_i^b \beta_i^b}{N^b} = \gamma_{0b} + \gamma_{1b} \sum_{i=1}^N \frac{N_i^b X_i}{N^b}\end{aligned}\quad \text{Eq. V.1.20.}$$

siendo N_i^n, N_i^b : número de residentes en edad de votar de la sección i, de raza negra y blanca, respectivamente.

N^n, N^b : número total de personas en edad de votar residentes en el espacio total considerado, de raza negra y blanca, respectivamente.

De este modo, la sustitución de los parámetros expandidos (Eq.V.1.19) en el modelo de Goodman daría lugar a la siguiente especificación:

$$\begin{aligned}
 T_i &= (\gamma_{0n} + \gamma_{1n} X_i) X_i + (\gamma_{0b} + \gamma_{1b} X_i)(1 - X_i) \\
 T_i &= \gamma_{0b} + (\gamma_{0n} + \gamma_{1b} - \gamma_{0b}) X_i + (\gamma_{1n} - \gamma_{1b}) X_i^2 \\
 T_i &= (\gamma_{0n} + \gamma_{1b}) X_i + \gamma_{0b} (1 - X_i) + (\gamma_{1n} - \gamma_{1b}) X_i^2
 \end{aligned}
 \tag{Eq. V.1.21.}$$

Basados en este modelo, otros autores (Shively, 1985; Freedman *et al.*, 1991) han propuesto nuevas especificaciones y nuevas soluciones al problema de heterogeneidad o parámetros incidentales. Aunque probablemente, una de las propuestas más interesantes es la que se expone a continuación.

3. Modelo de King

King (1997) propone un modelo que, a su juicio, constituye la solución definitiva al problema de las IE y que, como ya se ha indicado, ha suscitado una gran polémica entre los investigadores dedicados al tema. Este autor parte, en realidad, del modelo de Goodman y del supuesto de que existen unos valores medios para los parámetros incidentales del mismo (β^n, β^b), que se mantienen constantes a través de las observaciones muestrales. Alrededor de estas medias constantes de cada parámetro, se produce una variación aleatoria o heterogeneidad de valores. Lo que King propone es la estimación tanto de los valores medios constantes de los parámetros, como de su correspondiente matriz asociada de covarianzas. De esta forma, la aplicación de este paradigma de coeficientes aleatorios hará posible la construcción de un modelo estadístico que permita, a su vez, la estimación de los mejores estimadores de los parámetros individuales (β_i^n, β_i^b), a partir de las estimaciones de los valores medios constantes (β^n, β^b) y de la matriz asociada de covarianzas⁶⁷.

La solución de King podría decirse que es una formulación que combina el enfoque de coeficientes aleatorios con el citado método de los límites, procedente del esquema bayesiano. El modelo parte, como ya se ha indicado, del supuesto de que los parámetros presentan una variación aleatoria en torno a un valor medio común, del modo siguiente:

$$\begin{aligned}
 \beta_i^n &= \beta^n + \mu_i^n \\
 \beta_i^b &= \beta^b + \mu_i^b
 \end{aligned}
 \tag{Eq. V.1.22.}$$

siendo μ_i^n, μ_i^b las perturbaciones aleatorias esféricas, e incorrelacionadas con la variable X,

tales que:

$$\begin{aligned}
 \mu_i^n &\approx N(0, \sigma_n^2) \\
 \mu_i^b &\approx N(0, \sigma_b^2) \\
 E[\mu_i X_i] &= 0
 \end{aligned}$$

La introducción de estas expresiones en la regresión de Goodman dará lugar a lo que, en términos de econometría espacial, es un caso de modelo de heterogeneidad espacial

⁶⁷ En el Anexo III, se incluye información acerca de los programas informáticos, EI y EzI, desarrollados por King para la estimación de las IE a través de este modelo.

que suele ser denominado como modelo del error heteroscedástico, cuya especificación y estimación ha sido ya presentado en los Apartados IV.1.2. y IV.2.3., respectivamente.

$$T_i = \beta^n X_i + \beta^b (1 - X_i) + u_i \quad \text{Eq. V.1.23.}$$

siendo u_i la perturbación aleatoria no esférica correspondiente a la sección i , tal que:

$$E[u_i] = E[\mu_i^n X_i + \mu_i^b (1 - X_i)] = 0$$

$$Var[u_i] = E[u_i]^2 = \sigma_n^2 X_i^2 + \sigma_b^2 (1 - X_i)^2 + 2\sigma_{nb} (1 - X_i)$$

La media de la perturbación u_i será nula cuando la variable X cumpla la condición de ser realmente exógena, es decir, siempre que no se produzca el sesgo de la agregación que relaciona los parámetros con las variables explicativas. En cuanto a la varianza de u_i , como puede observarse es heteroscedástica dado que varía con los distintos valores de X en cada sección electoral. La importancia que tiene este problema está en relación directa con la variabilidad de X , que podría contrastarse en la forma habitual a través, por ejemplo, del estadístico Breush-Pagan derivado del multiplicador de Lagrange.

Por otro lado, a diferencia de lo que suele establecerse para los modelos espaciales de coeficientes aleatorios, en el modelo de King la covarianza de las perturbaciones aleatorias correspondientes a los parámetros, $Cov[\mu_i^n \mu_i^b] = \sigma_{nb} \neq 0$, por definición, puesto que la identidad contable de Goodman conlleva la existencia de una relación lineal entre los mismos, del modo siguiente:

$$\beta_i^n = \frac{T_i}{X_i} - \beta_i^b \frac{(1 - X_i)}{X_i} \quad \text{Eq. V.1.24.}$$

Es decir, excepto en el caso que $X_i = 1 \forall i$, esta relación supone la existencia de una covarianza no nula entre los dos parámetros del modelo. Es decir, dado un valor de β_i^n , es posible obtener un valor para β_i^b , y viceversa, pudiendo dibujarse sobre una recta los siguientes pares de parámetros (β_i^n, β_i^b) . Esta representación, que es denominada por King como "gráfico tomográfico", permite observar gráficamente las restricciones sobre los pares de parámetros aceptables.

El valor medio de los parámetros (β^n, β^b) puede estimarse, en la Eq. V.1.20, consistentemente sin ninguna hipótesis previa sobre su correspondiente distribución, por lo que se opta por el método de mínimos cuadrados generalizados factibles (MCGF) que, a diferencia de MCO y MV, no exige el supuesto de normalidad en los mismos. Una vez obtenidos los coeficientes correspondientes a los parámetros generales $(\hat{\beta}^n, \hat{\beta}^b)$ éstos serán utilizados en la construcción de los estimadores óptimos $(\hat{\beta}_i^n, \hat{\beta}_i^b)$ de los parámetros individuales, a través de la siguiente relación (propuesta para el modelo de coeficientes aleatorios por Griffiths, 1972):

$$\hat{\beta}_i = \hat{\beta} + \hat{\Sigma}x_i [x_i' \hat{\Sigma}x_i]^{-1} [y_i - x_i' \hat{\beta}] \quad \text{Eq. V.1.25.}$$

siendo $\hat{\beta}$: vector de parámetros medios, comunes a toda la muestra
 $\hat{\Sigma}$: matriz que contiene los estimadores de los elementos de la matriz no escalar de varianzas y covarianzas de la perturbación aleatoria (u_i , en el Eq.V.1.20).
 x_i : vector de observaciones de las variables explicativas
 $[y_i - x_i' \hat{\beta}]$: residuo correspondiente a i (de la perturbación aleatoria u_i)

Esta expresión indica que el estimador ELIO correspondiente a una unidad i de un parámetro teórico, $\hat{\beta}_i$, se obtiene sumando al estimador del valor medio general de dicho parámetro, $\hat{\beta}$, el error de la regresión estimada para los valores medios de los parámetros, debidamente ponderado por una función de los valores de las variables exógenas, x_i , y las covarianzas de la correspondiente perturbación aleatoria.

En resumen, el modelo de King es una especificación particular del fenómeno de heterogeneidad espacial, en la que la expresión de los parámetros en cada unidad muestral se obtiene como una variación aleatoria en torno a una media general.

Anselin y Cho (2002A) dudan de que la imposición de una estructura paramétrica aleatoria, muchas veces adecuada, pueda ser la panacea para el problema general de las IE, ya que no es una solución general, ni flexible, ni robusta e, incluso, ignora que existen otras estructuras espaciales que pueden ser también causa de heterogeneidad en un modelo (ver Apartados IV.1.2. y IV.2.3.). Por ejemplo, Cho (2001) presenta la heterogeneidad de los coeficientes, no tanto como una variación en torno a una media común, sino con un criterio “espacial” suponiendo diferencias entre diversos conjuntos de unidades geográficas. Por otro lado, estos autores demuestran que el papel jugado por el fenómeno de autocorrelación espacial en los modelos de IE, es muy superior al estimado por King, por lo que debería también ser expresamente considerado, tal como se expondrá a continuación.

V.1.4. Efectos espaciales y el sesgo de la agregación

Pese a que algunos autores (por ejemplo, Arbia, 1989; King, 1997) han destacado el efecto que el fenómeno de autocorrelación espacial puede tener sobre las IE y todos coinciden en la muy posible existencia de heteroscedasticidad en los datos agregados, los primeros que de un modo científico analizan el papel de los efectos de dependencia y heterogeneidad espacial en las IE son Anselin y Cho (2002A y B) y Cho (2001). A través de experimentos de Monte Carlo, ambos autores se proponen cuantificar las consecuencias que determinadas estructuras de autocorrelación y heterogeneidad espacial, frecuentemente existentes en la realidad socioeconómica, tienen sobre la agregación de datos y, por tanto, las IE⁶⁸.

⁶⁸ Para ello, se considerará el modelo básico de Goodman como punto de partida, dejando abierta a futuras investigaciones el análisis de los efectos espaciales en modelos de IE más complejos y generales.

Según King (1997) y Cho (1998), el sesgo de la agregación, que está en el origen de las IE, y el efecto de **autocorrelación espacial** suelen estar estrechamente vinculados en la realidad. Como ya se ha expuesto, en el marco de las IE, el sesgo de la agregación se produce cuando los parámetros del modelo están correlacionados con los regresores y, por tanto, la consideración de parámetros comunes a toda la muestra (B^n , B^b) dará lugar a falsas inferencias en el ámbito microterritorial. La solución vendría dada por una reespecificación del modelo que evitara esta falta de exogeneidad en las variables independientes, por ejemplo, incorporando expresamente las relaciones existentes entre éstas y los parámetros del modelo (como en los ya presentados modelos de Duncan y King).

Resulta interesante, en este sentido, el estudio en el que Achen y Shively (1995) establecen un paralelismo entre el sesgo de la agregación y el efecto de autocorrelación espacial. En concreto, se refieren a esta última como “autocorrelación espacial intra-circunscripción” que consiste, por un lado, en la existencia de una evidente similitud entre los votantes de una misma circunscripción o distrito electoral y, por otro, en las diferencias que se producen, entre distritos, en cuestiones como la lealtad o las deserciones políticas del electorado, y que no tienen por qué ser achacadas a las variaciones de la variable independiente.

Es decir, según estos autores, el fenómeno de autocorrelación espacial es un indicador de la existencia del sesgo de la agregación en una variable de datos regionales (no individuales), de forma que, una correcta consideración de dicha autocorrelación espacial en una IE (mediante métodos de filtrado o con una especificación adecuada), evitaría dicho sesgo en la misma.

King (1997), por su parte, contempla también la posibilidad de existencia de autocorrelación espacial en las IE, aunque el tratamiento que le concede a este punto no es adecuado ni completo, lo que conducirá a minusvalorar este problema, como se expondrá a continuación.

Efectivamente, este autor propone una especificación para la dependencia espacial en los parámetros de su modelo que no considera que las relaciones entre unidades espaciales son simultáneas, por lo que la expresión resultante es más propia del contexto temporal (unidimensional/unidireccional) que espacial (bidimensional/multidireccional):

$$\begin{aligned}\beta_i^n &= \delta\beta_{i-1}^n + (1-\delta)U^n \\ \beta_i^b &= \delta\beta_{i-1}^b + (1-\delta)U^b\end{aligned}\quad \text{Eq. V.1.26.}$$

siendo β_i^n, β_i^b : parámetros del modelo de IE (proporción de votantes de raza negra y blanca, respectivamente), correspondientes a cada unidad i
 $\beta_{i-1}^n, \beta_{i-1}^b$: parámetros del modelo de IE correspondientes a la unidad inmediatamente “anterior” (según un orden previo de la muestra) a i
 U^n, U^b : realizaciones de una distribución normal bivalente truncada correspondientes a β_i^n, β_i^b , respectivamente.

- δ : coeficiente autorregresivo de primer orden, tal que:
- $\delta = 0$: ausencia de autocorrelación espacial (el modelo quedaría reducido al modelo inicial de King)
 - $\delta = 1$: se produce una equivalencia total entre parámetros, para toda unidad muestral i : $\beta_i^n = \beta_{i-1}^n = B^n$; $\beta_i^b = \beta_{i-1}^b = B^b$ (el modelo quedaría reducido al modelo de Goodman).

Obviamente, esta especificación no expresa correctamente un proceso SAR espacial autorregresivo, tal como se ha presentado ampliamente en el Apartado IV.1.1 (en todo caso, como indican Anselin y Cho, sería una expresión más relacionada con la especificación espacial de medias móviles, SMA, o autorregresiva en el término de la perturbación aleatoria). La simultaneidad y multidireccionalidad propias del contexto espacial, ya se ha demostrado (ver Apartado III.2. y Anexo V) que exige la introducción en el modelo de una matriz de pesos espaciales que exprese la naturaleza de las interacciones existentes entre las observaciones de la muestra.

Junto con el fenómeno de autocorrelación, la heterogeneidad espacial es también un efecto propio de los datos espaciales que debe considerarse explícitamente. En este caso, las IE deben ser consideradas como un caso particular de **heterogeneidad espacial**, en cuanto fenómeno por el que un modelo (forma funcional, parámetros y perturbación aleatoria) no es constante para todas las observaciones muestrales. En términos de econometría espacial, podría decirse que el modelo inicial de Goodman expresa una situación de heterogeneidad extrema en la que debe estimarse un parámetro (β_i^n, β_i^b) para cada unidad muestral. Es cierto que, algunos modelos ya presentados, como los formulados por Duncan *et al.* (1961) y King (1997), incorporan la heteroscedasticidad a través de una expansión de parámetros o con coeficientes aleatorios, respectivamente. Sin embargo, ninguna de estas formulaciones es propiamente “espacial”, como en el caso anterior.

Por tanto, dado que los efectos espaciales están íntimamente relacionados con el sesgo de la agregación y, por tanto, con las IE, resulta del todo evidente que éstos deban ser incorporados explícitamente en la especificación de cualquier modelo de este tipo. Por eso, a continuación, siguiendo las indicaciones de Anselin y Cho (2002A) y Cho (2001), se proponen algunas especificaciones de modelos de heterogeneidad espacial en los que se considera también la presencia de algunas formas de autocorrelación espacial.

V.1.5. Problemas y vías de solución

Tras la exposición de los principales conceptos relativos al MAUP y las IE, se ha querido poner de manifiesto la gran dificultad que entraña toda regresión ecológica (efectuada con datos espacialmente agregados), como estimación significativa para ámbitos desagregados, sólo aceptable bajo hipótesis muy restrictivas que, sin invalidar este tipo de aplicaciones, invitan a interpretar los resultados con cierta precaución. El ejercicio de predicción-extrapolación espacial consiste precisamente en obtener, a partir de datos y relaciones procedentes de un ámbito espacial agregado (provincias, regiones), información estadística de ámbito microterritorial (municipios, distritos, barrios, secciones).

Los principales problemas propios de cualquier ejercicio de extrapolación espacial podrían resumirse del modo siguiente: práctica imposibilidad de verificación de resultados, ausencia de teorías válidas en el ámbito microterritorial, incorrecto tratamiento de los efectos espaciales y problema de los límites.

1. Práctica imposibilidad de verificación de resultados

La predicción-extrapolación de datos microterritoriales plantean un problema que no suele producirse en el campo temporal: la imposibilidad de verificación de la acuracidad de los valores de predicción. Efectivamente, así como en la dimensión temporal llega un momento en el que el futuro deja de serlo, pudiendo disponerse de información estadística real sobre la predicción realizada en un momento anterior del tiempo, en el espacio esto no suele producirse debido, en parte, al elevado coste de obtención de datos microterritoriales, que impide poder llegar a disponer alguna vez de esta información.

Las soluciones que podrían proponerse a estos problemas son únicamente parciales, en el sentido de que ayudan a paliar las consecuencias que el sesgo de la agregación produce en los fenómenos económicos y, por tanto, en cualquier ejercicio de predicción-extrapolación de datos microterritoriales a partir de inferencias ecológicas.

Por ejemplo, un modo de verificar los resultados de la predicción podría ser su **comparación con estudios similares**, como sucede en el caso presentado en el Apartado V.3., de predicción de la renta disponible municipal de la Comunidad de Madrid en 2000, ya realizada por el Instituto de Estadística de la Comunidad de Madrid, aunque referida a períodos anteriores.

2. Posible ausencia de teorías válidas en el ámbito microterritorial

En muchos fenómenos sociales, como la economía, las principales teorías se han elaborado en ámbitos territoriales superiores (países o regiones) no siendo siempre directamente aplicables en una escala microterritorial, cercana al individuo. Efectivamente, el citado sesgo de la agregación consiste precisamente en una pérdida de información (o desvirtuación de la misma) que se produce en el proceso de agregación de datos más o menos individuales para su expresión en una escala superior.

Por eso, como ya se ha demostrado, algunos estadísticos básicos cambian de tamaño e incluso de signo, pudiendo no ser válidas las teorías propuestas para ámbitos agregados. Así, por ejemplo, la relación de causa-efecto entre actividad económica per capita y renta familiar per capita, respectivamente, no siempre se produce en el ámbito microterritorial.

Por ejemplo, los municipios residenciales, en los que se concentra la población con mayor poder adquisitivo, muchas veces presentan una menor capacidad productiva que otras localidades industriales, en las que reside una población con menor nivel de renta.

Muchas veces, estas diferencias en las relaciones económicas, pueden producirse a causa de la mayor heterogeneidad propia del ámbito microterritorial que obliga a distinguir diversas estructuras o regímenes. En el caso anterior, la relación entre actividad económica y renta familiar es perfectamente inferible, desde ámbitos superiores, a municipios de un cierto tamaño poblacional (capitales de provincia, cabeceras comerciales, etc.).

En cualquier caso, la bondad de una determinada relación económica, en cuanto destinada a ser inferida a un ámbito microterritorial, no siempre podrá ser evaluada a partir de los estadísticos básicos, como el coeficiente de correlación lineal, o ajustes de bondad como el coeficiente de determinación (R^2), pues como ya se ha indicado, éstos experimentan aumentos de tamaño con aumentos de escala. Es decir:

- Cuanto menor sea la escala espacial de un fenómeno, mayor será la variabilidad del mismo y, por tanto, menos preciso será el dato agregado como estimador del desagregado.
- Y cuanto mayor sea la dispersión de la variable explicativa en el ámbito individual, o menor la dispersión de la misma en el ámbito agregado o ecológico, mayor incremento experimentará el coeficiente de determinación en la regresión ecológica.

Estos problemas podrían ser mitigados mediante algunas acciones, como la búsqueda de información muestral, la cuidadosa elección de las variables explicativas y la aplicación de técnicas del AEDE.

La **búsqueda de información muestral** sobre los agentes individuales, cosa que no siempre es posible, sería una información auxiliar que podría ayudar a verificar las relaciones económicas en ámbitos microterritoriales. Por ejemplo, los datos de explotación de la Encuesta de Presupuestos Familiares constituyen un punto de partida interesante para el análisis microterritorial de la función de consumo.

Otra vía de solución sería la **elección cuidadosa de las variables independientes**, en el sentido de que sean verdaderamente explicativas del fenómeno en el ámbito microterritorial pues, tal como indica Bivand (1998), podría ayudar a mitigar el sesgo de la agregación. Es decir, habría que rechazar variables que, aunque con buen poder explicativo en una escala agregada, pudieran presentar algún sesgo en el ámbito microterritorial.

Por ejemplo, la variable de oficinas de crédito por habitante, habitualmente seleccionada por su capacidad explicativa de la renta disponible per capita en el ámbito provincial, produce sesgos en el ámbito municipal debido, entre otras causas, a la gran cantidad de municipios rurales sin oficinas bancarias, o con una oficina de horario reducido, o con un radio de influencia de varias localidades rurales.

Además, la **aplicación de técnicas de AEDE** sobre las variables del modelo, tanto en el ámbito agregado como microterritorial, permite obtener una mejor información sobre su distribución, los efectos espaciales y la capacidad explicativa de las mismas.

3. Incorrecto tratamiento de los efectos espaciales

Cualquier ejercicio de predicción-extrapolación de datos individuales exige una modelización adecuada de los efectos espaciales de dependencia y heterogeneidad espacial intrínsecas a dicho fenómeno. Como ya se ha indicado, existe una práctica identificación entre el sesgo de la agregación y presencia de autocorrelación espacial en los fenómenos. Por otro lado, el mismo sesgo de la agregación produce heterogeneidad espacial en los modelos de regresiones ecológicas.

En este sentido, resultan fundamentales las especificaciones de los modelos de heterogeneidad espacial, presentados en el Apartado IV.1.2., con algunas modificaciones para implementar, también en ellos, algunas especificaciones de autocorrelación espacial.

No debe tampoco olvidarse que las alteraciones producidas por el cambio de escala afectan también a la intensidad y forma de los propios efectos espaciales. Efectivamente, la distribución de la renta familiar disponible per capita, en el ámbito provincial español, experimenta clara dependencia espacial, así como dos estructuras o regímenes espaciales determinados por “dos Españas”: la mitad sur-suroeste más deprimida (con una “zona fría” en las provincias del extremo suroeste peninsular) y la mitad rica del norte-nordeste (con una “zona caliente” en algunas provincias del norte y nordeste). Sin embargo, el análisis de los resultados de renta familiar per capita para el ámbito municipal de la Comunidad de Madrid, por ejemplo, aunque también con autocorrelación espacial, presenta otra estructura de heterogeneidad espacial, más del tipo centro-periferia (ver Apartado V.3.).

La **especificación de modelos de econometría espacial** es la única que permite el tratamiento adecuado de los efectos de autocorrelación y heteroscedasticidad espacial. Este tratamiento podría consistir en la eliminación o filtrado de dichos efectos, como el filtro del retardo espacial propuesto por Getis (ver Apartado IV.1.1.), o en la incorporación de los efectos espaciales en la especificación del modelo, como sucede, por ejemplo, con la deriva de los parámetros en algunos modelos de heterogeneidad espacial.

Se han propuesto también otras soluciones, como el modelo de King (1997) para las inferencias ecológicas que, aunque reconoce la importancia de los efectos espaciales en las IE y especifica la heteroscedasticidad a través de un modelo de parámetros aleatorios, no tiene en cuenta adecuadamente la dependencia espacial y otras formas de heterogeneidad espacial.

Además, los modelos jerárquicos de regresión lineal, propuestos inicialmente por Bryk y Raudebush (1992) y sus compañeros bayesianos, podrían constituir también una solución a este problema, a través de la integración de múltiples escalas de análisis, aunque aún queda mucho por hacer en el ámbito espacial de análisis, siendo Langford *et al.* (1999) uno de los pioneros en este campo. Esta metodología se revela interesante, sobre todo, en variables categóricas y discretas para las que no se han formulado aún soluciones analíticas que incorporen la dependencia espacial. Se trata de métodos basados en estimadores de simulación, como el muestreo de Gibbs y las cadenas de Markov de Monte Carlo (MCMC), tal como exponen Gilks *et al.* (1996), LeSage (2000), Beron y Vijverberg (2000). Quedan,

sin embargo, por resolver importantes cuestiones metodológicas e informáticas, sobre todo, para implementar estas técnicas en grandes bases de datos.

4. El problema de los límites

Al igual que en las IE, el ejercicio de predicción-extrapolación se encuentra también condicionado por unos límites, que no se establecen sobre el valor de los parámetros como en el caso de IE. Ahora, los límites se encuentran en el valor agregado de la magnitud a predecir que, para cada unidad territorial superior, debe coincidir con la suma de los valores correspondientes a dicha variable en las unidades inferiores.

Por ejemplo, si se pretende predecir la renta bruta disponible municipal a partir de datos oficiales de renta disponible provincial, debe asegurarse que la suma de los valores estimados de renta municipal para una provincia, sea igual al valor oficial de la renta en dicha provincia.

El método habitual de ajuste consiste repartir proporcionalmente, entre cada unidad microterritorial (por ejemplo, municipios), la diferencia entre la suma de los valores estimados para una variable en el microterritorio (municipios de una provincia) y el valor de partida correspondiente a dicha unidad superior (provincia) que, obviamente, no suelen coincidir, como ya se ha indicado.

Recientemente, en un estudio patrocinado por el Instituto de Estadística de Andalucía (1999), los autores propugnan una solución que, siendo interesante, no soluciona este problema. Se trata de un procedimiento por el que se realiza una inferencia ecológica a través de un modelo en el que todas las variables están relativizadas por la población y, en concreto, las variables exógenas se encuentran, además, en desviaciones a la media.

En el proceso de extrapolación microterritorial, los valores de las variables se relativizan por la población y se expresan en desviaciones a la media local, es decir, correspondiente a la unidad agregada en la que se encuentran (por ejemplo, provincias). Además, el término independiente del modelo ecológico, que sería una media de la variable endógena en el espacio general considerado (media nacional), se sustituye por el valor medio de las unidades agregadas (valores provinciales), de forma que se estima un modelo diferente para cada grupo agregado de unidades microterritoriales; es decir, un modelo distinto para cada provincia, debido a las diferentes ordenadas en el origen.

Este procedimiento no conduce a unos valores microterritoriales estimados cuya suma dé como resultado el valor total correspondiente a la unidad agregada, debido a que la dispersión de los valores, entre unidades agregadas, en torno a la media general sigue siendo menor que la dispersión de valores, entre unidades microterritoriales, en torno a una media local.

V.2. MODELOS DE PREDICCIÓN-EXTRAPOLACIÓN DE DATOS MICROTERRITORIALES

En España, existe un interés creciente por la información estadística microterritorial y, por tanto, por la obtención de magnitudes socioeconómicas para las que no se dispone de información oficial. El caso más habitual serían las estimaciones de renta familiar disponible, de municipios y comarcas, que se han llevado a cabo en los últimos años, patrocinados por diversos organismos e instituciones publicos y privados, de ámbito nacional y regional.

Estos ejercicios de predicción-extrapolación espacial son conocidos como “métodos indirectos” de estimación, en contraposición con los llamados “métodos directos”, a través de los cuales algunas instituciones importantes han obtenido información más o menos microterritorial de ciertas magnitudes económicas. En este sentido, es conocida la estimación, por el método directo, del valor añadido bruto y renta de los hogares que realizan el INE y la Fundación BBVA para las provincias y FUNCAS para las comunidades autónomas, así como las estimaciones de renta bruta disponible de los hogares que, desde hace años, elabora SADEI para los municipios asturianos.

Los métodos directos requieren de un gran volumen de estadísticas básicas que, por lo general, no están fácilmente disponibles⁶⁹. Por eso, los investigadores han recurrido a procedimientos de predicción-extrapolación que, a partir de inferencias ecológicas (en un ámbito agregado para el que se disponga de datos estadísticos), permitan estimar datos microterritoriales. Es decir, a partir de la estructura de dependencia que liga, en un ámbito ecológico (por ejemplo, provincial) y en un momento dado, una magnitud económica (la renta) con un grupo de indicadores socio-económicos (instrucción, empleo, actividad), disponibles también para todas las unidades microterritoriales (municipios), será posible estimar dicha magnitud económica en la escala espacial desagregada.

Para ello, deben superarse las dificultades expuestas en el Apartado anterior, cosa que apenas se ha considerado hasta ahora por los investigadores. En este sentido, deben destacarse los esfuerzos que el Área de Economía Espacial Microterritorial del Instituto Lawrence R. Klein (Universidad Autónoma de Madrid), viene realizando, desde 1992, en esta línea de investigación. Son conocidos sus ejercicios de predicción-extrapolación de los datos de renta bruta disponible municipal de España que, con la máxima actualización, son anualmente publicados en el Anuario Económico de España, patrocinado por “la Caixa” (2001). Desde las primeras estimaciones de renta bruta disponible municipal de 1992, publicadas en el Atlas Comercial de España (Banco Central-Hispano, 1994), hasta hoy, la metodología empleada ha ido experimentando importantes cambios, en un esfuerzo continuado por superar los problemas que plantea el sesgo de la agregación, así como incorporar correctamente los efectos espaciales en los modelos utilizados.

⁶⁹ Para una mayor información sobre los procesos directos de elaboración de estadísticas, ver SADEI (1994). En el Anexo II, se ha hecho un resumen del proceso de obtención de la renta bruta disponible municipal llevado a cabo por este organismo.

A continuación, se resumen algunas de las más interesantes propuestas realizadas en este campo de la predicción-extrapolación (ver Tabla V.2.1. y Anexo II, para una mayor información al respecto), que se han clasificado en modelos no espaciales y de econometría espacial, en función de su consideración explícita de los efectos espaciales de dependencia y heterogeneidad espacial. Como podrá advertirse, exceptuando el trabajo realizado en el Instituto Lawrence R. Klein (“la Caixa”, 2001) y la tesis doctoral de Alañón (2001), ninguno de los trabajos consultados considera la existencia de efectos espaciales en los mismos. En todo caso, algunas investigaciones, como las llevadas a cabo por Otero *et al.* (Instituto de Estadística de Andalucía, 1999), advierten sobre la necesidad de contrastar la posible presencia de heteroscedasticidad en el modelo, aunque con instrumentos propios de la econometría tradicional (temporal).

Tabla V.2.1. Algunos modelos de predicción-extrapolación de la renta familiar disponible microterritorial (ver Anexo II para mayor información)

Publicación	Período	Escala	Ámbito	Variables explicativas
<i>“la Caixa” (Instituto L.R. Klein)</i>	1999	Municipios (más 1.000 habitantes)	España	. Parque automóv./motos . Oficinas bancarias . Establecim. turísticos . Líneas RDSI . Líneas telef. domésticas . Tasas demandas empleo . Tasa de instrucción . Tasa empleo especializ. . Distancia a cabeceras
<i>Instituto de Estadística de la Comunidad de Madrid</i>	1994- 1997	Municipios	Comunidad de Madrid	. Base imponible IRPF . Índice socioeconómico
<i>Instituto de Estadística de la Comunidad de Madrid</i>	1997	Distritos, barrios y secciones censales	Comunidad de Madrid	. Base imponible IRPF
<i>Ayuntamiento de Madrid</i>	1996	Distritos y barrios	Municipio de Madrid	. Tasa de universitarios . Parque de vehículos
<i>CES Región de Murcia</i>	1987, 1996	Municipios	Región de Murcia	. Líneas telefónicas . Oficinas bancarias . IVA/nº declarantes . Tasa de ocupación
<i>Generalitat de Catalunya</i>	1989- 1995	Municipios (más 5.000 habitantes) y comarcas	Cataluña	. Base imponible IRPF
<i>Reunión de Estudios Regionales (Bachero et al.)</i>	1995	Municipios (más 1.000 habitantes)	Comunidad Valenciana	. Líneas telef.corregidas . Oficinas bancarias . Parque de vehículos . Activ. comercial y turis.

Publicación	Período	Escala	Ámbito	Variables explicativas
<i>Diputación de Sevilla</i>	1986-1994	Municipios	Provincia de Sevilla	. Cons.doméstico electric.
<i>Instituto de Estadística de Andalucía</i>	1994	Municipios (más 10.000 habitantes) y comarcas	Andalucía	. Base imponible IRPF . Oficinas bancarias . Parque de vehículos . Consumo energía eléctric. . Líneas telefónicas . Activ. empresariales
<i>Asepelt (Heras et al.)</i>	1991	Municipios	España (no País Vasco y Navarra)	. Base imponible IRPF . Tasas de ocupación
<i>Banesto</i>	1991	Municipios	España	. Líneas telefónicas . Segunda vivienda . Oficinas bancarias
<i>Universidad Complutense (Alañón)</i>	1991	Municipios	España	. Locales activos . Edificios más 4 plantas . Nivel de instrucción
<i>Asepelt (Fernández y Sierra)</i>	1985	Municipios	La Rioja	. Cons.doméstico electric.
<i>Asepelt (Remírez)</i>	1985	Municipios	Navarra	. Presup. gastos municip.
CAZAR	1981	Municipios	Aragón	. Parque de vehículos . Activos en la industria . Saldos en cajas ahorro
<i>Caixa Galicia</i>	1982, 84, 86	Municipios	Galicia	. Empleo asalariado . Parque de vehículos . Licencias comerciales

Fuente: Elaboración propia.

1. Modelos no espaciales

Como ya se ha indicado, el sesgo de la agregación y los efectos espaciales han sido largamente ignorados en la mayoría de las aplicaciones de predicción-extrapolación de datos microterritoriales que han concentrado sus esfuerzos en la obtención de variables explicativas disponibles en el ámbito microterritorial, así como en la formulación, más o menos compleja, de modelos econométricos que son trasladados directamente del campo temporal al espacial.

En España, el origen de la predicción-extrapolación microterritorial se encuentra en los trabajos de estimación de la renta familiar disponible que Ayuso *et al.* realizaron, durante muchos años, para Banesto, y que eran publicados con carácter periódico en el

Anuario del Mercado Español. Como en el modelo de Goodman, los autores realizan una inferencia ecológica, a partir de los datos de renta familiar disponible provincial publicados por el Servicio de Estudios del entonces Banco de Bilbao (actual BBVA) y de un grupo de variables explicativas, disponibles también para el ámbito municipal de España, suponiendo una constancia paramétrica en todo el territorio nacional e ignorando posibles defectos de especificación por la presencia de dependencia espacial en el modelo.

Esta experiencia, durante muchos años única en España, ha constituido el punto de partida de otros muchos estudios (ver Tabla V.2.1.), sobre todo en los últimos años, en que ha crecido espectacularmente el interés por la información microterritorial, tanto por parte de los organismos públicos, como de las empresas. Sin duda que la obra más parecida al desaparecido Anuario del Mercado Español es hoy el Anuario Económico de España, elaborado por el Instituto L.R. Klein y patrocinado por “la Caixa”, de periodicidad anual, que incluye más de 40 datos estadísticos municipales, entre los que destaca la renta familiar disponible por habitante⁷⁰. En las primeras estimaciones de renta disponible realizadas por el Instituto Klein, aunque con una metodología novedosa (ver Vicéns y Chasco, 1998), las inferencias seguían realizándose a partir de un modelo de regresión lineal, de parámetros constantes, en el que no se contrastaban los efectos espaciales ni se consideraba el posible sesgo de la agregación. Estas cuestiones han empezado a abordarse recientemente, tal como se expone en el Apartado V.2.2., y creemos que reciben una mejor respuesta en el presente trabajo.

Como ya se ha indicado, las primeras experiencias de predicción-extrapolación de datos microterritoriales, fundamentalmente renta familiar disponible, han concentrado sus esfuerzos en la obtención de variables explicativas de ámbito microterritorial, así como en la formulación de modelos econométricos cada vez más complejos. Ambos puntos se desarrollan con más detalle a continuación.

- Obtención de variables explicativas de ámbito microterritorial

La obtención de datos estadísticos microterritoriales ha constituido un gran escollo en la realización de ejercicios de predicción-extrapolación. En países con estructuras políticas muy descentralizadas, como España, este problema se agrava en cuanto que la responsabilidad de obtención de muchas estadísticas queda en manos de las comunidades autónomas, con sus planes estadísticos propios y diferenciados del resto de comunidades autónomas, no existiendo apenas bases de datos microterritoriales de ámbito nacional, homogéneas y actualizadas.

Las primeras experiencias de predicción-extrapolación tuvieron en el Anuario del Mercado Español (AME) su principal base de datos que, siendo única en aquel momento (exceptuando la información decenal publicada por el INE en los censos), resultaba escasa

⁷⁰ Al igual que Banesto, el Anuario de “la Caixa” proporciona los datos de renta familiar disponible por habitante, agrupados en una escala del 1 al 10. Además, aunque el Instituto L. R. Klein elabora la información de renta en términos monetarios y para la totalidad de los municipios del territorio nacional, “la Caixa” sólo presenta en el Anuario Económico información relativa a los municipios de mil y más habitantes, que constituyen más del 95% de la población española.

y posiblemente ineficiente para el propósito de estimar la renta familiar disponible. Dado que gran parte de la información municipal incluida en el AME estaba relacionada con la actividad económica (líneas telefónicas, parque de vehículos, oficinas de crédito, licencias comerciales, etc.), los autores de esta obra, así como muchos otros, han aplicado, también al campo microterritorial, la teoría macroeconómica que relaciona la renta de las familias con la generación de rentas o producción económica. Como se indicó en el apartado anterior, el sesgo de la agregación produce una cierta invalidación de las teorías procedentes de un ámbito territorial superior en la escala microterritorial, lo que ha llevado a situar en los primeros lugares del ranking de renta familiar disponible a localidades de pequeño tamaño con alguna especial actividad productiva, por delante de localidades residenciales en las que reside la población de alto nivel de vida. Por ejemplo, en el caso de la Comunidad de Madrid, se trataba de municipios de mediano tamaño que destacan por la presencia en los mismos de polígonos industriales o segunda vivienda de fin de semana, como Ajalvir y Pelayos de la Presa, respectivamente, por delante de Boadilla del Monte o Pozuelo de Alarcón.

Evidentemente, la relación entre renta de las familias y producción económica es perfectamente válida en ámbitos más agregados o en municipios urbanos, con un tamaño poblacional más cercano al de unidades territoriales superiores. Pero, a medida que se desciende en la escala o tamaño (indistintamente), las relaciones son más difusas y deben explorarse previamente, por ejemplo, aplicando técnicas del AEDE. Por eso, estudios con planteamientos metodológicos interesantes podrían llegar a resultados no muy fiables si no se analizan previamente las distintas estructuras o regímenes existentes.

Algunos autores han apostado por variables procedentes del impuesto sobre la renta de las personas físicas (IRPF), como indicadores “definitivos” en la estimación de la renta familiar de los municipios (Arcarons *et al.*, 1994; Heras, 1992; Heras *et al.*, 1998) o casi definitivos, tras la corrección de los mismos con otras variables socioeconómicas (Instituto de Estadística de la Comunidad de Madrid, 2000, 2001). Esta corrección viene motivada por las limitaciones del IRPF como indicador de renta de las familias, sobre todo, por su incapacidad para captar rentas no declaradas procedentes de la economía sumergida (que en algunas localidades pueden constituir un volumen muy importante) o de ciertas actividades económicas (agrícola, autónomos), así como por la exclusión de las rentas mínimas exentas del impuesto.

- Formulación de modelos

La práctica totalidad de trabajos consultados se basan en la estimación de un modelo uniecuacional de regresión lineal, con datos de corte transversal, estimado en un ámbito ecológico (provincial), siendo el prototipo las inferencias llevadas a cabo por los autores del AME.

A partir de este modelo básico, los esfuerzos se han concentrado en especificaciones diversas. Por ejemplo, algunos autores advierten de la presencia de heteroscedasticidad en la perturbación aleatoria, proponiendo su contraste por métodos de econometría temporal (Instituto de Estadística de Andalucía, 1999), así como de una deriva paramétrica discreta a

partir de varias estructuras no espaciales (Cazar, 1984, 1986; Bachero *et al.*, 1997). Otros autores, conscientes del posible error que entraña la inferencia de relaciones procedentes del campo ecológico en unidades desagregadas, proponen la especificación de un modelo temporal por cada provincia o ámbito ecológico, que permita, de este modo, la deriva paramétrica provincial a cambio de suponer la estabilidad temporal (Remírez, 1991; Fernández y Sierra, 1992).

Otras iniciativas intentan corregir un problema específico de los modelos de renta familiar disponible, el orden invertido de causalidad por el que la renta es función de variables de consumo cuando, en realidad, debería ser al revés. La solución planteada por algunos (Heras *et al.*, 1998) ha sido la especificación de un modelo simultáneo biecualcional que explique, en primer lugar, la renta en función de las tasas de ocupación y, en segundo, el IRPF en función de la renta de las familias.

Pese al interés y la novedad de algunos planteamientos, ninguno de los modelos citados puede considerarse como “bueno”, desde el punto de vista espacial, dado que todos ellos ignoran los efectos espaciales o, al menos, no son tratados con los instrumentos adecuados.

2. Modelos de econometría espacial

Las dos únicas experiencias consultadas que incorporan los conceptos propios de econometría espacial para la predicción-extrapolación de datos microterritoriales, como la renta disponible de los hogares a nivel municipal, son los trabajos del Instituto L. R. Klein (“la Caixa”, 2001) y Alañón (2001). Los primeros plantean un modelo de regresiones cambiantes en el que se especifica una deriva paramétrica discreta en dos regímenes espaciales, previamente contrastados por el test de Chow espacial de Anselin (1990B), que dividen las provincias españolas en dos subespacios denominados norte-nordeste rico y sur-suroeste más deprimido.

Es decir, se plantea un modelo con parámetros cambiantes según la localización (norte-nordeste, sur-suroeste) de las provincias/municipios. La crítica que podría hacerse a un modelo de estas características es que las diferentes sensibilidades no tienen por qué afectar a la totalidad de los municipios incluidos dentro de las provincias correspondientes, por lo que habría que replantear los regímenes de un modo más ajustado a la heterogeneidad real de la renta en los municipios españoles.

En cuanto a la experiencia de Alañón (2001), en su tesis doctoral, resulta interesante desde el punto de vista metodológico, ya que propone la filtración espacial de la variable endógena, con problemas de autocorrelación espacial, como solución a la extrapolación de un modelo del retardo espacial provincial al ámbito municipal.

Efectivamente, la inferencia al ámbito microterritorial de un modelo del retardo espacial estimado en el ámbito provincial sólo sería posible mediante el filtrado previo de dicha dependencia espacial (al transformar el modelo inicial en un MBRL, no se plantea el problema de obtener el retardo de la renta municipal cuando ésta aún no se conoce). Sin

embargo, el autor postpone esta solución, ante la imposibilidad de aplicarla, dadas las limitaciones del “software” disponible en el tratamiento de grandes bases de datos, y la sustituye por un modelo del error espacial que, aunque no plantea las mismas dificultades, se corresponde con un fenómeno espacial de externalidades diferente, que no debe equipararse al modelo del retardo espacial.

El ejercicio de aplicación que se propone en el Apartado siguiente supondría un avance en la consolidación de la metodología de estimación de renta disponible municipal llevada a cabo por el Instituto L. R. Klein, así como una propuesta de aplicación para cualquier ejercicio de predicción-extrapolación de datos microterritoriales.

V.3. PROCESO DE REALIZACIÓN DE UN EJERCICIO DE PREDICCIÓN-EXTRAPOLACIÓN DE LA RENTA FAMILIAR DISPONIBLE MUNICIPAL DE LA COMUNIDAD DE MADRID, EN 2000

Los conceptos fundamentales del MAUP y las IE (Apartado V.1.) ponen claramente de manifiesto que las relaciones estadísticas entre variables obtenidas en un nivel agregado nunca pueden identificarse directamente con las relaciones existentes entre las mismas variables en ámbitos geográficos inferiores. Esto es debido, sobre todo, a las variaciones que pueden sufrir los estadísticos básicos de una variable, con la agregación de datos microterritoriales (individuos, secciones, municipios) a una escala geográfica superior (provincias, regiones). En concreto, todo proceso de agregación de unidades geográficas tiene, al menos, las siguientes consecuencias:

- disminución de la varianza
- disminución del grado de autocorrelación espacial
- incremento (y, a veces, cambio de signo) del coeficiente de correlación lineal entre dos variables
- incremento del coeficiente de determinación, R^2 , de una regresión lineal

Las posturas más o menos pesimistas, adoptadas por algunos autores ante este tipo de dificultades (Robinson, 1950), deben ser superadas mediante instrumentos adecuados, pues el mercado demanda información microterritorial no existente (King, 1997). Los instrumentos o estrategias a seguir, para abordar con éxito un problema de predicción-extrapolación de datos, deberían comprender, tanto herramientas estadístico-econométricas propias del análisis espacial de datos (Anselin, 2000B, Anselin y Cho, 2002A y B), como todo tipo de información complementaria sobre el ámbito microterritorial analizado, ya que, como afirma King (1997), “debe evitarse aceptar, sin más, cualquier resultado que sea técnicamente posible o lógicamente viable”.

En esta línea, se propone, a continuación, un proceso de realización de un ejercicio de predicción-extrapolación de datos microterritoriales, ilustrado con una aplicación en la que se estima y analizan los datos de renta familiar bruta disponible de los municipios de la Comunidad de Madrid, referidos al año 2000. Este proceso consta de 6 etapas, que se enumeran a continuación:

1. Fundamentos teóricos
2. Selección y tratamiento de la información disponible
3. Análisis exploratorio espacial
4. Análisis confirmatorio espacial
5. Predicción-extrapolación de datos microterritoriales
6. Análisis y validación de resultados

Es decir, antes de realizar el ejercicio de modelización propiamente dicho, será necesario un primer paso de fundamentación teórica destinado a lograr una especificación adecuada del modelo ecológico que será posteriormente aplicado a la extrapolación de datos microterritoriales. Como sucede en el campo econométrico tradicional, esta primera etapa estaría destinada a esclarecer el concepto y condicionantes de la magnitud a estimar-extrapolar, a partir de una revisión de los antecedentes económicos y econométricos (Pulido y Pérez, 2001) existentes en la literatura. Este primer análisis debe concluir en una primera selección de factores con potencial poder explicativo, así como la formulación de una serie de hipótesis a contrastar.

A continuación, los factores explicativos anteriormente especificados habrán de convertirse en indicadores estadísticamente disponibles, tanto para el ámbito agregado como desagregado. La capacidad explicativa de dichas variables, así como la ausencia de sesgos de cualquier tipo en las mismas, habrá de ser contrastado en un análisis exploratorio que, contrariamente a lo habitual en otras experiencias similares, tendrá que ser doble, en el sentido de aplicado a los dos ámbitos territoriales analizados.

Efectivamente, dado el cierto nivel de incertidumbre que rodea este ejercicio, pues no siempre existe una relación directa entre relaciones funcionales establecidas en distintos ámbitos territoriales, el AEDE habrá de realizarse de forma que sea posible extraer, a partir del mismo, estructuras o formas más o menos estables, a partir de las cuales construir relaciones confirmatorias “válidas” en ambos niveles territoriales (dentro del cierto nivel de incertidumbre existente).

La aplicación de las técnicas del AEDE, tanto en el ámbito microterritorial como agregado, podrá o no poner de manifiesto la existencia de efectos de dependencia y/o heterogeneidad espacial, en todas o algunas de las variables. Dependiendo de los resultados obtenidos, se procederá a la etapa de análisis confirmatorio espacial, en la que habrá que identificar y estimar el modelo de regresión ecológico más adecuado al fenómeno, habida cuenta, no sólo sus características ecológicas, sino también su distribución microterritorial.

Por último, el ejercicio de predicción-extrapolación consiste en la aplicación, lo más “directa” posible, del modelo seleccionado en el análisis confirmatorio. Es decir, deben considerarse correctamente las posibles estructuras de dependencia y heterogeneidad que hayan sido detectadas, en el sentido de que pueda realizarse una extrapolación correcta de estos efectos desde el ámbito ecológico al microterritorial. Asimismo, deberá evitarse la presencia de la variable endógena RFD (objetivo a estimar) espacialmente retardada como variable explicativa del modelo, mediante el filtrado espacial de las variables exógenas y término independiente.

V.3.1. Fundamentos teóricos

Ante todo, el investigador debe tener claro el concepto de la magnitud que va a analizar, así como poseer un conocimiento suficiente de las teorías en que se fundamenta el mismo, no sólo para niveles territoriales agregados sino también para ámbitos microterritoriales. Esta etapa será fundamental para lograr una especificación ecológica del modelo lo más perfecta posible, que permita extrapolar sus resultados satisfactoriamente.

Así, en el caso que se propone, la magnitud económica que se va a extrapolar es la **renta disponible bruta de los hogares**, tal como la define el Instituto Nacional de Estadística (INE) en la Contabilidad Regional de España. En la Tabla V.3.1., se exponen los resultados de la Cuenta de Renta de los hogares de la Comunidad de Madrid que el INE publicó para las comunidades autónomas en julio de 2001, con los criterios del Sistema Europeo de Cuentas (SEC-95).

Tabla V.3.1. Cuenta de Renta de los hogares de la Comunidad de Madrid

Contabilidad Regional de España.Base 1995					
Cuentas de renta de los hogares.Serie 1995-1999					
MADRID (COMUNIDAD DE)					
Cuenta de asignación de la renta primaria	1995(P)	1996(P)	1997(P)	1998(P)	1999(A)
Unidad: millones de pesetas					
RECURSOS					
Excedente de explotación bruto / Renta mixta bruta	2.271.864	2.457.707	2.631.897	2.794.249	2.862.982
Remuneración de los asalariados	6.593.172	6.814.942	7.214.072	7.763.000	8.502.642
Rentas de la propiedad	931.017	940.231	841.674	818.063	807.402
EMPLEOS					
Rentas de la propiedad	360.697	379.194	361.099	346.885	352.177
Saldo de rentas primarias brutas	9.435.356	9.833.686	10.326.544	11.028.427	11.820.849
	6	6	4	7	9
Cuenta de distribución secundaria de la renta	1995(P)	1996(P)	1997(P)	1998(P)	1999(A)
Unidad: millones de pesetas					
RECURSOS					
Saldo de rentas primarias brutas	9.435.356	9.833.686	10.326.544	11.028.427	11.820.849
Prestacion. sociales distintas de las transf.soc.especie	1.437.002	1.502.155	1.554.481	1.597.175	1.680.351
Otras transferencias corrientes	680.913	704.795	728.287	793.041	849.530
EMPLEOS					
Impuestos corrientes sobre la renta, patrimonio, etc.	1.283.657	1.351.896	1.354.448	1.403.601	1.494.843
Cotizaciones sociales efectivas	1.686.572	1.767.001	1.844.985	1.951.427	2.155.892
Cotizaciones sociales imputadas	253.183	264.938	270.376	277.656	285.474
Otras transferencias corrientes	692.139	716.341	739.427	825.607	851.220
Renta disponible bruta	7.637.720	7.940.460	8.400.076	8.960.352	9.563.301

Fuente: INE.

El concepto de renta disponible bruta de los hogares suele ser también denominado como renta familiar disponible (RFD), y definido como “el nivel de renta de que disponen las economías domésticas para gastar y ahorrar, o bien como la suma de todos los ingresos efectivamente percibidos por las economías domésticas durante un período” (“la Caixa”, 2001). El conocimiento de la RFD suele constituir un objetivo de política económica (nacional, regional y local) de primer orden, como medida complementaria del nivel de bienestar de los ciudadanos. Efectivamente, su conocimiento y cuantificación, junto con otras macromagnitudes, resulta un instrumento adecuado para el establecimiento de políticas tendentes a la reducción de las desigualdades entre distintas zonas, en base a su localización o realidad socioeconómica. Además, las empresas están también interesadas en disponer de esta macromagnitud, para niveles espaciales cada vez más desagregados, por cuanto constituye un indicador del poder adquisitivo del mercado al que dirigen sus productos y servicios.

Según la contabilidad regional del INE (Tabla V.3.1.), la renta bruta disponible de los hogares sería el resultado de la siguiente identidad contable:

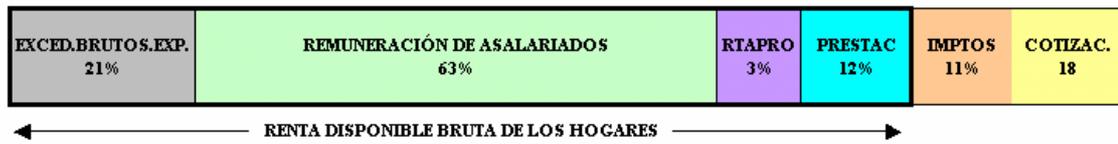
$$RFD = EBE + Remun + RtaPro. + Prestac + Transf. - Imptos - Cotiz \quad Eq. V.3.1.$$

siendo RFD:	renta familiar disponible o renta disponible bruta de los hogares
EBE:	excedente bruto de explotación (renta mixta bruta)
Remun:	remuneración de asalariados
RtaPro:	saldo de rentas de la propiedad
Prestac:	prestaciones sociales distintas de las transferencias sociales en especie
Transf:	saldo de transferencias corrientes
Imptos:	impuestos corrientes sobre la renta, patrimonio, etc.
Cotiz:	cotizaciones sociales efectivas, imputadas y corrientes

Es decir, la RFD podría considerarse como la suma total de ingresos procedentes del trabajo, más las rentas de capital, prestaciones sociales y transferencias, menos los impuestos directos pagados por las familias y las cuotas pagadas a la seguridad social. Éste debe constituir el punto de partida de cualquier estimación de RFD y, por tanto, de un ejercicio de predicción-extrapolación de la misma a territorios desagregados, como en el caso que nos ocupa.

De los datos de la Tabla V.3.1., puede obtenerse la participación en la renta bruta de los hogares (RFD más impuestos directos y cotizaciones sociales) de cada uno de los componentes de la misma en la Comunidad de Madrid, en 2000 (ver Figura V.3.1.).

Figura V.3.1. Composición de la renta bruta de los hogares en la Comunidad de Madrid (C.M.) en 1999



Fuente: Elaboración propia, a partir de INE.

- Puede observarse la importancia que, en general, tiene la partida de **remuneración de asalariados** (rentas salariales) en la composición de la renta bruta de los hogares que, en concreto, en la Comunidad de Madrid (C.M.) constituye el 63%. Este grupo representa un porcentaje aún mayor en zonas con una situación económica de crisis y malas perspectivas económicas.
- Los **excedentes brutos de explotación** (rentas mixtas) constituyen el 21% de la renta bruta total de los hogares de la CM, aunque aumenta este porcentaje en zonas con mayor número de agricultores propietarios o trabajadores autónomos.
- Las **rentas de la propiedad** o rentas de capital (RTAPRO) constituyen un apartado poco importante (3%, en la CM) de la renta bruta total, siendo mayores donde se concentran los rentistas y jubilados, así como en lugares prósperos, con buenas perspectivas económicas para la inversión.
- La partida correspondiente a las **rentas sociales** está constituida por las prestaciones sociales recibidas (PRESTAC) y las transferencias netas (TRANSF), grupo este último que, en la C.M., es prácticamente nulo. Las prestaciones sociales son un 12% de la renta bruta total y cobran especial importancia en lugares donde las pensiones y la protección al desempleo son superiores.
- Entre las deducciones sociales de la RFD, se encuentran los **impuestos directos** (IMPTOS) y las **cotizaciones a la seguridad social** (COTIZAC) que conjuntamente representan el 29% de la renta bruta total de los hogares.

Además, como se ha indicado al comienzo, esta primera etapa de fundamentación debe orientarse también a obtener un conocimiento suficiente de las teorías formuladas sobre esta magnitud económica.

En Pena (1996), se realiza un estudio muy completo sobre las **teorías microeconómicas** existentes, en el que se detallan los factores condicionantes de la renta personal (individual). Además, este trabajo constituye también una información complementaria de primera magnitud, por haber comprobado la eficacia de dichos condicionantes a través de información muy desagregada (a nivel de individuos y hogares) procedente de las Encuestas de Presupuestos Familiares (EPF) de 1973, 1980 y 1990, en España, además de las Encuestas Continuas de Presupuestos Familiares (ECPF) de 1985 a 1990. Pese al desfase temporal existente con el período de referencia de nuestro estudio

(2000), creemos que constituye una información de referencia válida, dada la inexistencia de estudios similares y la constancia en el tiempo de ciertos comportamientos humanos.

En la obra citada, se resumen las principales teorías microeconómicas explicativas de la distribución de la renta disponible personal. Aunque el sujeto del análisis es el individuo (o el hogar), creemos que podría constituir un buen punto de partida para un análisis microterritorial como el municipal, aunque la dimensión de algunas localidades, como el municipio de Madrid, las asemeje más al ámbito agregado que al microterritorio. Del grupo de teorías expuestas (teoría de la habilidad, teoría estocástica, teoría de la elección individual, teoría del capital humano, teorías de la desigualdad educativa, teoría del ciclo vital, teorías de la redistribución de los ingresos por las administraciones públicas y teorías de justicia distributiva), se podrían deducir los siguientes factores explicativos de la RFD:

- **Nivel de instrucción:** la educación formalizada y no formalizada es considerada, por unanimidad, como el factor más explicativo de la distribución de la renta disponible personal (individual), obteniendo en España una clara confirmación, a partir de las ECPF de 1985-1990.
- **Categoría socioprofesional:** combinación del nivel educativo con las ramas de actividad y situación profesional de los individuos. Los grupos con mayor RFD son los “directivos, gerentes, cuadros superiores no agrarios” y “empresarios no agrarios y profesionales liberales no agrarios”, habiéndose constatado un crecimiento importante de dicha renta en el grupo de profesionales dedicados a la agricultura.
- **Edad:** los ingresos individuales crecen con la edad hasta los años cercanos a la jubilación (60 ± 2 años), en que decrecen.
- **Sexo:** la mujer, con idénticas condiciones de nivel educativo y experiencia, tiene un diferencial negativo de ingresos respecto del hombre.
- **Factores macroeconómicos:** la actividad económica, especialmente los avances tecnológicos, así como la inflación y el desempleo, son asimismo condicionantes del nivel de RFD.
- **Sector público:** mediante sus políticas redistributivas, el sector público puede contribuir decisivamente a la reducción de las desigualdades sociales.
- **Factores locacionales:** que determinan diferentes oportunidades y rendimientos de los factores productivos y, por ende, de su distribución.
- **Clases de habitat:** el entorno en que se vive condiciona la forma de vida, el tipo de actividad y los ingresos que se perciben. Del análisis de las citadas EPF, se deduce que los hogares con mayor nivel de RFD suelen estar localizados en municipios mayores (más de 50.000 habitantes y capitales de provincia). Además, también se comprobó que las rentas salariales participan en la RFD en sentido inverso al tamaño de los municipios, mientras que los excedentes brutos de explotación

participan, en la RFD, en el mismo sentido. Efectivamente, en los municipios menores de 10.000 habitantes (rurales y semiurbanos) suelen residir los trabajadores independientes (zonas residenciales) y los empresarios agrícolas (zonas rurales), mientras que en los grandes municipios se concentran los asalariados.

- **Herencia:** aspectos genéticos, así como el entorno social y familiar, constituyen factores explicativos del mayor o menor nivel de riqueza personal.

Resulta también interesante la constatación, en España, en el período 1973-1990, de una evidente tendencia hacia una sociedad más igualitaria, dado que los índices de desigualdad se reducen a medida que pasa el tiempo y son también menores en las generaciones más jóvenes que en las más antiguas. Este resultado no es debido únicamente al pasivo transcurso del tiempo sino a los cambios de sensibilidad hacia el concepto de desigualdad que se han traducido en políticas de redistribución de las rentas y del aumento de la igualdad de oportunidades.

A modo de resumen de esta primera fase del proceso, podrían extraerse las siguientes conclusiones, que serán también **hipótesis** a contrastar en el análisis exploratorio espacial, que a su vez, estará encaminado a una posterior especificación del modelo ecológico:

- **Las rentas salariales constituyen el volumen principal de la RFD**, también en la CM, sobre todo en localidades de gran tamaño (como Madrid y los municipios metropolitanos de Alcalá de Henares, Fuenlabrada, Leganés o Móstoles), pero son menos representativas en municipios rurales, con mayor proporción de jubilados y rentistas, así como de profesionales y empresarios agrarios (Ambite, Berzosa del Lozoya y Puebla de la Sierra).
- **El nivel de instrucción es considerado como un factor de gran capacidad explicativa** de la RFD, siendo especialmente elevado en municipios residenciales, cercanos a los centros productivos (Boadilla del Monte y Las Rozas de Madrid) e inferior en núcleos rurales de la periferia norte y sureste de la CM (La Acebeda y Estremera).
- **La categoría socioprofesional es también explicativa de la RFD**, obteniendo los mayores niveles los grupos de “directivos y gerentes no agrarios”, así como los empresarios y profesionales liberales no agrarios” que, en la CM, se concentran en el Área Metropolitana y Sierra Central (Boadilla del Monte, Majadahonda, Pozuelo de Alarcón, Torreldones,...), siendo inferior peso relativo en pequeños núcleos rurales, como La Acebeda, Cervera de Buitrago, Madarcos, Puebla de la Sierra y Robledillo de la Jara, en la Sierra Norte.
- **Los ingresos individuales decrecen con la edad de la jubilación**, por lo que cabría esperar un menor nivel de RFD en municipios con mayor concentración de personas mayores, como La Acebeda, Navarredonda y San Mamés, Prádena del Rincón y Robregordo.

- **La actividad tecnológica es causa del incremento de la RFD** que, en la CM, es de gran importancia en ciertos municipios del Área Metropolitana (como Tres Cantos, Alcobendas y Las Rozas de Madrid) y Henares (Ajalvir).
- **El desempleo afecta negativamente a la RFD**, concentrándose las mayores tasas en núcleos rurales de la Sierra Norte (Berzosa del Lozoya, Cervera de Buitrago y Prádena del Rincón), así como en ciertos municipios semiurbanos de las zonas de Henares (Meco) y Suroeste (San Martín de Valdeiglesias). Por el contrario, los municipios con menores tasas de desempleo suelen situarse en municipios rurales con mayor concentración de personas jubiladas (La Hiruela y Brea de Tajo).
- **La localización geográfica es determinante en la distribución de la RFD**; es decir, los municipios situados cerca de núcleos con una población de alto nivel de RFD, tendrán más posibilidades de tener también mayor riqueza que las localidades próximas a zonas más deprimidas. Por tanto, cabrá esperar la existencia de efectos espaciales (autocorrelación y heterogeneidad espacial) en la distribución de la RFD de los municipios madrileños.
- **En los municipios rurales pueden producirse diferencias de cierta importancia en la distribución de la RFD**, dado que junto con núcleos prácticamente despoblados y muy envejecidos, como algunos municipios de la Sierra Norte de la CM (La Acebeda, Navarredonda y San Mamés, Prádena del Rincón y Robregordo), con apenas cien habitantes, existen también localidades en las que residen profesionales o empresarios agrícolas, como Batres, Berzosa del Lozoya y Olmeda de las Fuentes, con mayor nivel de renta.
- **El habitat poblacional suele condicionar los niveles de RFD**, dado que los hogares con mayor nivel de renta suelen situarse en grandes núcleos urbanos o en localidades cercanas a los mismos, residenciales, de tipo semiurbano. Por este motivo, sería de esperar también mayores concentraciones de RFD en torno a los municipios de Madrid capital y Área Metropolitana.

V.3.2. Selección y tratamiento de la información disponible

Las reflexiones derivadas del Apartado anterior deben constituir el punto de partida de esta segunda fase del proceso de predicción-extrapolación de datos microterritoriales, en la que se aborda el problema de la búsqueda de información relevante, como explicativa de la RFD⁷¹, y existente, no sólo para el ámbito ecológico de la estimación (provincial), sino también para el nivel desagregado de la extrapolación (municipios de la CM). Por eso, en primer lugar, se realiza un resumen de la relación de orígenes y factores explicativos de la RFD (Tabla V.3.2.), derivados tanto de la exposición del Apartado V.3.2., como de la relación de modelos propuestos para la estimación microterritorial de la RFD, realizada tanto en el Anexo II, de forma más extensa.

⁷¹ En adelante, nos referiremos con las siglas RFD a la renta disponible bruta de los hogares relativizada por la población (“per capita”).

A diferencia de los condicionantes y variables procedentes de las teorías económicas, que tienen, en principio, un carácter más general, las variables explicativas procedentes de modelos econométricos de estimación de renta son indicadores concretos, realmente disponibles para el ámbito municipal (total de España o para alguna comunidad autónoma o provincia). Como ya se avanzó en el Apartado V.2., la obtención de información estadística microterritorial ha constituido un gran escollo en la realización de ejercicios de predicción-extrapolación, sobre todo en países con estructuras políticas muy descentralizadas, como España, en los que muchas estadísticas quedan en manos de las comunidades autónomas, con planes estadísticos propios y diferenciados del resto.

Tabla V.3.2. Factores explicativos de la renta familiar disponible

Componentes de la RFD	Variables explicativas en las teorías microeconómicas	Variables explicativas en otros modelos de RFD
. Remuneración de asalariados: - salarios	. Nivel de instrucción elevado	. Actividades comerciales . Actividades empresariales
. Excedentes brutos de explot.: - beneficios empresarial.	. Categoría socioprofesional: - directiv., gerentes no agr. - profesionales lib. no agr.	. Actividades del turismo . Activos sector industrial . Base imponible IRPF
. Rentas de la propiedad: - intereses	. Edad: 65 y más años	. Cons. doméstico electricidad . Distancia a cabeceras servic. . Edificios más 4 plantas
. Prestaciones sociales: - jubilaciones - seguro de desempleo	. Sexo: mujer	. Empleo asalariado . Establecimientos turísticos
. Impuestos directos: - IRPF - Impuesto de sociedades	. Factores macroeconómicos: - actividad económica - avances tecnológicos - desempleo - inflación	. Índice socioeconómico . IVA/nº declarantes . Líneas telefónicas . Líneas teléfonos uso domést. . Líneas RDSI
. Cotizaciones seguridad social	. Políticas redistributivas sector público	. Locales activos . Parque de vehículos . Presup. de gastos municipales . Oficinas bancarias
	. Factores locacionales: cercanía a los centros generadores riqueza	. Saldos en cajas de ahorro . Segunda vivienda
	. Habitat: municipios mayores	. Tasas de demandas empleo . Tasa de instrucción superior . Tasa de empleo especializado
	. Herencia: familia, entorno	. Tasas de ocupación . Tasa de universitarios

Fuente: Elaboración propia.

Como puede apreciarse en la última columna de la Tabla V.3.2., la mayor parte de las variables explicativas de la RFD seleccionadas por los investigadores, son indicadores

de actividad económica, general o financiera, y variables relacionadas con el destino de la RFD (consumo y el ahorro) y variables de empleo. Por ejemplo:

- Variables de actividad económica general: líneas telefónicas, parque de camiones o parque total de vehículos, locales activos, licencias comerciales y establecimientos turísticos, IVA, segunda vivienda, etc.
- Variables de finanzas públicas: presupuestos de gastos municipales.
- Variables de consumo: líneas telefónicas de uso doméstico, parque de turismos, consumo de energía eléctrica de uso doméstico.
- Variables de ahorro: oficinas bancarias y saldos en cajas de ahorros.
- Variables de empleo: demandas de empleo (paro), tasa de ocupación, activos en la industria, categoría profesional.

Algunas de las variables citadas no se encuentran disponibles para todo el territorio nacional, conjuntamente para las 50 provincias españolas y los 179 municipios de la CM, como sucede con la recaudación del IVA, presupuestos de gastos municipales, consumo de energía eléctrica de uso doméstico, saldos en cajas de ahorros, tasas de ocupación y activos. Por ese motivo, quedarán descartadas de este estudio.

Por otro lado, ya se ha indicado que deben seleccionarse con precaución los indicadores de actividad económica general, dado que el sesgo de la agregación puede producir una cierta invalidación de la teoría económica, situando en los primeros lugares del ranking de RFD a localidades de pequeño tamaño con alguna actividad productiva de relevancia, por delante de localidades residenciales en las que se concentra la población de alto nivel de vida. Por ejemplo, en el caso de la Comunidad de Madrid, suele darse esta situación en municipios pequeños con presencia en los mismos de polígonos industriales (Ajalvir, Humanes de Madrid) o segunda vivienda de fin de semana (como Pelayos de la Presa), que son ordenados con mayor RFD que núcleos residenciales del estilo de Boadilla del Monte, Majadahonda, Las Rozas de Madrid o Pozuelo de Alarcón .

Este efecto distorsionante se ha comprobado que se produce muy habitualmente en todo el territorio nacional, con indicadores como actividades empresariales (industria, comercio, turismo), parque de automóviles y camiones, líneas telefónicas y segunda vivienda.

Por ejemplo, en la Tabla V.3.3., se presentan los 12 municipios con mayor/menor tasa de líneas telefónicas, actividades comerciales minoristas y parque de automóviles, por cada cien habitantes. Como puede observarse, la variable líneas telefónicas, que ha sido considerada una variable “clásica”, como explicativa de la RFD, por su gran capacidad explicativa en modelos ecológicos (ámbito provincial), resulta muy problemática en niveles microterritoriales, como ciertos municipios pequeños con alguna actividad económica importante, como el turismo, con altas tasas de segunda vivienda (Gargantilla del Lozoya y

Pinilla de Buitrago, Robregordo, Santa María de la Alameda, Horcajo de la Sierra, etc.) o polígonos industriales (Ajalvir).

En el extremo contrario, con bajas tasas de teléfonos por habitante, suelen situarse municipios dormitorio o, en general, densamente poblados, como Parla, Ciempozuelos, Móstoles o Alcalá de Henares. Por eso, sólo será válido el indicador de líneas telefónicas, como explicativo de la RFD municipal, si es previamente corregido de las líneas de uso no doméstico y aquéllas correspondientes a la segunda vivienda.

Tabla V.3.3. Municipios con tasas extremas de líneas telefónicas, actividades comerciales y parque de automóviles, por cien habitantes (año 2000)

Municipio	Población	Teléfonos	Municipio	Población	Teléfonos
Gargantilla del Lozoya y Pinilla de Buitrago	245	110%	Hiruela (La)	77	31%
Robregordo	88	108%	Parla	77.157	33%
Ambite	302	100%	Ciempozuelos	14.170	33%
Rozas de Puerto Real	273	98%	Puebla de la Sierra	96	34%
Pelayos de la Presa	1.422	91%	Móstoles	197.062	34%
Santa María de la Alameda	702	91%	Morata de Tajuña	5.485	35%
Madarcos	23	90%	Villaconejos	2.946	35%
Acebeda (La)	56	86%	Belmonte de Tajo	1.158	36%
Ribatejada	390	84%	Villarejo de Salvanés	5.753	36%
Venturada	919	82%	Tielmes	2.003	37%
Horcajo de la Sierra	146	82%	Alcalá de Henares	172.418	37%
Ajalvir	2.386	79%	Torres de la Alameda	4.541	37%

Municipio	Población	Comerc.	Municipio	Población	Comerc.
Puebla de la Sierra	96	5,6%	Atazar (El)	92	0,0%
Somosierra	98	5,4%	Cervera de Buitrago	88	0,0%
Berrueco (El)	334	4,8%	Madarcos	23	0,0%
Lozoyuela-Navas-Sieteiglesias	680	4,0%	Prádena del Rincón	103	0,0%
Lozoya	459	3,9%	Robregordo	88	0,0%
Pozuelo del Rey	181	3,7%	Villar del Olmo	1.517	0,6%
Serna del Monte (La)	115	3,5%	Valverde de Alcalá	309	0,7%
Humanes de Madrid	9.937	3,4%	Horcajo de la Sierra	146	0,7%
Ajalvir	2.386	3,4%	Meco	6.923	0,8%
Rascafría	1.637	3,3%	Berzosa del Lozoya	131	0,8%
Buitrago del Lozoya	1.565	3,3%	Anchuelo	579	0,9%
Olmeda de las Fuentes	146	3,3%	Navarredonda y San Mamés	106	0,9%

Municipio	Población	Autom.	Municipio	Población	Autom.
Torremocha de Jarama	351	283%	Lozoyuela-Navas-Sieteiglesias	680	19%
Colmenar del Arroyo	913	145%	Serna del Monte (La)	115	19%
Aldea del Fresno	1.503	88%	Horcajuelo de la Sierra	108	21%
Alcobendas	92.537	85%	Horcajo de la Sierra	146	23%
Escorial (El)	11.209	60%	Orusco de Tajuña	644	24%
Majadahonda	49.137	58%	Acebeda (La)	56	25%
Ajalvir	2.386	57%	Gargantilla del Lozoya y Pinilla de Buitrago	245	25%
Serranillos del Valle	901	56%	Boalo (El)	3.447	25%

Anchuelo	579	56%	Villavieja del Lozoya	185	27%
Pozuelo de Alarcón	68.470	56%	Cenicientos	1.835	27%
Cervera de Buitrago	88	56%	Fresno de Torote	722	27%
Torrejón de Ardoz	97.546	55%	Ambite	302	27%

Fuente: Elaboración propia

Parecidos resultados se obtienen con otras actividades económicas como, por ejemplo, la tasa de actividades comerciales minoristas, también muy empleada en modelos explicativos de RFD. Una vez más, los primeros municipios en el ranking de actividades comerciales minoristas son núcleos poco poblados, con una gran actividad turística (Puebla de la Sierra, Somosierra, Lozoyuela-Navas-Sieteiglesias, Lozoya, Rascafría, etc.) o industrial (Humanes de Madrid, Ajalvir).

En cuanto a la tasa de parque de automóviles, tan frecuentemente utilizada como indicador de consumo/renta de las familias, también llama a veces la atención por situar en los primeros puestos del ranking, a ciertos municipios de pequeña población en los que existen importantes actividades de fabricación, reparación o alquiler de vehículos (Torremocha del Jarama, Aldea del Fresno), junto a localidades residenciales como Alcobendas (que incluye la entidad de La Moraleja), El Escorial, Majadahonda o Pozuelo de Alarcón.

En el extremo opuesto, este indicador, al igual que el de actividades comerciales, suele situar en los últimos puestos a municipios rurales con escasa actividad económica, alejados de la capital y probablemente también con inferior RFD (El Atazar, Cenicientos, Madarcos, etc.).

Algunos autores han apostado por variables procedentes del impuesto sobre la renta de las personas físicas (IRPF), como indicadores especialmente relevantes en la explicación de la RFD aunque, como es bien sabido, padece algunas limitaciones por su incapacidad para captar rentas no declaradas procedentes de la economía sumergida (que en algunas localidades pueden constituir un volumen muy importante) o procedentes de ciertas actividades económicas (agrícola, autónomos), así como por la exclusión de las rentas mínimas exentas del impuesto.

Pese a las citadas limitaciones, también creemos que se trata de una variable importante y queda para futuras aplicaciones su empleo como variable explicativa de la RFD. No se ha podido incluir en este estudio por no encontrarse disponible para todas las provincias españolas (País Vasco y Navarra), ni estar suficientemente actualizada.

Por tanto, a la luz de la experiencia acumulada por el Instituto L. R. Klein en estos últimos años, así como de un análisis exhaustivo de gran cantidad de variables económicas y sociodemográficas⁷², se propone (Tabla V.3.4.) una relación de indicadores con buena

⁷² Entre las variables no citadas en la Tabla V.3.4., que han sido sometidas a pruebas y finalmente no seleccionadas como posibles explicativas de la RFD municipal, podrían citarse las siguientes: demográficas

capacidad explicativa, “a priori”, de la RFD provincial⁷³, disponibles además para el ámbito municipal de la CM, con una actualización lo más cercana posible al año de referencia (año 2000). Para cada una de las variables explicativas, se presenta también el coeficiente de correlación lineal procedente de su relación con la RFD en el ámbito provincial.

Tabla V.3.4. Selección de variables explicativas de la RFD

Sigla	Variable explicativa de la RFD	Año	Fuente estadística	Corr.
SE2T	<i>Actividades profesionales independientes</i>	2000	Ministerio Economía, Hacien. Forales del P. Vasco y Navarra	0,6
DEM00	<i>Demandas de empleo</i>	2000	INEM	-0,7
PPDSPE1	<i>Directivos, gerentes y empresarios no agrarios</i>	1996	Instituto de Estadística de la Comunidad de Madrid	0,7
DIST96	<i>Distancia a municipios cabeceras de servicios</i>	2000	Elaboración propia	-0,8
PPD23G	<i>Instrucción de 2º y 3º grado</i>	1996	Instituto de Estadística de la Comunidad de Madrid	0,7
TELFPO0	<i>Líneas telefónicas uso doméstico primera vivienda</i>	2000	Telefónica y elaboración propia	0,7
RDSI00	<i>Líneas RDSI</i>	2000	Telefónica	0,7
PREVI00	<i>Precio del metro cuadrado de la vivienda</i>	2000	Ministerio de Fomento	0,6

Fuente: Elaboración propia.

Un buen criterio de selección debería fundamentarse en los siguientes principios:

- Alto coeficiente de correlación entre las variables explicativas y la endógena (RFD) en el ámbito agregado (provincial).
- Detección y corrección (cuando sea posible) de posibles distorsiones sufridas por algunas variables que, con alto poder explicativo en el ámbito agregado (provincial)

(población femenina total, población de 65 y más años, índice de dependencia, población en municipios rurales), empleo (paro registrado total, sexo, sectores de actividad), actividad económica (empresarial, industrial, comercio mayorista, comercio minorista, grandes superficies, centros comerciales, turismo, total de líneas telefónicas, parque de automóviles, motocicletas y camiones, oficinas de crédito, segunda vivienda), culturales (centros educativos privados, pantallas de cine), sanitarias (camas de hospital, centros y residencias de mayores), políticas (votos al PP, PSOE), religiosas (número de comunidades religiosas y centros educativos católicos).

⁷³ Como se indica en el Apartado siguiente, se ha realizado una estimación de la RFD provincial del año 2000, a partir de las series provinciales proporcionadas por el INE para el período 1995-1999, mediante un modelo de datos de panel de efectos fijos. Esta nueva serie de RFD provincial de 2000 será la utilizada en el modelo ecológico para estimar la RFD municipal de la CM en 2000.

presentan valores atípicos en su distribución microterritorial (municipal), como los casos presentados en la Tabla V.3.3.

- Distribución microterritorial de las variables explicativas acorde con las hipótesis de partida, formuladas en el Apartado V.3.1.

Estas variables seleccionadas, así como la propia variable endógena de RFD, disponible para el ámbito provincial de España, deberán ser sometidas a un doble análisis exploratorio, tanto para el ámbito provincial como municipal de la CM, que se realizará en la tercera etapa del proceso.

V.3.3. Análisis exploratorio espacial

En esta etapa, se utilizarán las herramientas gráficas y estadísticas del AEDE que, como se indicó en el Capítulo III, están especialmente diseñadas para encontrar algún “sentido” en los datos, es decir, para la detección en ellos de estructuras o conglomerados (“clusters”), sobre todo en casos como éstos, en los que se produce una cierta incertidumbre sobre la distribución del fenómeno (RFD municipal). Efectivamente, dado que no siempre existe una relación directa entre relaciones funcionales establecidas en distintos ámbitos territoriales, el análisis exploratorio habrá de realizarse para el conjunto de variables en el doble nivel analizado (provincial y municipal, en este caso), de forma que sea posible extraer formas más o menos estables, a partir de las cuales construir relaciones confirmatorias “válidas” en ambos niveles territoriales.

V.3.4.1. Análisis exploratorio de datos municipales

En primer lugar, se realiza un AEDE de las variables explicativas seleccionadas con el objetivo de analizar si presentan puntos atípicos (“outliers”) o algún tipo de efecto de dependencia y/o heterogeneidad espacial. Obviamente, en el ámbito municipal, no será posible realizar este tipo de análisis sobre la variable endógena de RFD, puesto que se trata del objetivo a lograr (su estimación y análisis). Los instrumentos utilizados serán el gráfico de caja (“box map”), diagrama y mapa de dispersión de Moran (“Moran scatterplot”), y mapa LISA, para el análisis de dependencia espacial univariante local y global. El cálculo de los estadísticos se ha realizado con el programa SpaceStat y su representación gráfica, con la ayuda de las extensiones SpaceStat para ArcView y DynESDA.

Para contrastar la presencia de autocorrelación espacial, se ha empleado una matriz de interacciones o contigüidades espaciales, junto con una matriz de pesos espaciales calculada como el cuadrado de la inversa de la distancia entre cada par de municipios, ambas estandarizadas por filas. La elección de estas matrices puede justificarse por cuanto se ha considerado que las relaciones intermunicipales subyacentes a la distribución de la RFD en este ámbito, son muy dependientes de la cercanía física y el “contacto” entre municipios.

En la Tabla V.3.5., se han presentado únicamente los resultados obtenidos por el test I de Moran, para las dos matrices de pesos espaciales definidas, en cada una de las variables explicativas de la RFD municipal. Tanto en el caso del test I de Moran, como el de Geary y el definido como Nueva- G_i^* , de Getis y Ord, producen un resultado unánime, en el sentido de la existencia de autocorrelación espacial positiva en todas las variables seleccionadas. Este resultado se pone aún más claramente de manifiesto en los gráficos del AEDE que se presentan, variable a variable, a continuación. En estos gráficos, puede también observarse la clara existencia, en todas las variables, de puntos calientes en los que se concentran los valores extremos (altos/bajos) de cada una de ellas, indicativos de autocorrelación espacial local, tal como se expuso ampliamente en el Apartado III.1.3.

Tabla V.3.5. Selección de variables explicativas de la RFD

MORAN'S I TEST FOR SPATIAL AUTOCORRELATION (normal approximation)						
DATA SET:	MUNCAM					
VARIABLE	WEIGHT	I	MEAN	ST.DEV.	Z-VALUE	PROB
TELF00	CAMW1S	0.6004018	-0.006	0.047904	12.650693	0.000000
RDSI00	CAMW1S	0.4667796	-0.006	0.047904	9.861323	0.000000
SE2T	CAMW1S	0.454155	-0.006	0.047904	9.597784	0.000000
PPD23G96	CAMW1S	0.5650946	-0.006	0.047904	11.913653	0.000000
PPDSPE1	CAMW1S	0.4631261	-0.006	0.047904	9.785055	0.000000
DEM00	CAMW1S	0.2787047	-0.006	0.047904	5.935250	0.000000
DIST96	CAMW1S	0.9083302	-0.006	0.047904	19.078714	0.000000
PREVI00	CAMW1S	0.6808749	-0.006	0.047904	14.330573	0.000000
TELF00	DINVS_2	0.3358595	-0.006	0.026229	13.019107	0.000000
RDSI00	DINVS_2	0.2944231	-0.006	0.026229	11.439312	0.000000
SE2T	DINVS_2	0.3089416	-0.006	0.026229	11.992840	0.000000
TP6500	DINVS_2	0.5237079	-0.006	0.026229	20.180983	0.000000
PPD23G96	DINVS_2	0.355245	-0.006	0.026229	13.758194	0.000000
PPDSPE1	DINVS_2	0.2850065	-0.006	0.026229	11.080296	0.000000
DEM00	DINVS_2	0.1467121	-0.006	0.026229	5.807707	0.000000
DIST96	DINVS_2	0.6669727	-0.006	0.026229	25.643070	0.000000
PREVI00	DINVS_2	0.3705994	-0.006	0.026229	14.343593	0.000000

Fuente: Elaboración propia a partir de SpaceStat

Más adelante, se presentan las variables seleccionadas, para las que se incluye una definición, fuente estadística, cinco municipios con valores extremos (altos/bajos) y gráficos del AEDE: diagrama de caja, diagrama de dispersión de Moran, mapa de dispersión de Moran y mapa LISA de autocorrelación espacial local, que se describen seguidamente.

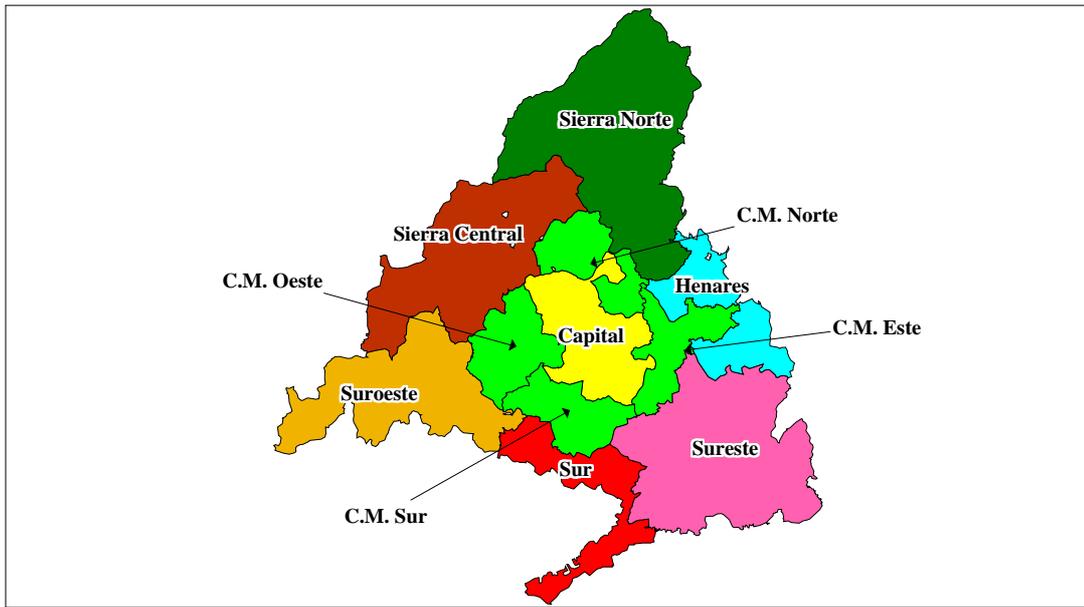
- **Mapa del diagrama de caja** (superior izquierdo): representa los cuartiles y las cotas superior e inferior, obtenidas como el producto de los valores del tercer y primer cuartil por 1,5 veces el recorrido intercuartílico, que vienen representados por los colores rojo oscuro (>99%) y azul oscuro (<1%), respectivamente, que destacan los valores atípicos. Por ejemplo, la variable de profesionales independientes registra valores atípicos en Olmeda de las Fuentes y Hoyo de

Manzanares, donde la proporción de impuestos pagados por los profesionales liberales, respecto del total del impuesto de actividades económicas, es muy superior a la registrada por el resto de municipios.

- **Diagrama** (superior derecho) y **mapa de dispersión de Moran** (inferior derecho): ambos dividen el tipo de asociación espacial en cuatro categorías. La dependencia espacial positiva se representa en los cuadrantes I y III del diagrama, es decir, valores altos de una variable rodeados de valores altos (color rojo en el mapa) o valores bajos rodeados de valores bajos (color rojo claro). La dependencia espacial negativa se sitúa en los cuadrantes II y IV del diagrama, es decir, municipios con valores altos/bajos de dicha variable (color azul oscuro) y bajos/altos (azul claro) rodeados por municipios con valores altos/bajos de dicha variable, respectivamente. Así, las variables positivamente relacionadas con la RFD (todas excepto la tasa de demandas de empleo y distancia al municipio cabecera) suelen experimentar una fuerte dependencia espacial positiva de valores altos en torno a la capital, el Área Metropolitana y Sierra Central, mientras que los valores bajos se sitúan en las periferias norte, sureste y suroeste.
- **Mapa LISA** (inferior izquierdo): destaca los municipios con valores significativos en el estadístico de asociación espacial local I_i local de Moran, poniendo de manifiesto, a través de una gradación de colores, la presencia de puntos calientes (“hot spots”) o atípicos espaciales, cuya mayor o menor intensidad dependerá de la significatividad asociada del citado estadístico. Por ejemplo, la variable de tasa de profesionales independientes experimenta una mayor significatividad del test I local ($p = 0,001$) en los municipios con altos valores en la misma, que están situados en la Sierra Central. Con menor significatividad se presentan las zonas de concentración de valores bajos de esta variable ($p = 0,01$ y $p = 0,05$), situadas en la Sierra Norte y Sureste de la CM.

Como puede observarse, nos referimos a menudo a la clasificación municipal del Instituto de Estadística de la Comunidad de Madrid, que divide esta Comunidad en 8 zonas: Capital, Área Metropolitana (a su vez dividida en Norte, Sur, Este y Oeste), Sierra Norte, Sierra Central, Suroeste, Sur, Sureste y Henares. Estas zonas, que se representan en la Figura V.3.2., se han dibujado también en los mapas del AEDE, para facilitar mejor la identificación de los municipios a través de las zonas.

Figura V.3.2. Zonas municipales de la Comunidad de Madrid

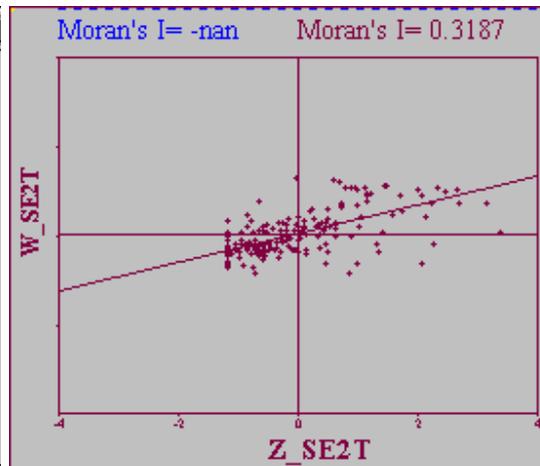
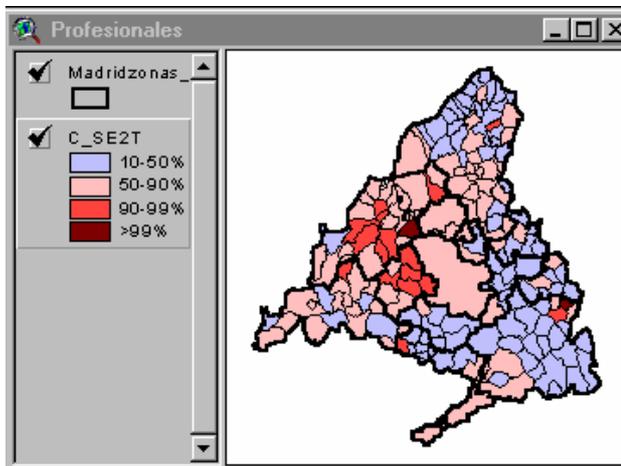


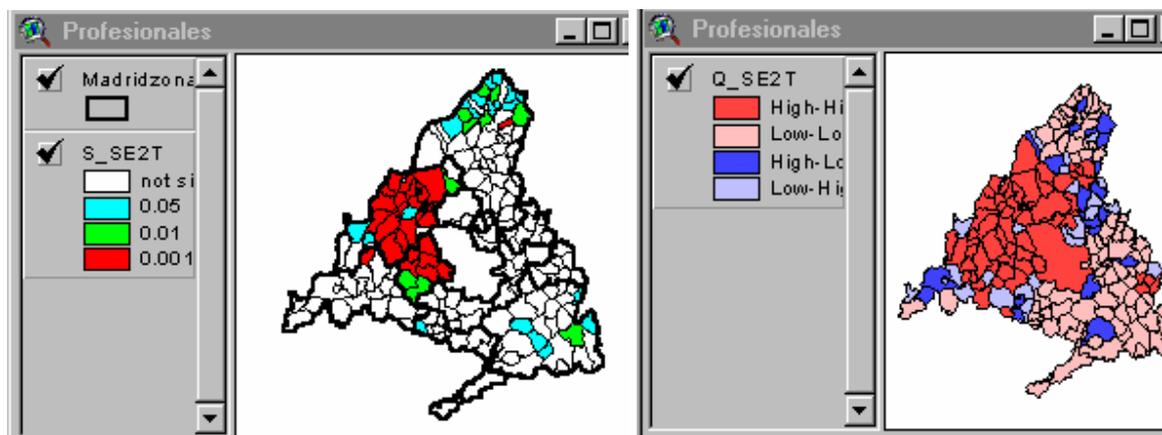
Fuente: Elaboración propia, a partir de MapInfo Professional.

➤ **Actividades profesionales independientes**

Impuesto de actividades económicas (IAE) pagado por los residentes que, en un municipio, ejercen una profesión liberal, así como empresarios y autónomos que realizan una actividad por cuenta propia. Esta variable (IAE de actividades profesionales) se relativiza por el impuesto total del IAE pagado en el municipio. Fuente: Ministerio de Economía y Hacienda.

Municipio	Población	Profes.	Municipio	Población	Profes.
Olmeda de las Fuentes	146	32%	Acebeda (La)	57	0%
Hoyo de Manzanares	6.013	30%	Braojos	160	0%
Villanueva de la Cañada	11.429	27%	Brea de Tajo	458	0%
Galapagar	23.693	27%	Canencia	441	0%
Collado Mediano	4.766	26%	Cervera de Buitrago	85	0%



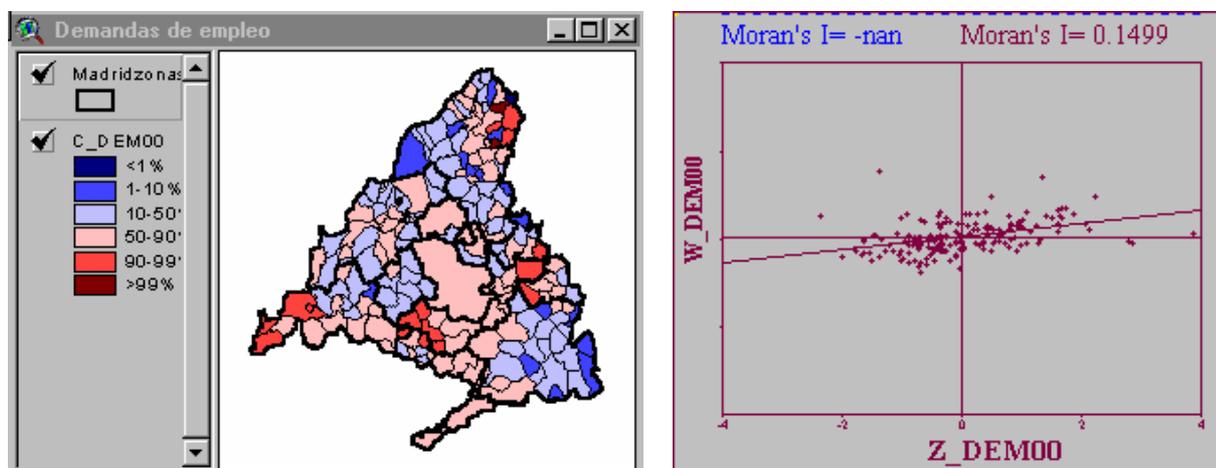


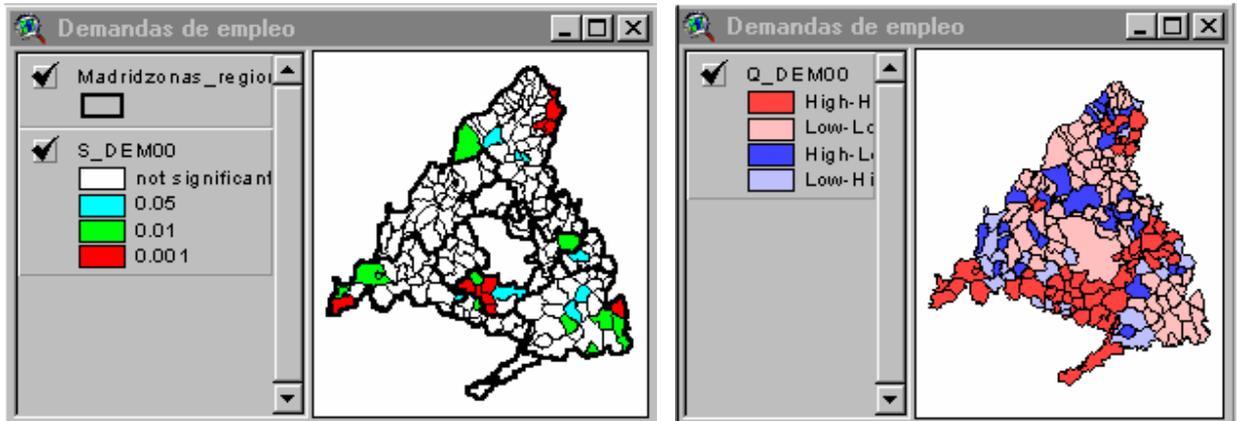
Como puede apreciarse en los mapas del diagrama de caja (superior izquierdo) y dispersión de Moran (inferior derecho), esta variable presenta autocorrelación positiva, sobre todo en la zona centro y oeste, produciéndose puntos atípicos calientes (gran concentración de valores altos de la variable) en los municipios de la Corona Metropolitana Oeste y Sierra Central (color rojo en la mapa LISA, en el cuadro inferior derecho).

➤ **Demandas de empleo**

Número total de demandantes de empleo en oficinas del INEM, relativizado por la población de 15 y más años (variable “proxy” de la población potencialmente activa). Fuente: INEM.

Municipio	Población Deman.	Municipio	Población Deman.
Cervera de Buitrago	85 15%	Hiruela (La)	74 2%
Prádena del Rincón	95 13%	Brea de Tajo	458 2%
Berzosa del Lozoya	118 13%	Redueña	179 3%
San Martín de Valdeiglesias	6.089 12%	Valdepiélagos	305 3%
Meco	6.328 11%	Valdelaguna	560 3%



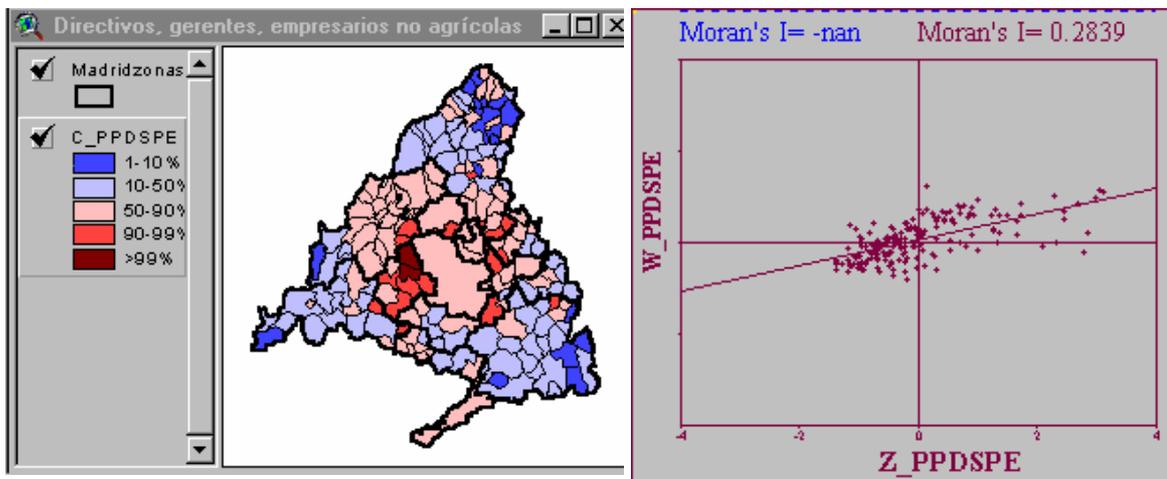


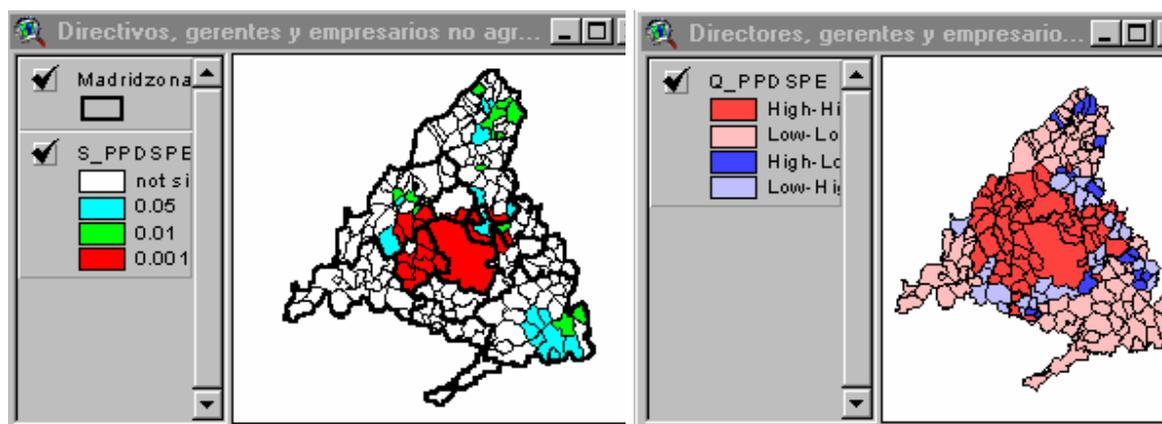
En el mapa del diagrama de caja (superior izquierdo) se observa que, en la Sierra Norte hay municipios con valores atípicamente elevados (color rojo oscuro) en la tasa de demandas de empleo, como Cervera de Buitrado y Prádena del Rincón, y valores bajos (color azul oscuro), como es el caso de La Hiruela. Asimismo, el mapa LISA (inferior izquierdo) destaca también la existencia de zonas calientes de municipios con altas tasas de paro en el Suroeste (Cenicientos) y Corona Metropolitana Sur (Móstoles, Parla y Fuenlabrada).

➤ **Directivos, gerentes y empresarios no agrarios**

Número de directores, gerentes y empresarios, no agrarios, relativizado por la población de 15 y más años. Fuente: Instituto de Estadística de la Comunidad de Madrid.

Municipio	Población Direct.	Municipio	Población Direct.
Rozas de Madrid (Las)	54.676 15%	Cervera de Buitrago	85 0%
Majadahonda	45.819 15%	Robledillo de la Jara	91 0%
Boadilla del Monte	20.686 15%	Madarcos	21 0%
Valdeolmos-Alalpartdo	1.775 14%	Acebeda (La)	57 0%
Venturada	819 14%	Puebla de la Sierra	71 0%





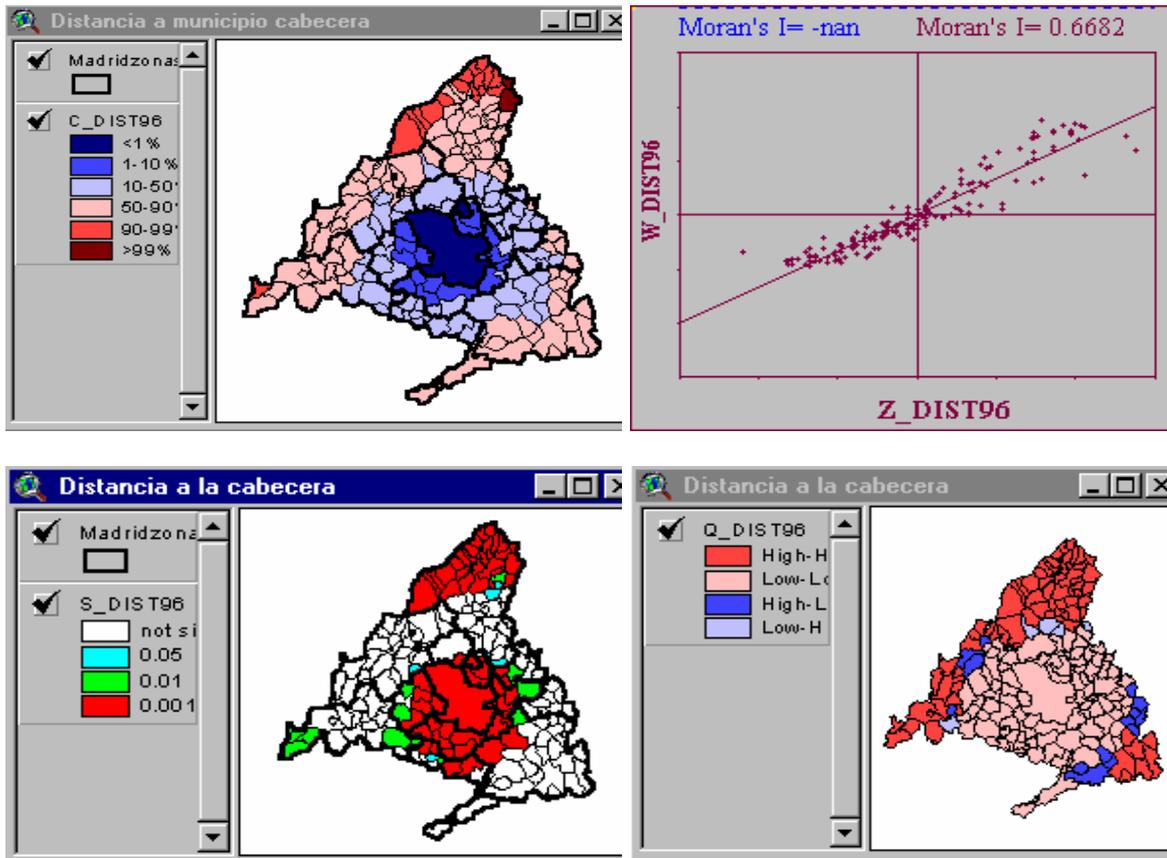
La distribución espacial de esta variable es similar a la experimentada por la tasa de actividades profesionales independientes, con una gran autocorrelación espacial positiva, produciéndose una concentración de valores altos en las zonas centro y oeste, que es lo que se desprende del mapa de dispersión de Moran (en el cuadro inferior derecho). Por su parte, el mapa del diagrama de caja destaca como puntos atípicos los valores obtenidos por este indicador en Las Rozas de Madrid y Majadahonda (Corona Metropolitana Oeste), no existiendo en este caso municipios con tasas por debajo de la cota inferior. La menor proporción de directivos se produce en los municipios de la periferia.

➤ Distancia a municipios cabeceras de servicios

Distancia, en kilómetros, a Madrid capital, en cuanto que se trata del municipio cabecera de área comercial⁷⁴ sobre el que gravitan comercialmente los 179 municipios de la CM. Fuente: elaboración propia, a partir de la actualización de los flujos comerciales realizada, con carácter anual, por el Instituto L. R. Klein.

Municipio	Población Deman.	Municipio	Población Deman.
Puebla de la Sierra	71 108,0	Madrid	2.882.860 0,0
Hiruela (La)	74 105,0	Coslada	77.057 11,5
Horcajuelo de la Sierra	103 94,0	Leganés	172.049 12,0
Rascafría	1.615 94,0	Pozuelo de Alarcón	66.298 12,0
Somosierra	93 93,0	Alcorcón	144.636 13,0

⁷⁴ Se entiende por municipio cabecera de área comercial aquel municipio sobre el que gravita comercialmente, o se siente atraída, la población del resto de municipios integrados en la misma área comercial. Estas localidades pueden ser consideradas, en general, como centros comerciales y de servicios ("la Caixa", 2001; Chasco, 1988; Chasco, 1997).

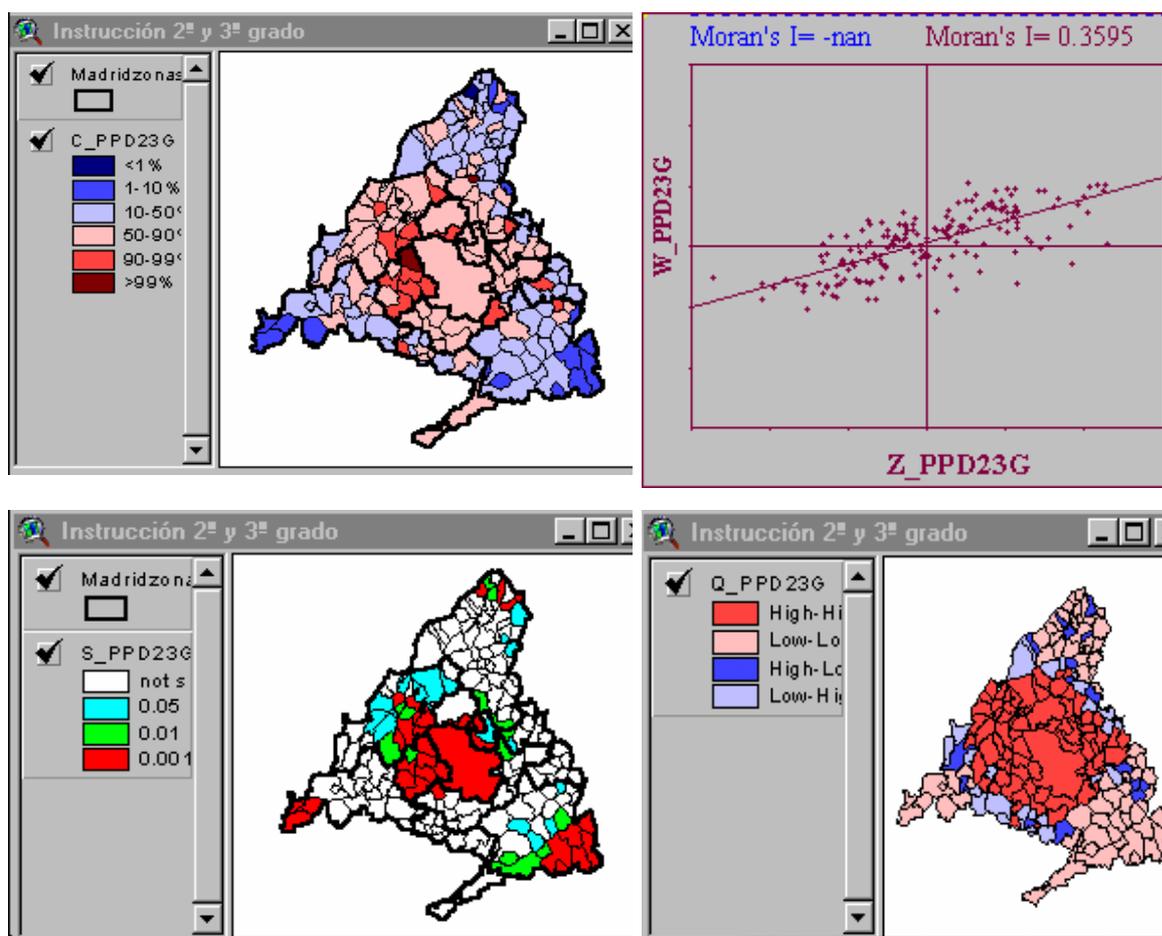


Dado que el municipio de Madrid es la cabecera de área comercial de todos los municipios de la CM, esta variable experimenta una distribución concéntrica uniforme, obteniendo valores muy altos en las localidades más alejadas del centro, como es el caso de Puebla de la Sierra y La Hiruela, situadas a más de 100 km. de la capital.

➤ **Instrucción de 2º y 3º grado**

Número de personas con título universitario o de bachiller (y asimilados) sobre el total de población de 15 y más años. Fuente: Instituto de Estadística de la Comunidad de Madrid.

Municipio	Población Instruc.	Municipio	Población Instruc.
Venturada	819 71%	Acebeda (La)	57 19%
Rozas de Madrid (Las)	54.676 70%	Prádena del Rincón	95 25%
Boadilla del Monte	20.686 70%	Estremera	1.028 25%
Majadahonda	45.819 69%	Valdaracete	617 27%
Torrelodones	13.467 68%	Horcajuelo de la Sierra	103 28%



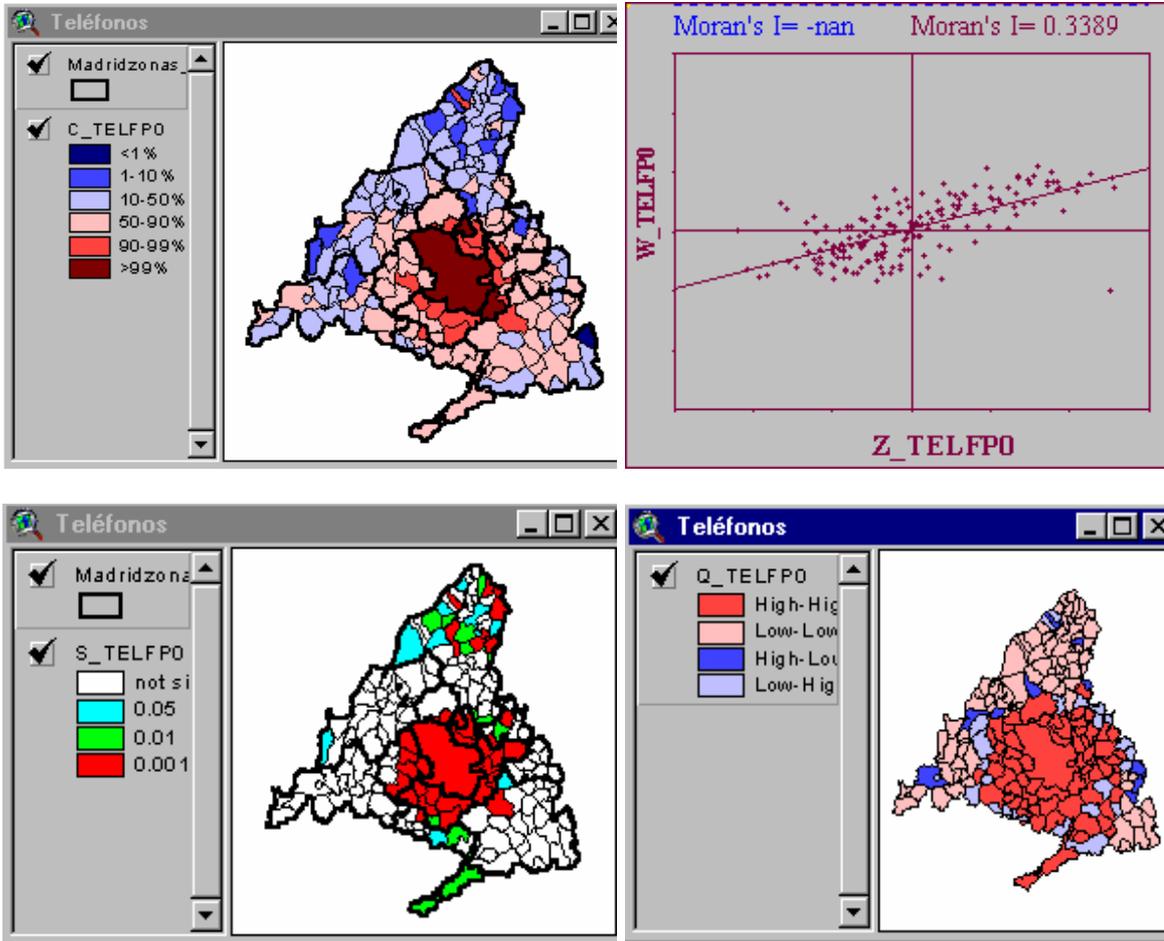
En esta variable, como sucede también con las tasas de actividades profesionales y directivos, el mapa de dispersión de Moran (inferior derecho) presenta una gran mancha roja (valores altos de la variable rodeados de valores altos) en los municipios del centro metropolitano y, sobre todo, en la Sierra Oeste. Por otro lado, del mapa LISA se deduce también que las tasas más bajas de instrucción se localizan en los extremos Norte (La Acebeda, Horcajuelo de la Sierra y Prádena del Rincón), Suroeste (Cenicientos y Las Rozas de Puerto Real) y Sureste (Brea de Tajo, Estremera y Valdaracete).

➤ **Líneas telefónicas de uso doméstico en primera vivienda**

Número total de líneas telefónicas instaladas, corregido por la proporción de líneas de uso no doméstico (negocios, centralitas,...) y por la proporción de segunda vivienda existente en un municipio. Esta variable está relativizada por la población. Fuente: Telefónica y elaboración propia.

Municipio	Población	Telef.	Municipio	Población	Telef.
Madrid	2.882.860	35%	Brea de Tajo	458	8%
Rivas-Vaciamadrid	29.092	35%	Puebla de la Sierra	71	9%
Gascones	107	35%	Hiruella (La)	74	9%

San Fernando de Henares	32.364	34%	Montejo de la Sierra	313	10%
Alcobendas	89.612	33%	Cervera de Buitrago	85	10%



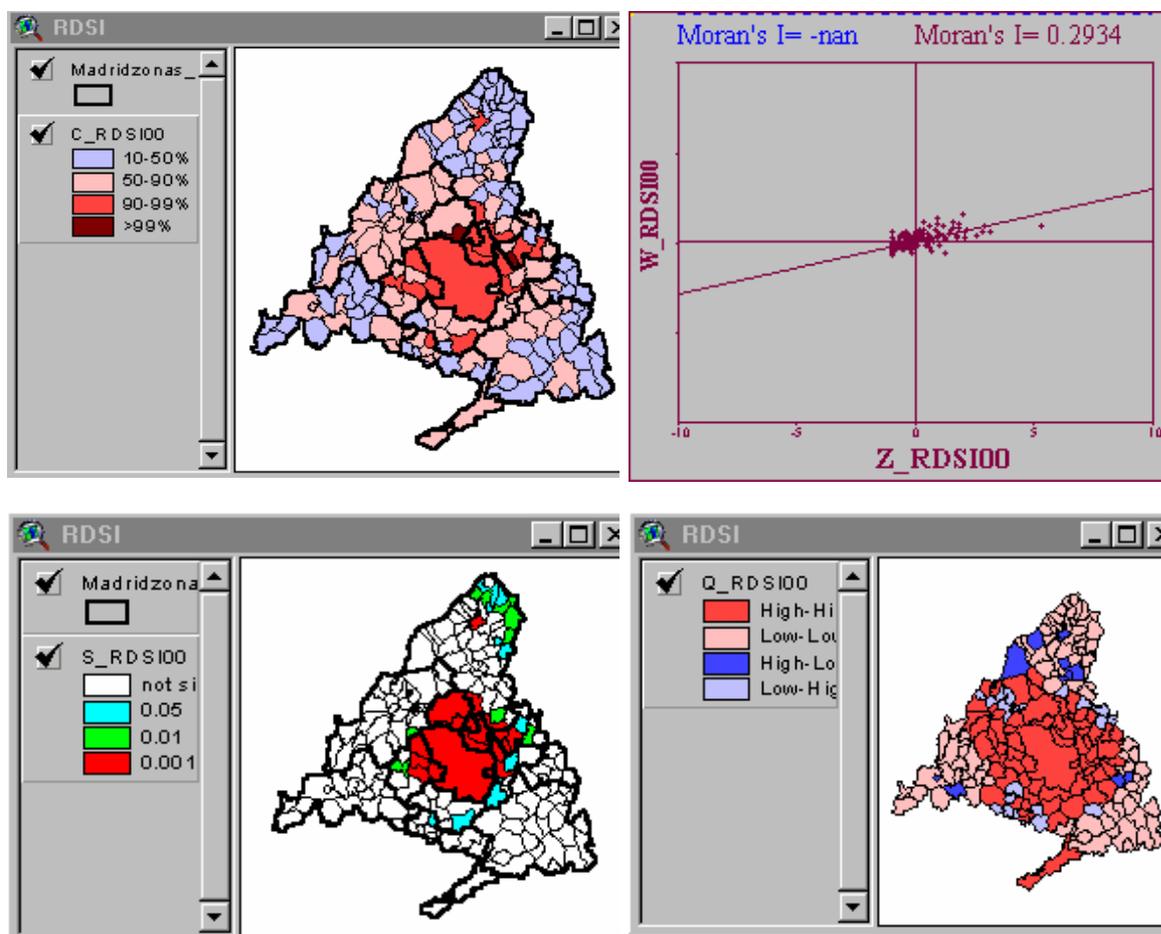
La tasa de teléfonos domésticos es una variable espacialmente autocorrelacionada, cuyos más altos valores se localizan en el centro, este y sur, donde se concentra gran parte de la población madrileña. La propia capital, con Rivas-Vaciamadrid, constituye un punto atípico de alta concentración de teléfonos (color rojo oscuro en el mapa del diagrama de caja), destacando en el extremo contrario la localidad de Brea de Tajo (Sureste). El mapa LISA muestra también en el extremo Norte zonas frías de valores bajos de esta variable.

➤ **Líneas RDSI**

Número de líneas RDSI por habitante. Fuente: Telefónica.

Municipio	Población	RDSI	Municipio	Población	RDSI
Ajalvir	2.205	7%	Acebeda (La)	57	0%
Tres Cantos	35.046	5%	Alameda del Valle	198	0%
Alcobendas	89.612	4%	Anchuelo	579	0%
Rozas de Madrid (Las)	54.676	4%	Atazar (El)	87	0%

San Agustín de Guadalix	5.859	4%	Belmonte de Tajo	1.158	0%
-------------------------	-------	----	------------------	-------	----



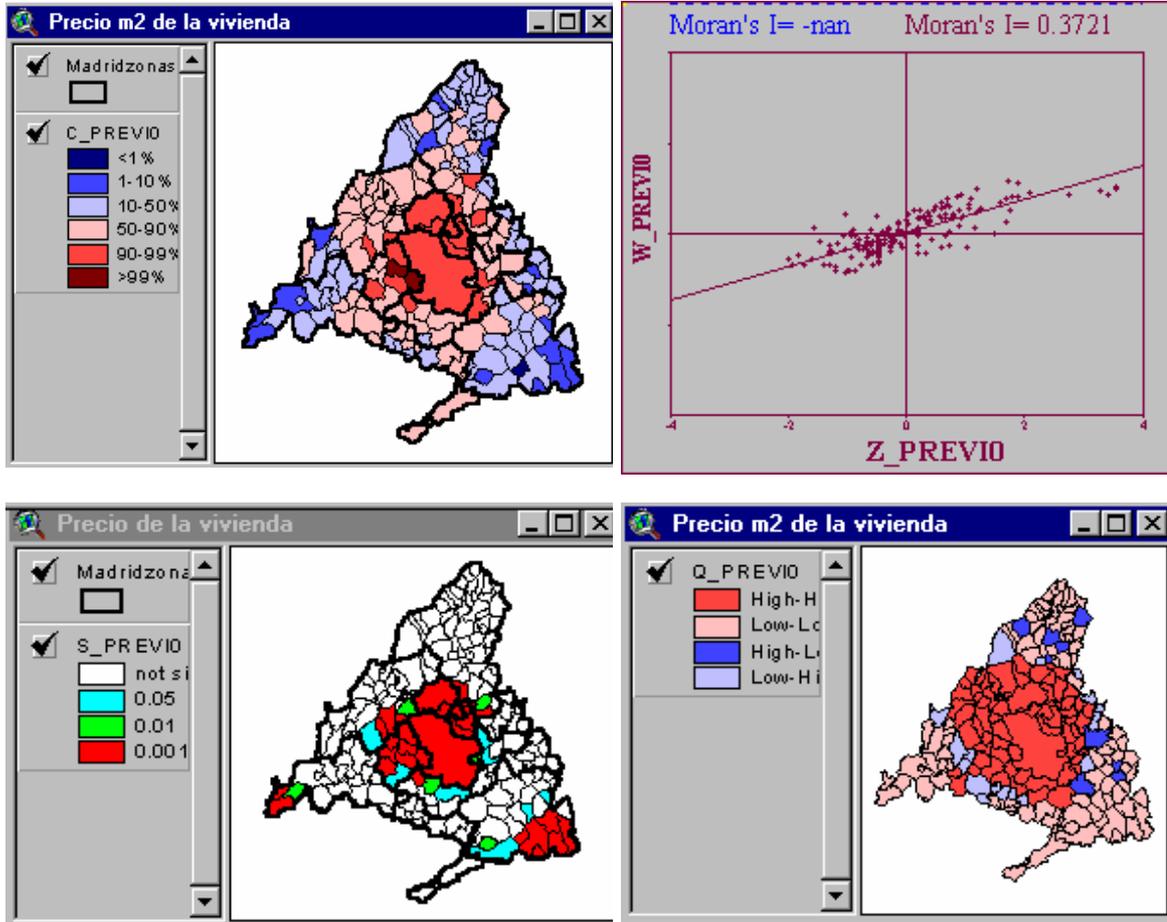
La tasa de líneas RDSI vuelve a presentar el mismo esquema de autocorrelación espacial, con gran concentración de valores altos en los municipios del Área Metropolitana, Sierra Central, Henares y Sur. En estas zonas reside la población joven, con mayor nivel de instrucción, así como también se localizan los parques empresariales que son, a la vez, causa y consecuencia de la implantación en el municipio de zonas residenciales. En esta línea, el municipio residencial y empresarial de Tres Cantos destaca por ser el único que supera la cota superior del diagrama de caja de esta variable (mapa superior izquierdo). Una vez más, en color azul, este mismo mapa pone al descubierto que los municipios situados en los extremos de la CM son los más alejados de las nuevas tecnologías.

➤ **Precio del metro cuadrado de la vivienda**

Precio del metro cuadrado de la vivienda. Fuente: Ministerio de Fomento y elaboración propia.

Municipio	Población Vivien.	Municipio	Población Vivien.
Majadahonda	45.819 305.673	Belmonte de Tajo	1.158 68.863

Pozuelo de Alarcón	66.298	304.758	Rozas de Puerto Real	283	71.800
Madrid	2.882.860	300.052	Villaconejos	2.870	76.393
Alcobendas	89.612	294.594	Cenicientos	1.864	80.633
Tres Cantos	35.046	271.820	Valdaracete	617	86.732



Por último, la variable precio del metro cuadrado de la vivienda vuelve a presentar una distribución muy similar a la experimentada por la mayoría de las variables anteriores, con altos valores en los municipios del Área Metropolitana y Sierra Central, entre los que destacan con atípicos Majadahonda y Pozuelo de Alarcón. El menor precio de la vivienda se produce en los municipios rurales de Norte, Suroeste y Sureste, obteniendo un valor atípicamente bajo la localidad de Belmonte de Tajo (Sureste). El mapa LISA localiza tres zonas calientes: el centro (capital, Corona Metropolitana Norte y Oeste), en los que se produce una concentración de valores muy altos de la variable, y los extremos suroeste y sureste, dos agrupaciones de municipios con bajo valor de la vivienda.

V.3.4.2. Análisis exploratorio de datos provinciales

La distribución de las variables endógena (RFD) y exógenas del modelo ecológico (provincial) será muy determinante de los resultados del mismo, por lo que dichas variables serán también analizadas con técnicas del AEDE. Para ello, se han calculado gráficos y estadísticos para el análisis univariante (diagrama y mapa de dispersión de Moran, mapa LISA), así como un diagrama de dispersión entre cada variable exógena y la endógena. El cálculo de los estadísticos se ha realizado con el programa SpaceStat, y su representación gráfica, con la ayuda de las extensiones SpaceStat para ArcView y DynESDA.

Para contrastar la presencia de autocorrelación espacial, se han empleado las mismas matrices, de contigüidades e inversa de la distancia cuadrática, entre provincias, en este caso, también estandarizadas por filas.

En la Tabla V.3.6., se han presentado los resultados obtenidos por el test I de Moran, para las dos matrices de pesos espaciales definidas en cada una de las variables explicativas y la propia RFD provincial (aunque también se han calculado, como en el caso anterior, los tests c de Geary y Nueva- G_i^* , de Getis y Ord), que indican unánimemente la existencia de autocorrelación espacial positiva en todas las variables. Este resultado se pone aún más claramente de manifiesto en los gráficos del AEDE, en los que puede también observarse la existencia de puntos calientes donde se concentran los valores extremos (altos/bajos) de cada una de ellas, indicativos de autocorrelación espacial local.

Tabla V.3.6. Selección de variables explicativas de la RFD

MORAN'S I TEST FOR SPATIAL AUTOCORRELATION (normal approximation)						
DATA SET:	PROV					
VARIABLE	WEIGHT	I	MEAN	ST.DEV.	Z-VALUE	PROB
RTAH00	PROVW1S	0.7779849	-0.020	0.095534	8.357133	0.000000
TELF00	PROVW1S	0.3985707	-0.020	0.095534	4.385637	0.000012
RDSI00	PROVW1S	0.5112612	-0.020	0.095534	5.565218	0.000000
SE2H	PROVW1S	0.2926161	-0.020	0.095534	3.276564	0.001051
PD23G91	PROVW1S	0.6318992	-0.020	0.095534	6.827989	0.000000
DEM00	PROVW1S	0.7416907	-0.020	0.095534	7.977225	0.000000
PDSPE91	PROVW1S	0.3965698	-0.020	0.095534	4.364693	0.000013
DIST96	PROVW1S	0.3030228	-0.020	0.095534	3.385494	0.000711
PREVIV01	PROVW1S	0.4470561	-0.020	0.095534	4.893155	0.000001
RTAH00	DINVS_2	0.5356706	-0.020	0.064942	8.562686	0.000000
TELF00	DINVS_2	0.2980209	-0.020	0.064942	4.903277	0.000001
RDSI00	DINVS_2	0.2877365	-0.020	0.064942	4.744914	0.000002
SE2H	DINVS_2	0.173116	-0.020	0.064942	2.979950	0.002883
PD23G91	DINVS_2	0.2678185	-0.020	0.064942	4.438210	0.000009
DEM00	DINVS_2	0.5605685	-0.020	0.064942	8.946072	0.000000
PDSPE91	DINVS_2	0.4639305	-0.020	0.070939	6.827547	0.000000
DIST96	DINVS_2	0.2119736	-0.020	0.064942	3.578291	0.000346
PREVIV01	DINVS_2	0.270226	-0.020	0.064942	4.475282	0.000008

Fuente: Elaboración propia a partir de SpaceStat

A continuación, se presentan las variables seleccionadas, para las que se incluyen algunos gráficos del AEDE, como el diagrama de caja, diagrama y mapa de dispersión de Moran, mapa LISA y, para las variables explicativas de la RFD, además, el diagrama de dispersión de cada una de ellas respecto de la endógena.

En general, de la observación de los gráficos y mapas se deduce que, también en el ámbito provincial, las variables explicativas de la RFD, y la propia renta, presentan una fuerte autocorrelación espacial positiva, con una concentración clara de valores altos/bajos de las mismas en determinadas zonas geográficas. En general, la RFD y las explicativas que mantienen una relación directa con ella (que son todas ellas excepto la tasa de demandas de empleo y distancia media a las cabeceras de servicios) suelen experimentar mayores valores en la mitad norte-nordeste nacional y viceversa, aunque las variables explicativas más relacionadas con la actividad económica, como teléfonos (pese a la corrección efectuada), líneas RDSI y profesionales independientes, también destacan en las provincias del Arco Mediterráneo.

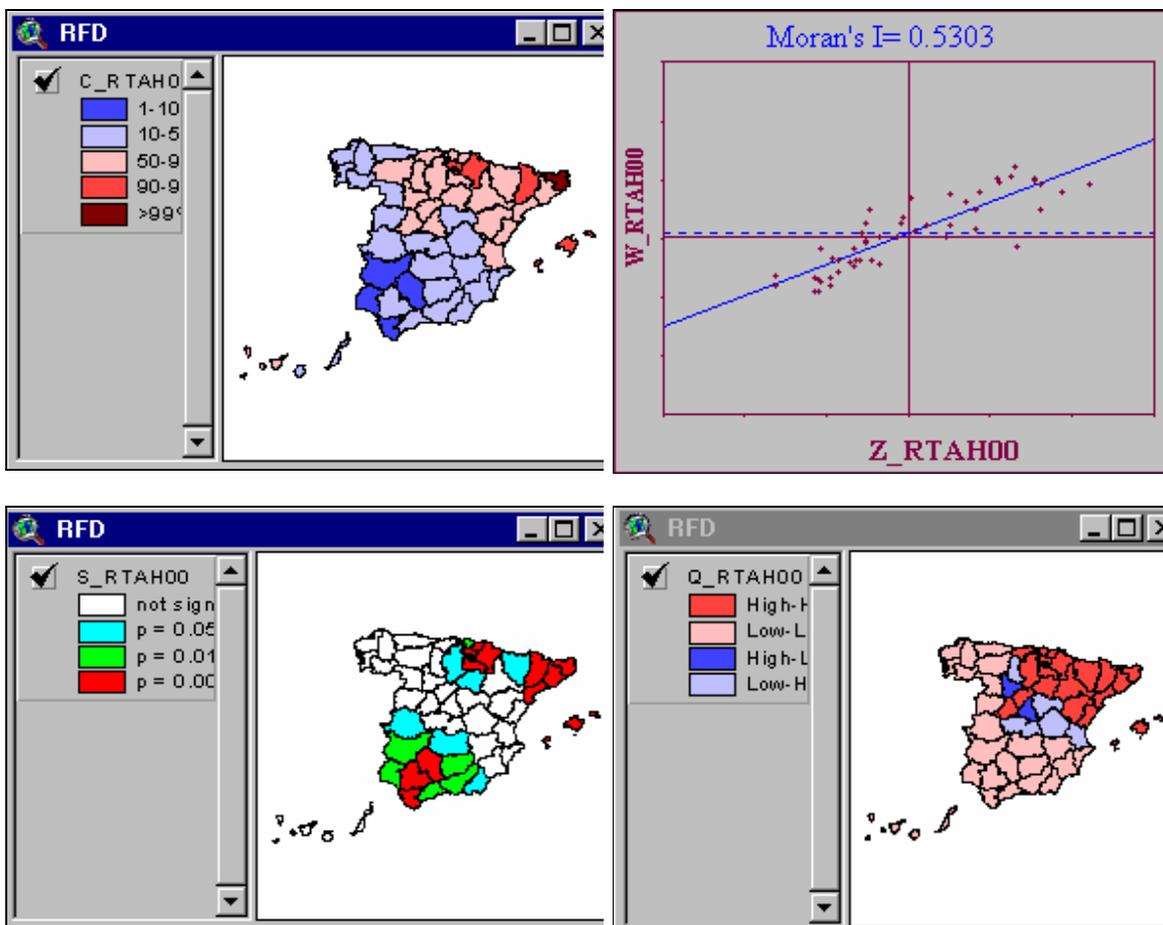
En concreto, los mapas LISA de autocorrelación espacial local suelen destacar, como provincias con máxima concentración de renta, las provincias vascas y catalanas, Navarra y Madrid.

Además, se presenta también el diagrama de dispersión de cada variable exógena respecto de la endógena (RFD), que puede ser indicativo de la presencia de heterogeneidad estructural en el modelo. Efectivamente, podrían diferenciarse dos estructuras o regímenes espaciales, entre las provincias con RFD superior a la media nacional, localizadas a modo de gran cuadrado en la mitad norte-este de España (desde Cantabria hasta Ávila, y desde Ávila hasta Illes Balears, pasando por Valencia/València, excepto Guadalajara) y el resto de provincias con inferior RFD. Como puede observarse, el coeficiente de correlación lineal de todas las variables explicativas respecto de la renta varía, a veces sensiblemente, de valor según que se obtenga en uno u otro subespacio, lo que podría estar indicando la existencia de algún tipo de discontinuidad geográfica. Es decir, podría resultar que la sensibilidad de los factores condicionantes de la RFD fuera significativamente distinto en cada una de las “dos Españas”.

Será ya propio del análisis confirmatorio dilucidar si la causa de estas estructuras o “clusters” son causadas por un efecto de dependencia o heterogeneidad espacial en el modelo explicativo de la RFD.

➤ **Renta disponible bruta de los hogares (RFD)**

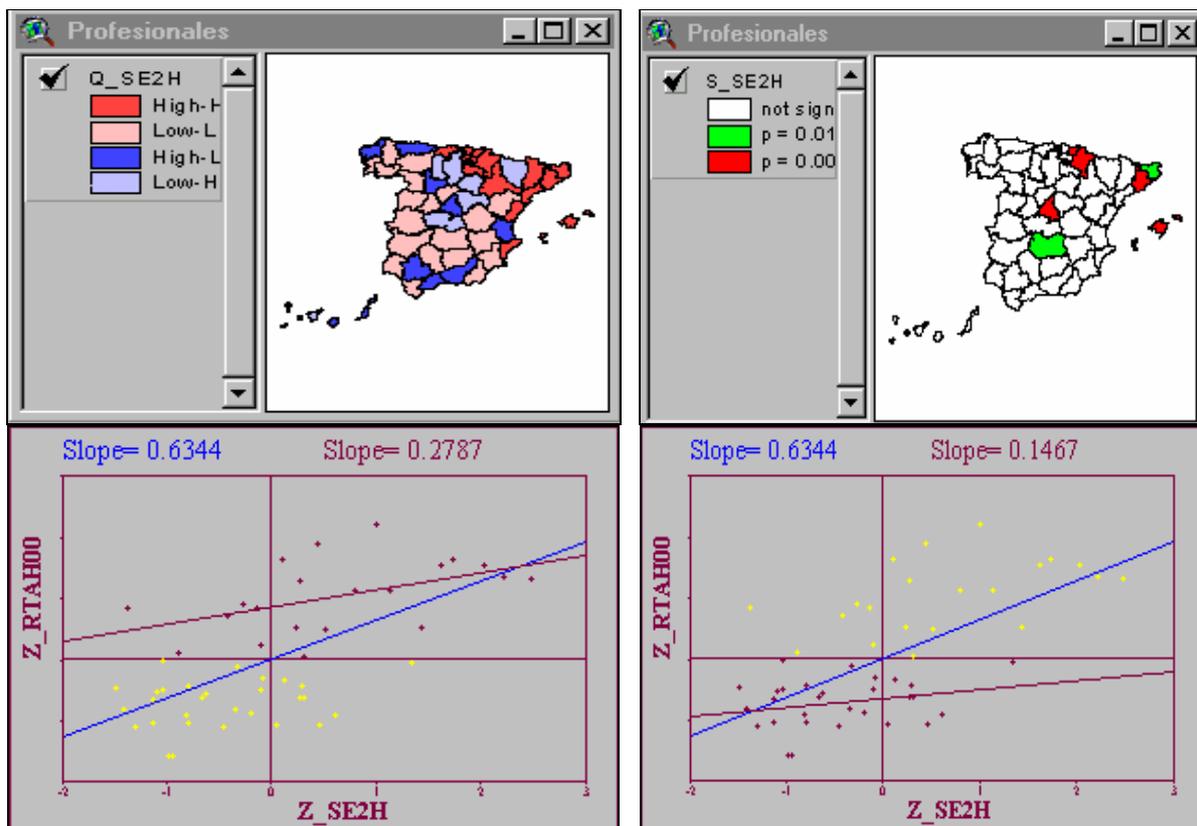
La RFD es la suma total de ingresos procedentes del trabajo, más las rentas de capital, prestaciones sociales y transferencias, menos los impuestos directos pagados por las familias y las cuotas pagadas a la seguridad social. Como la última actualización, para el conjunto de las provincias españolas, corresponde a 1999, se ha estimado la RFD provincial de 2000. Para ello, se ha seguido la metodología habitualmente empleada por el Instituto L. R. Klein (Vicéns y Chasco, 1997) en la que se realiza una predicción temporal de la RFD provincial de 2000, a partir de un modelo de datos de panel de efectos fijos, en el que las series de RFD, correspondientes a 50 provincias y 5 años (1995-1999), son explicadas mediante un grupo de variables.



Tanto del diagrama de caja (superior izquierda), como del mapa LISA (inferior izquierda) y del diagrama y mapa de dispersión de Moran (superior e inferior derecha, respectivamente) se deduce que, en el ámbito provincial, la RFD presenta autocorrelación espacial positiva con una concentración clara de valores altos en las provincias del norte/nordeste de España, frente a las provincias del sur/suroeste, en las que se agrupan los menores niveles de RFD.

➤ **Actividades profesionales independientes**

Los mapas de dispersión de Moran y LISA (superior izquierdo y derecho) ponen de manifiesto la existencia de autocorrelación espacial positiva en la distribución provincial de esta variable, produciéndose agrupamientos de valores altos en el entorno de las provincias de Illes Balears, Barcelona, Guipúzcoa, Madrid y Navarra, y de valores bajos de la variable alrededor de la provincia de Ciudad Real.



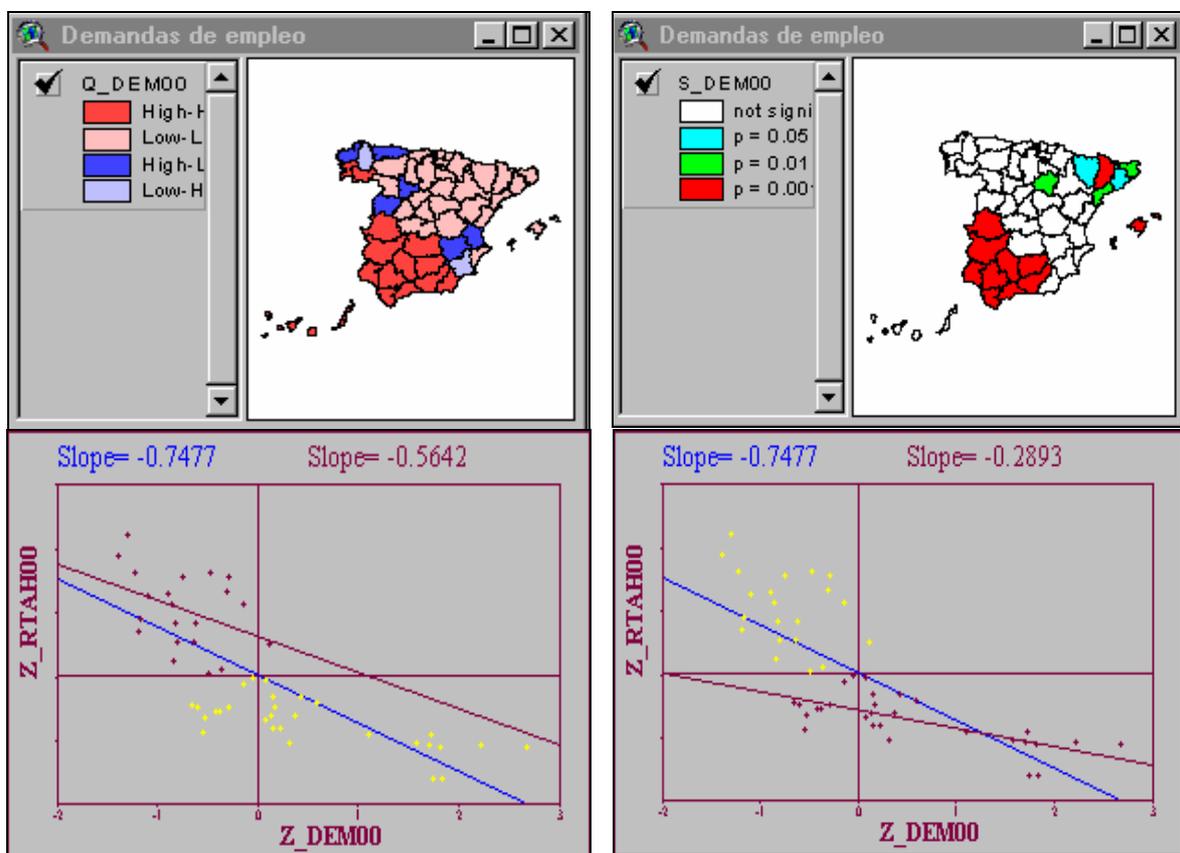
Además, se presenta también el diagrama de dispersión de esta variable exógena respecto de la endógena (RFD), diferenciando dos estructuras o regímenes espaciales, entre las provincias con RFD superior a la media nacional, localizadas en la mitad norte-este de España y el resto de provincias con inferior RFD. Como puede observarse, el coeficiente de correlación lineal de la variable actividades profesionales independientes respecto de la renta es mayor cuando se obtiene en el subespacio norte-este (0,2787) que en el resto de España (0,1467), tal como se aprecia, respectivamente, en los gráficos inferior izquierdo y derecho. Este resultado indica que la sensibilidad de esta variable, como explicativa de la RFD, es diferente en los dos subespacios considerados, y mayor en el régimen provincial norte-este.

Asimismo ambos coeficientes de correlación son sensiblemente inferiores al obtenido para el conjunto global de provincias (0,6344), de lo que se deduce que la

capacidad explicativa de esta variable en ambos subespacios es inferior a la obtenida en el espacio total de provincias.

➤ Demandas de empleo

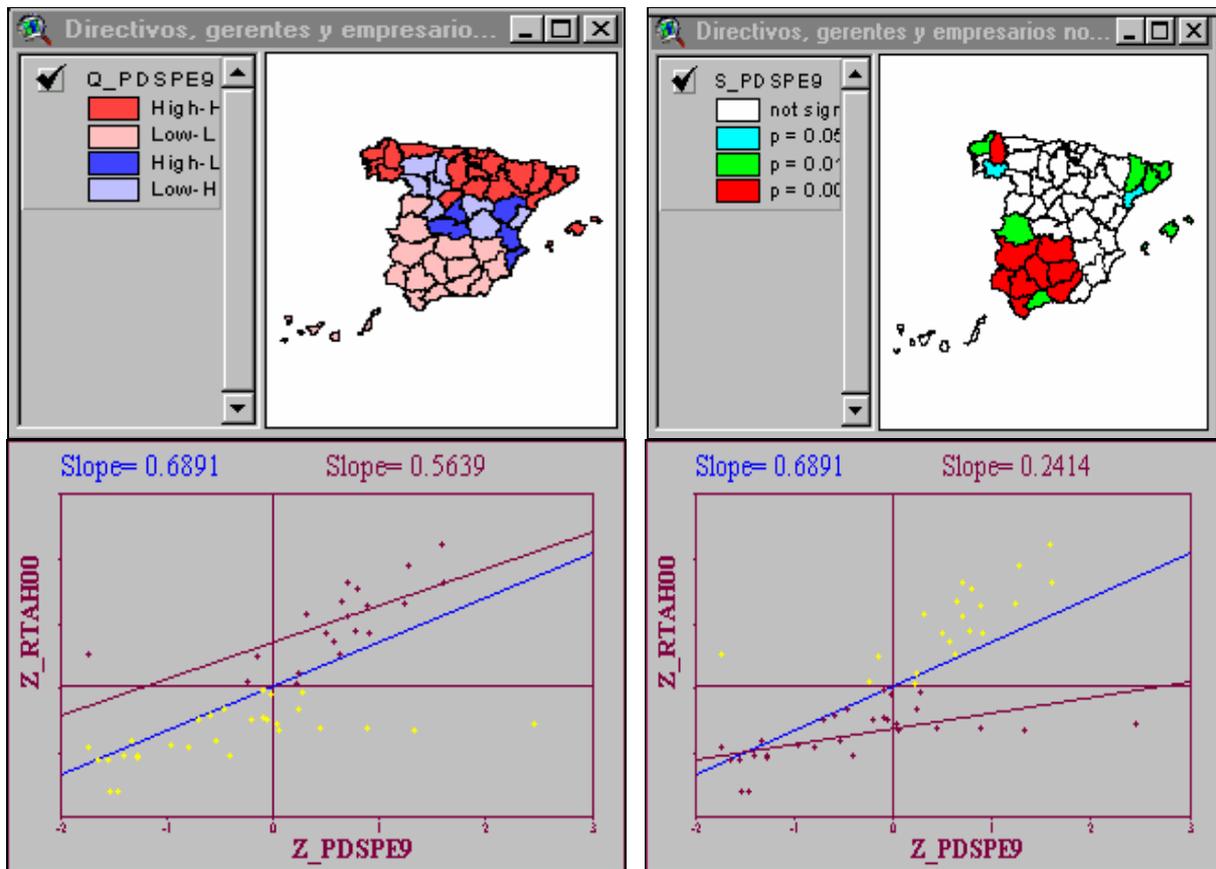
Los mapas de dispersión de Moran y LISA (superior izquierdo y derecho) ponen de manifiesto la existencia de autocorrelación espacial positiva en la distribución provincial de esta variable, produciéndose agrupamientos de valores altos de demandas de empleo (paro registrado) en el vértice suroeste peninsular de las provincias de Extremadura y Andalucía (Cádiz, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga y Sevilla) y de valores bajos de la variable alrededor de las provincias de Illes Balears y Lleida.



Además, el coeficiente de correlación lineal de esta variable de demandas de empleo respecto de la renta es también mayor (en valor absoluto) cuando se obtiene en la zona norte-este (-0,5642) que en el resto de España (-0,2893), lo que pone de manifiesto una mayor capacidad explicativa de la renta de esta variable en la “España rica” que en la pobre. En este caso, el grado de explicación de la variable en el subespacio norte-este es similar a la capacidad explicativa de la misma en el total del territorio nacional (con un coeficiente de correlación lineal de -0,7477).

➤ **Directivos, gerentes y empresarios no agrarios**

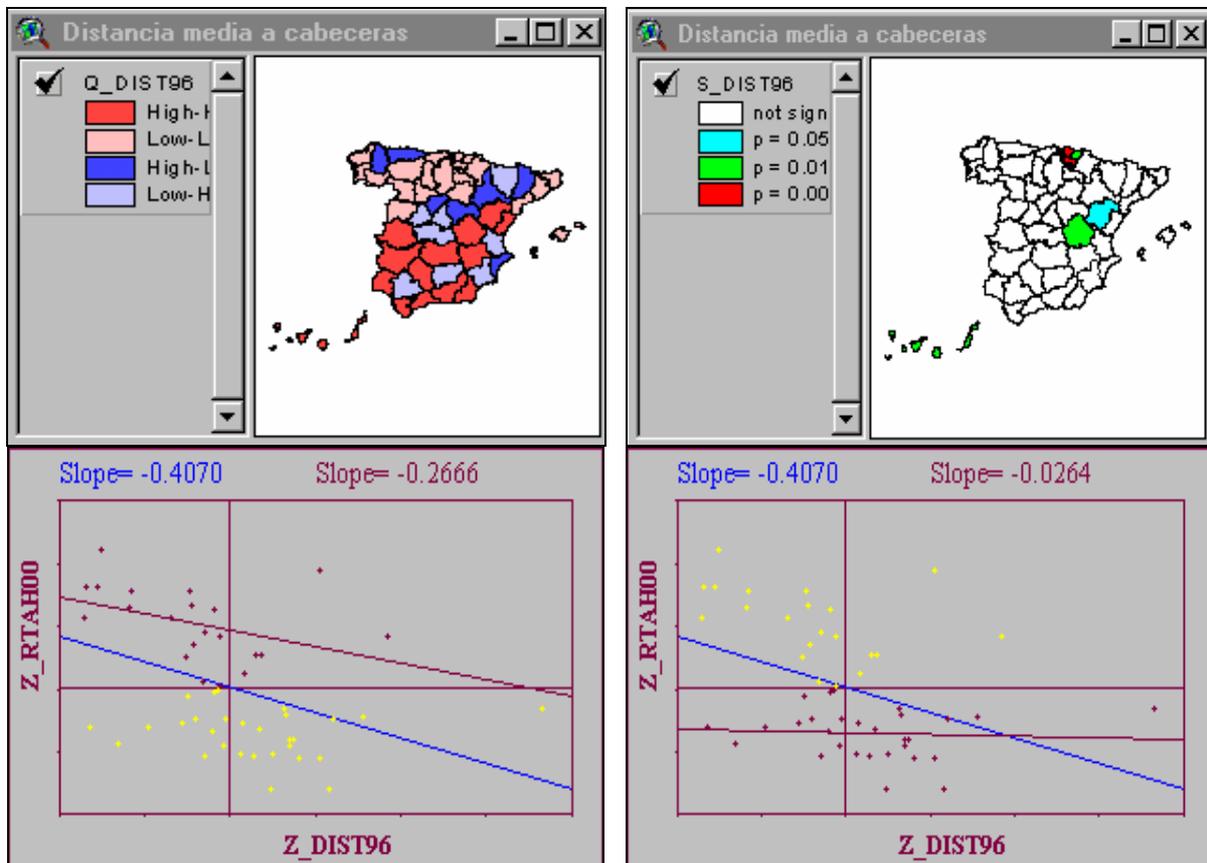
De los resultados de los mapas de dispersión de Moran y LISA se desprende la existencia de autocorrelación espacial positiva en la distribución provincial de esta variable, produciéndose agrupamientos de valores altos de directivos y empresarios no agrarios en los extremos nordeste (A Coruña, Lugo y Ourense) y noroeste (Illes Balears y Cataluña). Los valores bajos de la variable se localizan entorno a las provincias de suroeste peninsular (Badajoz, Cádiz, Ciudad Real, Córdoba, Granada, Huelva, Jaén y Sevilla).



Los diagramas de dispersión de esta variable respecto de la RFD en los dos espacios considerados vuelven a poner de manifiesto la mayor capacidad explicativa de la misma en la zona “rica”, con un coeficiente de correlación lineal de 0,5639 (cercano al coeficiente de correlación general, 0,6891).

➤ **Distancia media a municipios cabeceras de servicios**

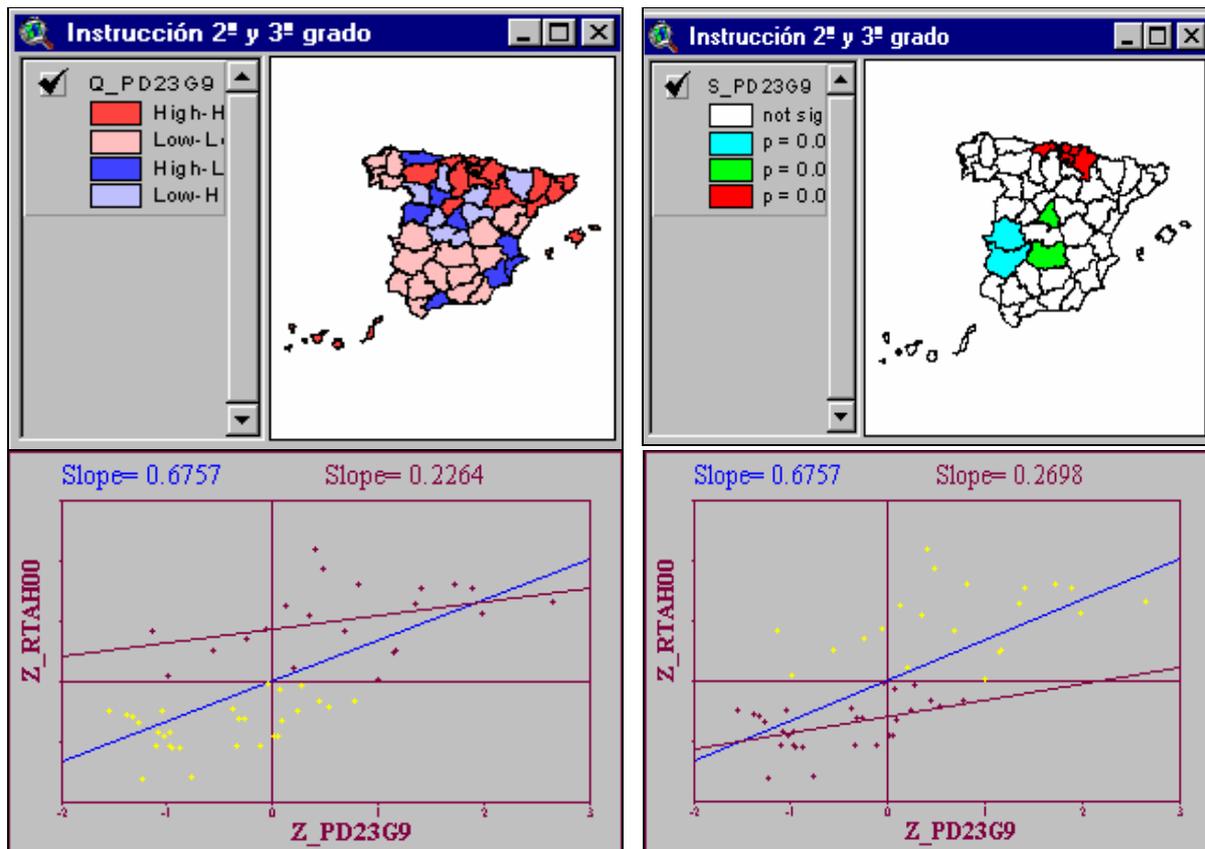
Para el ámbito provincial, esta variable se obtiene como la distancia media entre cada municipio de una provincia y el municipio cabecera de área comercial más cercano, ponderada por la población. Aunque presenta un menor grado de autocorrelación espacial, de los resultados expuestos en los mapas de dispersión de Moran y LISA se advierte una mayor presencia de provincias con menores distancias a los municipios cabeceras en la mitad norte de la Península, en el entorno del País Vasco como punto caliente. Y, por el contrario, en la mitad sur de España se sitúan las provincias con mayores distancias a los municipios cabeceras de servicios, sobre todo en la zona de las provincias interiores de Cuenca y Teruel.



La correlación lineal existente entre la variable de distancia media a las cabeceras y la RFD, en el espacio total de las provincias españolas, no es muy alta y negativa (-0,407). Esta relación es algo menor en valor absoluto en las provincias “ricas” (-0,2666), pero casi nulo en el subespacio del resto de provincias (-0,0264), lo que pone de manifiesto una menor sensibilidad de esta variable a la existencia de las dos estructuras o regímenes espaciales considerados.

➤ **Instrucción de 2º y 3º grado**

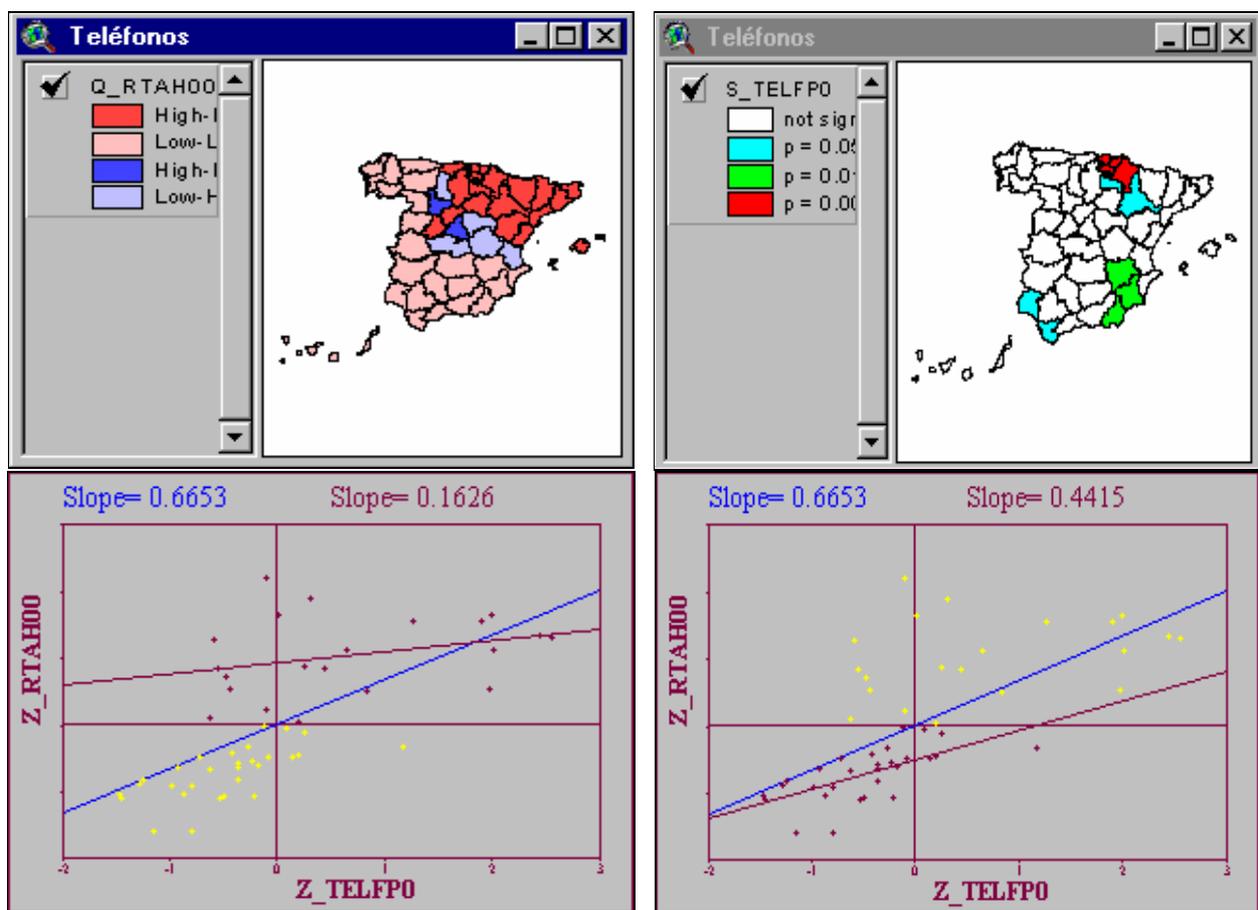
La variable del nivel de instrucción de grado superior presenta una distribución provincial muy parecida a la observada en el resto de variables, con una concentración de valores altos de esta variable en las provincias del País Vasco, Cantabria, Navarra y Madrid, estando localizadas las menores tasas de instrucción superior en las provincias cercanas a Badajoz, Cáceres y Ciudad Real.



En este caso, la correlación lineal o grado de explicación existente entre la variable de instrucción de 2º y 3º grado y la RFD, en los dos subespacios considerados es muy similar e incluso algo mayor en la zona “pobre” (0,2698) que en la “rica” (0,2264) y, en todo caso, bastante menor que la existente para el conjunto global de España (0,6757).

➤ **Líneas telefónicas de uso doméstico en primera vivienda**

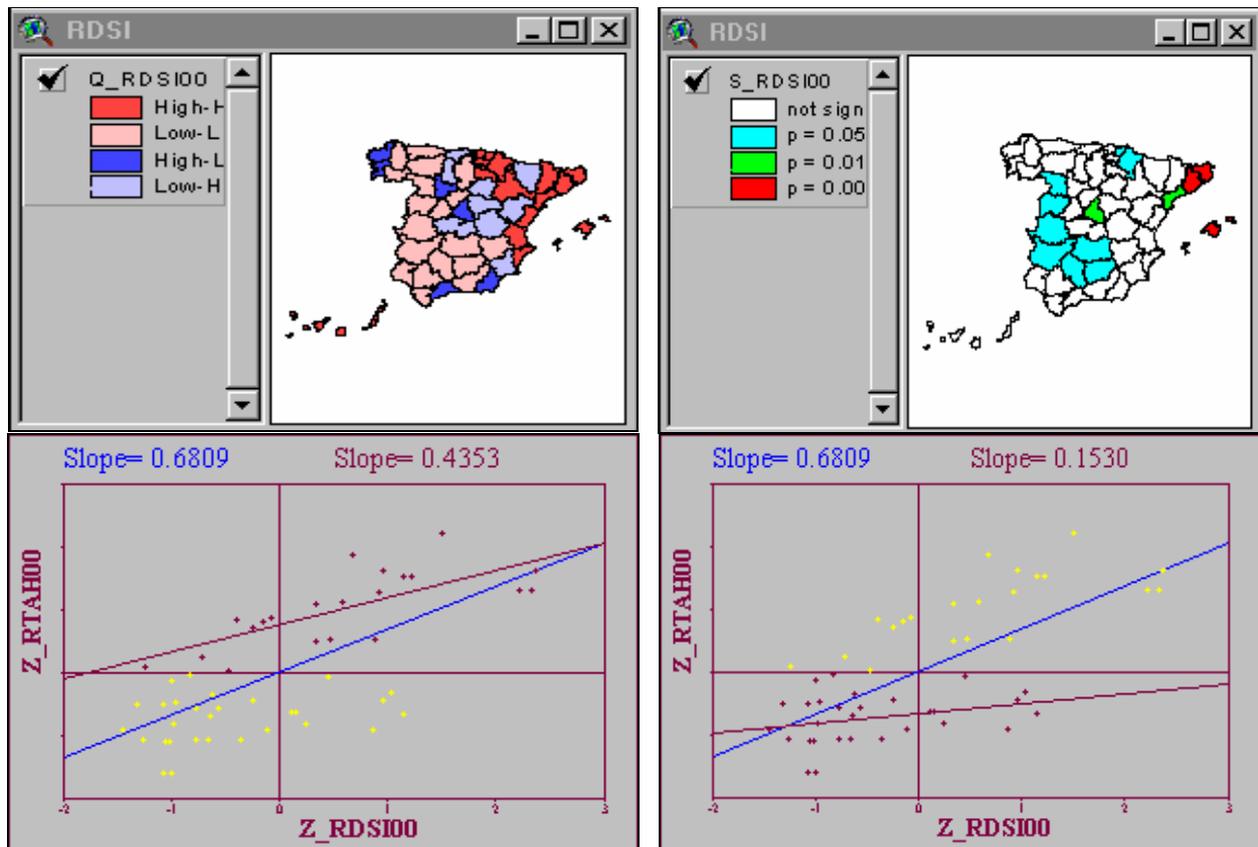
Una vez más, esta variable se distribuye en forma autocorrelacionada en el espacio de las provincias españolas, registrando los mayores valores en el cuadrante norte-este y los menores, en las provincias del sur de España. El mapa LISA indica que la zona caliente de la variable líneas telefónicas de uso doméstico en primera vivienda se localiza en las provincias del Valle del Ebro (País Vasco, Navarra, La Rioja y Zaragoza). Sin embargo, la zona fría de bajas tasas de teléfonos domésticos se encuentra no sólo en la zona suroeste peninsular (Cádiz y Huelva), sino también en el sureste (Albacete, Murcia y Almería).



Como en caso anterior, aunque la sensibilidad de esta variable a la RFD es elevada (0,6653), el coeficiente de correlación lineal es significativamente mayor en el subespacio “pobre” (0,4415) que en el “rico” (0,1626), en el que el indicador de líneas telefónicas de uso doméstico en primera vivienda tiene menor capacidad explicativa.

➤ Líneas RDSI

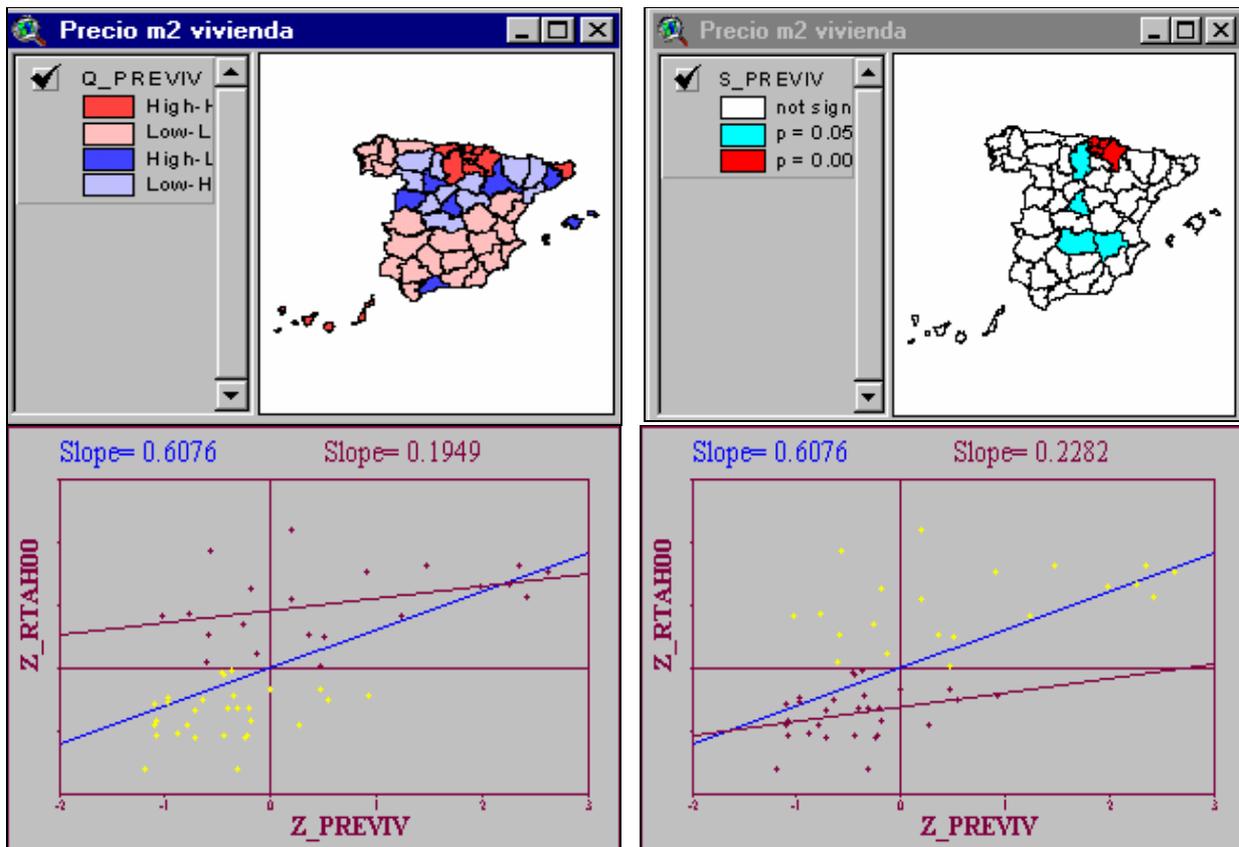
La variable de líneas RDSI presenta también autocorrelación espacial positiva, con un claro desplazamiento de los valores altos de la misma desde Madrid y las provincias del norte (País Vasco y Navarra) hacia el Arco Mediterráneo (Illes Balears y Cataluña). Por su parte, las más bajas tasas de líneas RDSI se localizan en las provincias interiores de la mitad sur peninsular (Ciudad Real, Córdoba y Jaén), así como en las provincias limítrofes con Portugal (Badajoz, Cáceres, Salamanca y Zamora).



En este caso, como en la variable actividades profesiones independientes, demandas de empleo, directivos y empresarios no agrarios y distancia media a municipios cabeceras de servicios, la sensibilidad de la variable líneas RDSI a la RFD es mayor en el subespacio norte-este (0,4353) que en el sur-oeste (0,1530).

➤ **Precio del metro cuadrado de la vivienda**

El precio del metro cuadrado de la vivienda es una variable que también depende de las coordenadas geográficas, siendo, en términos generales (una vez más) las provincias del cuadrante norte-este de España las que registran un mayor precio medio de la vivienda (con algunas excepciones, como Málaga y las provincias canarias). Por su parte, la Meseta sur (Albacete y Ciudad Real) es la zona en las que se concentran las provincias con un menor precio de la vivienda.



Al igual que las variables instrucción de 2º y 3º ciclo y teléfonos de uso doméstico en primera vivienda, el grado de explicación de la RFD que presenta esta variable es mayor en el subespacio de la “España pobre” (0,2282) que en el de la “rica” (0,1949), siendo en ambos casos muy inferior al coeficiente de correlación lineal correspondiente al total de provincias (0,6076).

V.3.5. Análisis confirmatorio espacial

La aplicación del AEDE, tanto en el ámbito municipal como provincial, pone de manifiesto la existencia de un elevado grado de dependencia espacial en todas las variables, incluida la RFD provincial; es decir, ninguna de ellas tiene precisamente una distribución aleatoria sobre el espacio geográfico español. Además, sería también probable que esta dependencia fuera en realidad (en lugar de o además) un efecto de heterogeneidad espacial, en forma de dos regímenes o estructuras de provincias, que podrían ser denominadas, de forma muy simplista, la España rica y la España pobre.

Estos procesos pudieran estar motivados por efectos de autocorrelación espacial en los factores explicativos de la RFD (presentes o ausentes del modelo), así como en la propia distribución de la RFD, provincial y municipal, que, como se indicaba en el Apartado de fundamentos teóricos, suele estar fuertemente condicionada por la localización geográfica y factores hereditarios. De ser así, la especificación del modelo debería incluir, de forma correcta, la presencia de los efectos espaciales, y los estimadores MCO no tendrían ya las buenas propiedades que se les exige.

En cualquier caso, previo a cualquier ejercicio de modelización, debe realizarse un análisis factorial entre las variables explicativas, que debido al grado de interrelación existente entre las mismas, causarían problemas añadidos de multicolinealidad en el modelo. La extracción de los factores se ha llevado a cabo a través del método de componentes principales que, mediante combinaciones lineales de los indicadores iniciales, logra que estén incorrelacionados entre sí. Se ha seguido el criterio de selección de los factores (2 factores, en este caso) con un valor propio superior a 1. En la Tabla V.3.7., se presenta la composición de dichos factores, diferenciándose en cada uno de ellos entre variables de factor puro y variables complejas, según el valor de la carga factorial.

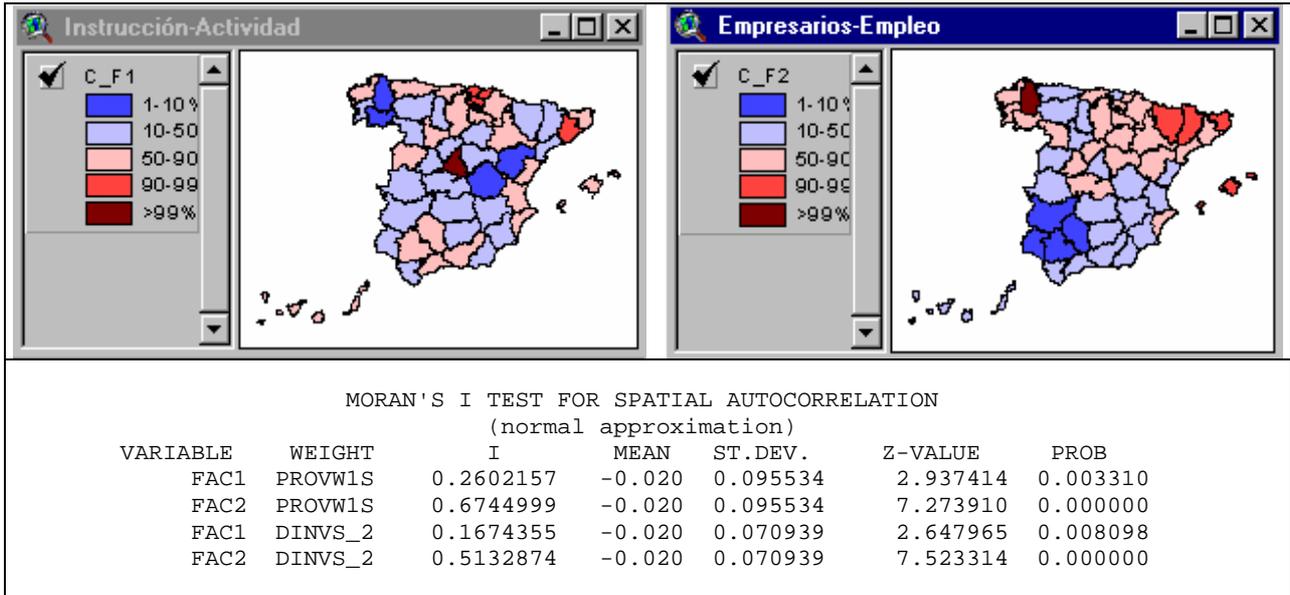
Tabla V.3.7. Composición de los factores rotados

Factor	Tipo de variable		Nombre Factor
	Factor puro (>0,7)	Variables complejas (>0,3)	
1	Instrucción (0,94) Precio vivienda (0,91) Profesionales (0,89) Teléfonos (0,81) Líneas RDSI (0,75)	Distancia cabeceras (-0,40)	INSTRUCCIÓN- ACTIVIDAD
2	Demandas empleo (-0,91) Directivos (0,89)	Distancia cabeceras (-0,38) Líneas RDSI (0,36)	EMPRESARIOS- EMPLEO

Fuente: Elaboración propia a partir de SPSS.

La varianza total explicada por los factores alcanza un porcentaje acumulado del 76%. Asimismo, se ha optado por el método de rotación Varimax, que minimiza el número de factores con carga alta en cada factor.

Figura V.3.3. Distribución geográfica y contraste de autocorrelación espacial de los factores Instrucción-Actividad (F1) y Empresarios-Empleo (F2)



Fuente: Elaboración propia a partir de SpaceStat y ArcView.

Los resultados obtenidos por el test I de Moran, para las dos matrices de pesos espaciales definidas, en cada los dos factores explicativos ponen de manifiesto la existencia de autocorrelación espacial positiva en todas las variables (Figura V.3.3.).

A partir de aquí, se aplica la estrategia de modelización clásica presentada en el Apartado IV.2. que parte de la especificación del **MBRL estimado por MCO**, sobre el que se aplican todo tipo de contrastes (normalidad, multicolinealidad, heteroscedasticidad, autocorrelación espacial). En concreto, se plantea el siguiente modelo inicial:

$$RFD_i = \beta_0 + \beta_1 F1_i + \beta_2 F2_i + u_i \quad \text{Eq. V.3.2.}$$

siendo RFD_i: renta familiar disponible o renta disponible bruta de los hogares, de la provincia i

- F1: factor de Instrucción-Actividad
- F2: factor de Empresarios-Empleo
- β₀, β₁, β₂: parámetros a estimar
- u: perturbación aleatoria esférica

Para la realización de los contrastes de autocorrelación espacial se han utilizado las citadas matrices de interacción espacial (PROVW1S) y pesos espaciales (inversa de la distancia cuadrática o DINV_2S), ambas estandarizadas por filas. Los resultados obtenidos de la estimación, presentados en la Tabla V.3.8., ponen de manifiesto una significatividad individual de los parámetros y conjunta del modelo, obtenida a través del test t de Student (t-value) y coeficiente de determinación (R2). El valor del término constante indica el valor medio de RFD per capita de España, en miles de pesetas (1.562,96), y la mayor cuantía del coeficiente estimado del factor 2 (Empresarios-Empleo), dada la estandarización de dichos factores, otorga un mayor peso en la RFD por habitante provincial a esta variable que a la anterior (Instrucción-Actividad).

Tabla V.3.8. Estimación del MBRL por MCO

DEPENDENT VARIABLE	RTAH00	OBS	50	VARS	3	DF	47		
R2	0.7581	R2-adj	0.7478						
LIK	-311.441	AIC	628.882	SC	634.618				
RSS	752971.	F-test	73.6598	Prob	3.26709e-15				
SIG-SQ	16020.7	(126.573)	SIG-SQ(ML)	15059.4	(122.717)
VARIABLE	COEFF	S.D.	t-value	Prob					
CONSTANT	1562.96	17.9001	87.315520	0.000000					
F1	142.511	18.0818	7.881478	0.000000					
F2	166.904	18.0818	9.230486	0.000000					
DIAGNOSTICS FOR HETEROSKEDASTICITY									
RANDOM COEFFICIENTS									
TEST	DF	VALUE	PROB						
Breusch-Pagan test	2	1.151653	0.562240						
SPECIFICATION ROBUST TEST									
TEST	DF	VALUE	PROB						
White	5	17.787006	0.003226						
DIAGNOSTICS FOR SPATIAL DEPENDENCE									
FOR WEIGHTS MATRIX DINVS_2 (row-standardized weights)									
TEST	MI/DF	VALUE	PROB						
Moran's I (error)	0.201750	3.820859	0.000133						
Lagrange Multiplier (error)	1	8.050387	0.004549						
Robust LM (error)	1	0.030278	0.861860						
Kelejian-Robinson (error)	3	1.738482	0.628413						
Lagrange Multiplier (lag)	1	17.412263	0.000030						
Robust LM (lag)	1	9.392154	0.002179						
Lagrange Multiplier (SARMA)	2	17.442542	0.000163						
FOR WEIGHTS MATRIX PROVW1S (row-standardized weights)									
TEST	MI/DF	VALUE	PROB						
Moran's I (error)	0.339774	4.124065	0.000037						
Lagrange Multiplier (error)	1	11.570512	0.000670						
Robust LM (error)	1	0.032656	0.856595						
Kelejian-Robinson (error)	3	8.361855	0.039096						
Lagrange Multiplier (lag)	1	24.070493	0.000001						
Robust LM (lag)	1	12.532637	0.000400						
Lagrange Multiplier (SARMA)	2	24.103149	0.000006						

Fuente: Elaboración propia a partir de SpaceStat.

Pero estos resultados quedan invalidados por los resultados de los contrastes de autocorrelación espacial, tanto en el caso de la matriz de distancias como de interacciones binarias. En concreto, parece ser que el tipo de autocorrelación espacial presente en el modelo no es residual (“error”), sino sustantiva (“lag”), debido al esquema de dependencia detectado en la variable de RFD por habitante gracias a las técnicas del AEDE. El test LM sobre el retardo espacial de la variable endógena, robusto a la presencia de dependencia residual, o test LM-LE, “Robust LM (lag)”, así parece confirmarlo.

Además, el test robusto de White de heteroscedasticidad rechaza la hipótesis nula de no existencia de este efecto en la perturbación aleatoria, por lo que habrá también que contrastar la existencia de regímenes o estructuras espaciales de heterogeneidad espacial, tal como se puso de manifiesto también en el AEDE.

Por tanto, el paso siguiente consistiría en la re-especificación del modelo anterior como un **modelo del retardo espacial o modelo mixto autorregresivo de regresión espacial**:

$$RFD_i = \rho.W.RFD_i + \beta_0 + \beta_1 F1_i + \beta_2 F2_i + u_i \quad Eq. V.3.3.$$

siendo ρ : parámetro autorregresivo correspondiente a la variable RFD espacialmente retardada

W: matriz de pesos espaciales especificada como la inversa de la distancia cuadrática, que expresa con mayor precisión las relaciones interprovinciales que la matriz binaria de interacciones o contigüidades.

Sin embargo, la realización de los contrastes de dependencia espacial se han llevado a cabo con las dos matrices de pesos espaciales (interacciones y distancias).

Dado que la estimación MCO ya no produce coeficientes con buenas propiedades, se ha utilizado el método MV. En la Tabla V.3.9., puede apreciarse el buen nivel de significatividad individual de los parámetros, incluido el término autorregresivo espacial ρ , correspondiente a la variable de RFD per capita desplazada (W_RTAH00). En este caso, el coeficiente de determinación (R2) ya no resulta un instrumento adecuado para comparar este resultado con el obtenido en la estimación MCO, por lo que se utilizará el valor del logaritmo de verosimilitud (LIK); se observa una mayor cuantía del LIK en el modelo MV (-303,307) que MCO (-311,441), que pone de manifiesto su superioridad.

Además, se ha realizado también la estimación del modelo del error espacial, constatándose también su inferioridad respecto de la especificación del modelo anterior, del retardo espacial⁷⁵.

⁷⁵ Efectivamente, el logaritmo de verosimilitud (LIK) no supera al obtenido por el modelo del retardo espacial y, además, el coeficiente autorregresivo espacial de primer orden correspondiente a la perturbación aleatoria (λ) no resulta significativo.

Tabla V.3.9. Estimación del modelo del retardo espacial por MV

DATA SET	MERL		SPATIAL WEIGHTS MATRIX	DINVS_2	
DEPENDENT VARIABLE	RTAH00		OBS 50	DF 46	
R2	0.8107	Sq. Corr.	0.8365		
LIK	-303.307	AIC	614.614	SC 622.262	
SIG-SQ	10190.6	(100.949)	
VARIABLE	COEFF	S.D.	z-value	Prob	
W_RTAH00	0.646456	0.132783	4.868506	0.000001	
CONSTANT	544.36	208.409	2.611981	0.009002	
F1	99.4995	16.1267	6.169871	0.000000	
F2	96.7089	21.2041	4.560850	0.000005	
REGRESSION DIAGNOSTICS					
DIAGNOSTICS FOR HETEROSKEDASTICITY					
RANDOM COEFFICIENTS					
TEST	DF	VALUE	PROB		
Breusch-Pagan test	2	0.421621	0.809928		
Spatial B-P test	2	0.421621	0.809928		
DIAGNOSTICS FOR SPATIAL DEPENDENCE					
SPATIAL LAG DEPENDENCE FOR WEIGHTS MATRIX DINVS_2 (row-standardized weights)					
TEST	DF	VALUE	PROB		
Likelihood Ratio Test	1	16.267861	0.000055		
LAGRANGE MULTIPLIER TEST ON SPATIAL ERROR DEPENDENCE					
WEIGHT	STAND	ZERO	DF	VALUE	PROB
DINVS_2	yes	no	1	0.130892	0.717509
PROVW1S	yes	no	1	0.751120	0.386122

Fuente: Elaboración propia a partir de SpaceStat.

Aunque los contrastes de heteroscedasticidad (Breush-Pagan y B-P espacial) parecen aceptar, en este caso, la hipótesis nula de ausencia de este efecto en el modelo, se ha llevado a cabo una especificación más, de heterogeneidad espacial, que contraste la existencia de las dos estructuras espaciales (España rica “versus” España pobre) anteriormente detectadas en el AEDE. Se trata de un **modelo de regresiones cambiantes (“switching regressions”) del modelo del retardo espacial** anterior, en el que se estiman coeficientes diferentes para cada estructura o régimen espacial considerado (P, R), a través de una variable dicotómica (NS), que adopta el valor “0” para la zona P (España pobre) y “1” para la zona R (España rica). En concreto, se trata de la siguiente especificación:

$$RFD_i = \rho.W.RFD_i + \beta_{0P} + \beta_{1P} F1_i + \beta_{2P} F2_i + \beta_{0R} + \beta_{1R} F1_i + \beta_{2R} F2_i + u_i \quad Eq. V.3.4.$$

siendo P: subíndice correspondiente a la estructura de la España pobre

R: subíndice correspondiente a la estructura de la España rica

La estimación MV da lugar a los resultados presentados en la Tabla V.3.10., en los que se advierte, una vez más, la significatividad individual de los coeficientes estimados,

así como una superioridad del contraste conjunto (LIK = -283,901) respecto de los modelos anteriores.

Tabla V.3.10. Estimación de un modelo del retardo espacial de regresiones cambiantes, por MV

SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION					
STRUCTURAL CHANGE - DIFFERENT SLOPES FOLLOWING VARIABLE NS					
DATA SET	MERL		SPATIAL WEIGHTS MATRIX	DINVS_2	
DEPENDENT VARIABLE	RTAH00	OBS	50	VARS	7 DF 43
R2	0.9190	Sq. Corr.	0.9202		
LIK	-283.901	AIC	581.801	SC	595.185
SIG-SQ	4966.82	(70.4757)	
VARIABLE	COEFF	S.D.	z-value	Prob	
W_RTAH00	0.240575	0.129554	1.856946	0.063319	
CONST_0	1062.75	193.202	5.500709	0.000000	
F1_0	63.549	25.2134	2.520450	0.011720	
F2_0	54.9765	16.5618	3.319474	0.000902	
CONST_1	1226.6	211.105	5.810380	0.000000	
F1_1	92.7165	14.9246	6.212325	0.000000	
F2_1	193.421	29.8148	6.487433	0.000000	
TEST ON STRUCTURAL INSTABILITY FOR 2 REGIMES DEFINED BY NS					
TEST	DF	VALUE	PROB		
Chow - Wald	3	54.571111	0.000000		
STABILITY OF INDIVIDUAL COEFFICIENTS					
TEST	DF	VALUE	PROB		
CONST_0	1	21.020717	0.000005		
F1_0	1	0.981544	0.321818		
F2_0	1	19.851663	0.000008		
DIAGNOSTICS FOR HETEROSKEDASTICITY					
LINEAR SPECIFICATION USING VARIABLES					
CONSTANT	NS_1				
TEST	DF	VALUE	PROB		
Breusch-Pagan test	1	7.304992	0.006876		
Spatial B-P test	1	7.304994	0.006876		
DIAGNOSTICS FOR SPATIAL DEPENDENCE					
SPATIAL LAG DEPENDENCE FOR WEIGHTS MATRIX DINVS_2 (row-standardized weights)					
TEST	DF	VALUE	PROB		
Likelihood Ratio Test	1	3.203576	0.073477		
LAGRANGE MULTIPLIER TEST ON SPATIAL ERROR DEPENDENCE					
WEIGHT	STAND	ZERO	DF	VALUE	PROB
DINVS_2	yes	no	1	3.025319	0.081974
PROVWLS	yes	no	1	0.181790	0.669840

Fuente: Elaboración propia a partir de SpaceStat.

Por otro lado, el test de Chow espacial de Anselin, de heterogeneidad espacial conjunta en el modelo, rechaza la hipótesis nula de ausencia de deriva paramétrica discreta en los parámetros, lo que viene también a confirmar la bondad de esta especificación (aunque el contraste de estabilidad individual sobre el parámetro del factor de Instrucción-Actividad acepta la hipótesis nula de constancia). Por último, los contrastes de dependencia

espacial sustancial (“spatial lag”) y heteroscedasticidad ponen de manifiesto la necesidad de especificar ambos efectos en el modelo, tal como se ha hecho. Además, el contraste LR confirma también la ausencia de dependencia espacial residual (“spatial error”), por lo que éste será el modelo ecológico seleccionado para realizar la predicción-extrapolación de la RFD per capita de los municipios de la CM.

Tabla V.3.11. Residuos del modelo estimado en la Tabla V.3.10. (datos en €)

OBS	RTAH00	PREDIC	RESID
1	11.835	11.284	551
2	7.781	8.021	-240
3	8.391	8.610	-219
4	8.123	8.179	-56
5	9.531	9.617	-86
6	6.948	7.478	-531
7	11.835	12.195	-360
8	11.338	12.168	-829
9	10.624	10.413	211
10	7.775	7.675	100
11	6.953	7.717	-764
12	10.160	9.452	708
13	8.117	7.817	301
14	7.696	7.669	27
15	8.423	8.757	-334
16	8.636	7.932	705
17	12.710	12.249	461
18	7.739	8.130	-392
19	8.690	8.345	344
20	11.697	11.543	154
21	7.692	7.529	163
22	10.706	10.974	-269
23	7.980	7.770	209
24	9.196	9.780	-584
25	12.210	11.343	868
26	11.049	11.008	40
27	8.507	8.651	-145
28	11.382	11.053	328
29	7.987	8.271	-285
30	8.015	8.479	-464
31	11.698	10.729	969
32	8.347	8.426	-79
33	8.861	8.518	342
34	9.320	9.368	-48
35	8.719	8.475	244
36	8.419	8.576	-157
37	8.625	8.499	126
38	8.874	8.148	725
39	9.437	10.400	-963
40	9.725	10.195	-469
41	7.752	7.960	-208
42	10.446	10.765	-319
43	11.288	11.288	0
44	10.643	9.989	655
45	8.535	8.335	200
46	9.280	10.198	-918
47	10.110	9.767	344
48	11.074	10.971	102
49	8.616	8.230	386
50	10.173	10.718	-546

Residuos RTAH00

- 2 a 1 Sigma
- 1 a 0 Sigma
- Media
- 0 a -1 Sigma
- -1 a -2 Sigma

Fuente: Elaboración propia a partir de SpaceStat y MapInfo.

Los principales resultados, el valor real (RTAH00), estimado (PREDICTED) y el residuo de la regresión (RESIDUAL), se ofrecen en la Tabla V.3.11, así como el mapa de los residuos, en el que se destacan las provincias con un error superior a los límites de confianza $\pm 2\sigma$. Los mayores errores, atípicamente altos en valor absoluto, se producen en las provincias de Cantabria (-963 €) y Navarra (969 €). En el primer caso, el modelo estima por exceso el valor de la RFD, mientras que en el segundo, lo hace por defecto. Es decir, la predicción-extrapolación de los datos de RFD municipal, a partir del modelo seleccionado, resultará sobrevalorada en los municipios de Cantabria y subvalorada en los municipios de Navarra, lo que obligará a una posterior corrección de los mismos para que se ajusten a las cifras oficiales proporcionadas por el INE para las provincias españolas en la Contabilidad Regional de España.

Además, también pueden destacarse otros errores por exceso de menor tamaño, en los intervalos $(\sigma, 2\sigma)$ o $(-\sigma, -2\sigma)$ cometidos en Badajoz (-531 €), Barcelona (-829 €), Cádiz (-764 €), León (-584 €), Murcia (-464 €), Segovia (-464 €), Valencia/València (-918 €) y Zaragoza (-546 €). En cuanto a los errores por defecto, además del ya mencionado en la provincia de Navarra, éstos se han producido en Álava (551 €), Castellón/Castelló (708 €), Cuenca (705 €), Girona (461 €), Huesca (868 €) y Santa Cruz de Tenerife (725 €). Por su parte, el tamaño del residuo correspondiente a la provincia de Madrid (328 €) es bajo y se considera situado en la media que, por definición, es cero.

V.3.6. Predicción-extrapolación de datos microterritoriales

Por último, el ejercicio de predicción-extrapolación consiste en la aplicación, lo más “directa” posible, del modelo seleccionado en el análisis confirmatorio. Por ejemplo, en este caso, las dos estructuras espaciales detectadas en la distribución provincial de la RFD per capita (España rica/España pobre) podrían corresponderse con las dos estructuras que parecen también producirse en la CM, del tipo Centro-Periferia. Por otro lado, debe también re-especificarse el modelo para evitar la presencia de la variable endógena RFD (objetivo a estimar) espacialmente retardada como variable explicativa del modelo, mediante el filtrado espacial de las variables exógenas y término independiente, del modo siguiente:

$$\hat{y} = [I - \rho W]^{-1} X \hat{\beta}$$

$$RFD_i = [I - \rho W]^{-1} \left[\hat{\beta}_{0P} + \hat{\beta}_{1P} F1_i + \hat{\beta}_{2P} F2_i + \hat{\beta}_{0R} + \hat{\beta}_{1R} F1_i + \hat{\beta}_{2R} F2_i \right] \quad Eq. V.3.5.$$

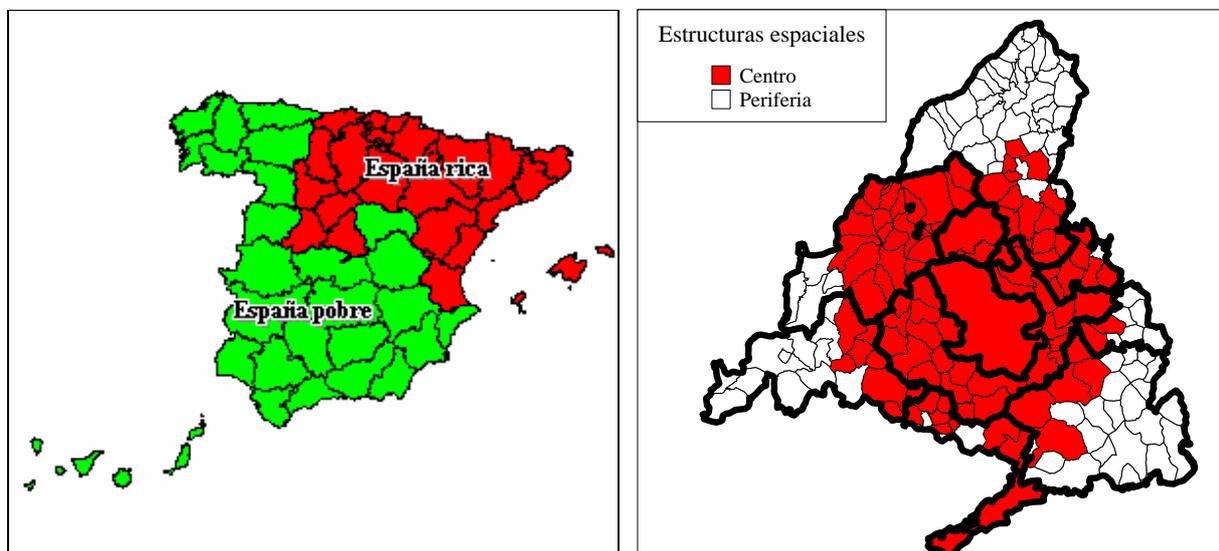
siendo $[I - \rho W]^{-1}$ un multiplicador espacial global, según la terminología de la segunda taxonomía de Anselin (ver Apartado IV.1.). Efectivamente, de esta expresión se desprende que el valor de la RFD en el municipio “i” depende no sólo de los valores de los factores explicativos en “i”, sino también del valor que adoptan dichas variables en el resto de municipios de la Comunidad de Madrid (dimensión global), aunque obviamente, debido a la composición de la matriz W (cuadrado de la distancia inversa), el influjo que ejercen los municipios cercanos a “i” sobre dicha localidad será mucho mayor que el realizado por los

municipios lejanos, que se considera prácticamente nulo. Este efecto tiene pleno sentido en cuanto que las variables que forman parte de los factores explicativos (nivel de instrucción superior, tasa de paro registrado, tasa de directivos y empresarios no agrarios, precio de la vivienda, etc.) presentan autocorrelación espacial positiva, es decir que si, por ejemplo, en un municipio se localiza un importante volumen de población con alto nivel de estudios será también probable encontrar altas tasas de población instruida en los municipios cercanos al mismo, y viceversa.

Por otro lado, como ya se indicado, se propone también una identificación entre la estructura que hemos denominado “España rica” (R), formada por las provincias del noreste de España, afectadas positivamente por su cercanía a los grandes centros de producción (Madrid, Barcelona, Valencia/València, Bilbao, Zaragoza) y los municipios ricos situados en el “Centro” de la CM, así como también podría existir un paralelismo, en la distribución de la RFD entre las provincias de la “España pobre” (P) y los municipios localizados en la “Periferia” de la CM.

Para ello, se ha realizado un análisis exploratorio de la distribución municipal de los factores explicativos de la RFD en la CM (Instrucción-Actividad y Empresarios-Emple), identificando como municipios del “Centro” a aquéllos que registran valores superiores a la media de la Comunidad en alguno de los dos factores mencionados. De este modo, los municipios integrantes de la zona denominada como “Periferia” serán aquellas localidades madrileñas que obtienen valores inferiores de la media en ambos factores explicativos.

Figura V.3.4. Estructuras espaciales en los ámbitos provincial y municipal



Fuente: Elaboración propia a partir de MapInfo.

En la Figura V.3.4. se advierte claramente que la distribución de los factores F1, F2, explicativos de la RFD en los municipios de la CM, siguen un esquema espacial del tipo

Centro-Periferia. En concreto, el Centro estaría formado por la Capital, los municipios del Área Metropolitana, parte de la Sierra Central (excepto Robledo de Chavela, Santa María de la Alameda, Valdemaqueda y Zarzalejo), parte del Henares (excepto Anchuelo, Corpa, Pezuela de las Torres, Santorcaz, Los Santos de la Humosa, Valdeavero y Valverde de Alcalá), así como también 13 municipios de la Sierra Norte, 9 municipios de la zona Sur, 8 del Sureste y 8 del Suroeste.

V.3.7. Análisis y validación de los resultados

Los resultados obtenidos para la RFD per capita de los municipios de la CM, en 2000, siguen también un esquema Centro-Periferia, bastante homogéneo desde el punto de vista geográfico, como puede observarse en la Figura V.3.5.

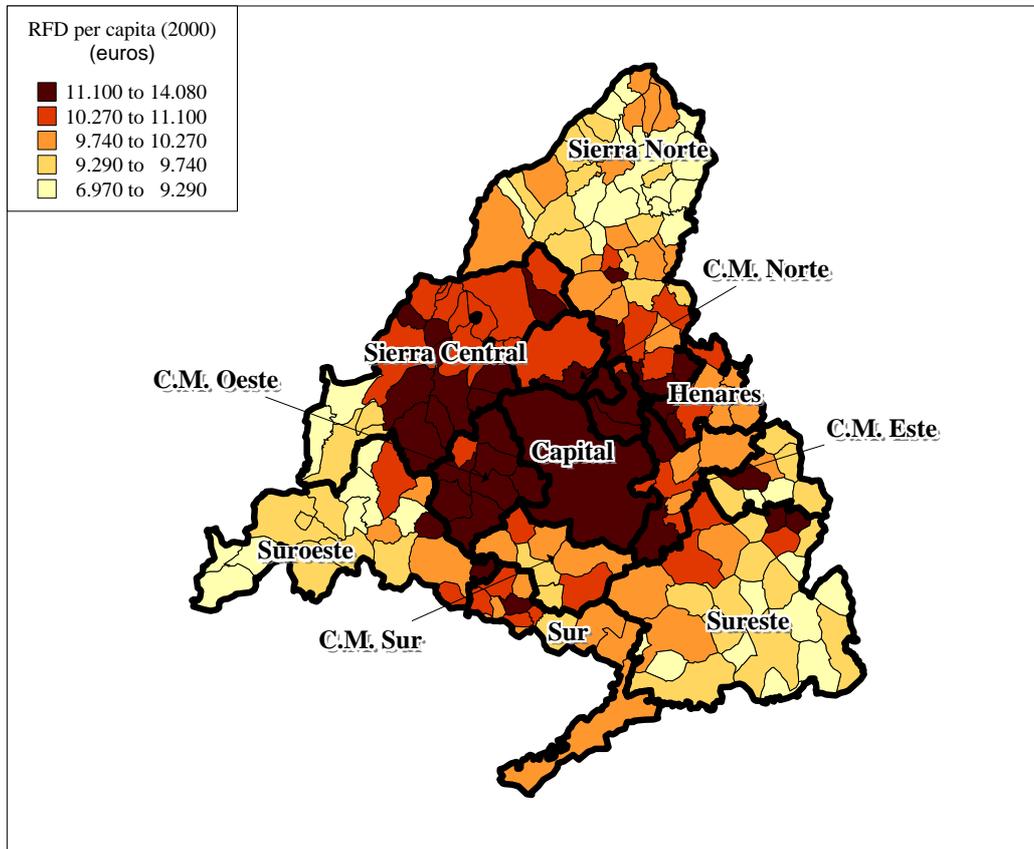
Los municipios con mayor nivel de RFD (ver Tabla V.3.12.) se localizan en Madrid capital, Corona Metropolitanas Oeste y Norte, y parte de la Sierra Central y Henares. En concreto, hay localidades en la CM que registran valores máximos de RFD en España, por encima de los 13.000 €/hab. (la media nacional es de 9.627 €/hab., en 2000), como es el caso de los municipios de la Corona Metropolitana Oeste (Las Rozas de Madrid, Boadilla del Monte, Majadahonda, Pozuelo de Alarcón y Villaviciosa de Odón).

En el otro extremo, los municipios situados en las “esquinas” o vértices de la CM (Sierra Norte, Suroeste y Sureste) son los que registran menores valores de RFD per capita, con menos de 8.500 € (1,4 millones de pesetas). En general, se trata de municipios rurales o de escasa población, con un nivel de RFD inferior a la media de España, algunos de ellos con menos de 7.000 €/per capita (1,2 millones de pesetas), como es el caso de Cervera de Buitrago, Prádena del Rincón y Robledillo de la Jara, todos ellos en la zona de la Sierra Norte madrileña.

Tabla V.3.12. Municipios con valores extremos de RFD (€) per capita (2000)

MUNICIPIO	ZONAS	RFD pc	MUNICIPIO	ZONAS	RFD pc
Rozas de Madrid (Las)	Área Metropolitana	14.075	Cervera de Buitrago	Sierra Norte	6.974
Boadilla del Monte	Área Metropolitana	13.754	Prádena del Rincón	Sierra Norte	6.974
Majadahonda	Área Metropolitana	13.634	Robledillo de la Jara	Sierra Norte	7.301
Pozuelo de Alarcón	Área Metropolitana	13.563	Puebla de la Sierra	Sierra Norte	8.151
Villanueva de la Cañada	Área Metropolitana	13.409	Corpa	Henares	8.275
Ajalvir	Henares	12.999	Hiruela (La)	Sierra Norte	8.275
Tres Cantos	Área Metropolitana	12.999	Serna del Monte (La)	Sierra Norte	8.379
Valdeolmos-Alalpardo	Sierra Norte	12.942	Cenicientos	Suroeste	8.458
Venturada	Sierra Norte	12.717	Robregordo	Sierra Norte	8.484
Rivas-Vaciamadrid	Área Metropolitana	12.655	Acebeda (La)	Sierra Norte	8.552

Figura V.3.5. Distribución de la RFD por habitante en los municipios de la Comunidad de Madrid (año 2000)



Fuente: Elaboración propia a partir de MapInfo.

Una forma de **validación** de los resultados sería la comparación de los mismos con los obtenidos, con otra metodología, por el Instituto de Estadística de la Comunidad de Madrid (IECAM). Efectivamente, tal como se desarrolla en el Anexo II, el IECAM publica desde 2001 un CD-ROM en el que se incluyen los valores de RFD por habitante de los municipios de la CM. Para ello, se ha escogido un modelo único basado en la explotación del IRPF municipal, como en Arcarons *et al.* (1998). En concreto, la base imponible del IRPF ha sido estimada en un 85% del total de la renta familiar disponible en la Comunidad de Madrid, por lo que se considera adecuada como indicador.

Sin embargo, el IRPF (aunque corregido) no recoge adecuadamente las rentas de los hogares agrarios ni la economía sumergida. Por este motivo, como se aprecia en la Tabla V.3.13., algunos municipios aparecen muy infravalorados en renta, como La Hiruela y Madarcos, con una RFD por habitante en 1999, inferior a 6.000 € (1 millón de pesetas), mientras que otras localidades, debido al ajuste realizado con el dato de renta per capita de la Comunidad procedente de la Contabilidad Regional, presentan valores de RFD bastante sobrevalorados, con valores medios superiores a 16.500 € (2,7 millones de pesetas) en los municipios de Boadilla del Monte, Pozuelo de Alarcón y Las Rozas de Madrid.

Tabla V.3.13. Comparativa de las estimaciones de RFD realizadas por el modelo (para 2000) y el IECAM (para 1999)

MUNICIPIO	ZONAS	RFD pc (modelo)	RFD pc (IECAM)	MUNICIPIO	ZONAS	RFD pc (modelo)	RFD pc (IECAM)
Rozas de Madrid (Las)	Área Metropolitana	14.075	16.768	Cervera de Buitrago	Sierra Norte	6.974	7.629
Boadilla del Monte	Área Metropolitana	13.754	16.603	Prádena del Rincón	Sierra Norte	6.974	6.820
Majadahonda	Área Metropolitana	13.634	16.537	Robledillo de la Jara	Sierra Norte	7.301	9.502
Pozuelo de Alarcón	Área Metropolitana	13.563	18.844	Puebla de la Sierra	Sierra Norte	8.151	7.755
Villanueva de la Cañada	Área Metropolitana	13.409	16.444	Corpa	Henares	8.275	7.132
Tres Cantos	Área Metropolitana	12.999	15.103	Hiruela (La)	Sierra Norte	8.275	5.350
Ajalvir	Henares	12.999	10.441	Serna del Monte (La)	Sierra Norte	8.379	6.998
Valdeolmos-Alalpardo	Sierra Norte	12.942	14.557	Cenicientos	Suroeste	8.458	6.569
Venturada	Sierra Norte	12.717	14.743	Robregordo	Sierra Norte	8.484	6.692
Rivas-Vaciamadrid	Área Metropolitana	12.655	12.032	Acebeda (La)	Sierra Norte	8.552	6.638
Torreloñones	Sierra Central	12.631	16.337	Alameda del Valle	Sierra Norte	8.701	7.718
Hoyo de Manzanares	Sierra Central	12.581	13.348	Valdeañuela	Sierra Central	8.721	7.425
Algete	Sierra Norte	12.296	13.445	Puentes Viejas	Sierra Norte	8.823	8.143
Villaviciosa de Odón	Área Metropolitana	12.285	13.522	Piñuécar-Gandullas	Sierra Norte	8.827	7.346
Paracuellos de Jarama	Área Metropolitana	12.240	12.457	Madarcos	Sierra Norte	8.871	5.853

En consecuencia, las estimaciones realizadas por el IECAM para la CM muestran una realidad más dispar entre municipios, con mayores diferencias de renta dentro de esta Comunidad de las realmente existentes. Efectivamente, según el IECAM, la diferencia entre la RFD obtenida, durante el año 1999, por un residente medio en el municipio más rico y más pobre de la CM es de 13.494 € (más de 2,2 millones de pesetas), muy superior al recorrido que presenta el dato de RFD estimado en este trabajo, de 7.101 € (1,2 millones de pesetas). Parece difícil admitir tales diferencias, en términos de RFD, entre municipios de una misma comunidad autónoma, debido a la labor redistributiva del Estado.

Por último, en la Tabla V.3.14, se presentan las estimaciones de la RFD per capita, para el año 2000, realizadas en este trabajo para todos los municipios de la CM.

Tabla V.3.14. RFD per capita de los municipios de la Comunidad de Madrid (2000)

Municipio	Zonas	Población (1.1.01)	RFD pc (€)
Madrid	Capital	2.957.058	11.917
Alcalá de Henares	Área Metropolitana	172.418	10.127
Alcobendas	Área Metropolitana	92.537	11.956
Alcorcón	Área Metropolitana	147.787	10.344
Boadilla del Monte	Área Metropolitana	23.654	13.754
Brunete	Área Metropolitana	5.414	11.276
Colmenar Viejo	Área Metropolitana	34.194	10.662
Coslada	Área Metropolitana	78.774	10.544

Municipio	Zonas	Población (1.1.01)	RFD pc (€)
Fuenlabrada	Área Metropolitana	178.221	9.737
Getafe	Área Metropolitana	150.532	10.141
Leganés	Área Metropolitana	173.426	10.041
Majadahonda	Área Metropolitana	49.137	13.634
Mejorada del Campo	Área Metropolitana	16.565	10.049
Móstoles	Área Metropolitana	197.062	10.122
Paracuellos de Jarama	Área Metropolitana	6.334	12.240
Parla	Área Metropolitana	77.157	9.533
Pinto	Área Metropolitana	30.114	10.508
Pozuelo de Alarcón	Área Metropolitana	68.470	13.563
Rivas-Vaciamadrid	Área Metropolitana	32.228	12.655
Rozas de Madrid (Las)	Área Metropolitana	59.002	14.075
San Fernando de Henares	Área Metropolitana	35.089	10.772
San Sebastián de los Reyes	Área Metropolitana	59.646	11.157
Torrejón de Ardoz	Área Metropolitana	97.546	10.268
Tres Cantos	Área Metropolitana	36.598	12.999
Velilla de San Antonio	Área Metropolitana	7.447	10.462
Villanueva de la Cañada	Área Metropolitana	11.429	13.409
Villanueva del Pardillo	Área Metropolitana	5.296	10.576
Villaviciosa de Odón	Área Metropolitana	20.832	12.285
Ajalvir	Henares	2.386	12.999
Anchuelo	Henares	579	9.756
Camarma de Esteruelas	Henares	2.543	10.152
Cobeña	Henares	3.060	11.559
Corpa	Henares	418	8.275
Daganzo de Arriba	Henares	4.208	11.024
Fresno de Torote	Henares	722	10.126
Meco	Henares	6.923	9.748
Pezuela de las Torres	Henares	491	9.499
Ribatejada	Henares	390	10.417
Santorcaz	Henares	583	9.474
Santos de la Humosa (Los)	Henares	930	9.383
Torres de la Alameda	Henares	4.541	9.409
Valdeavero	Henares	636	9.984
Valverde de Alcalá	Henares	309	8.992
Villalbilla	Henares	4.595	11.324
Alpedrete	Sierra Central	8.121	11.447
Becerril de la Sierra	Sierra Central	3.571	10.884
Boalo (El)	Sierra Central	3.447	10.760
Cercedilla	Sierra Central	5.826	11.075
Collado Mediano	Sierra Central	4.766	11.870
Collado Villalba	Sierra Central	44.872	11.073
Colmenarejo	Sierra Central	4.968	11.854
Escorial (El)	Sierra Central	11.209	11.185
Galapagar	Sierra Central	23.693	11.763
Guadarrama	Sierra Central	10.275	10.757
Hoyo de Manzanares	Sierra Central	6.013	12.581

Municipio	Zonas	Población (1.1.01)	RFD pc (€)
Manzanares el Real	Sierra Central	4.292	11.079
Miraflores de la Sierra	Sierra Central	3.854	10.748
Molinos (Los)	Sierra Central	3.497	11.225
Moralzarzal	Sierra Central	6.134	11.097
Navacerrada	Sierra Central	1.913	10.568
Robledo de Chavela	Sierra Central	2.439	9.720
San Lorenzo de El Escorial	Sierra Central	12.455	10.984
Santa María de la Alameda	Sierra Central	702	9.097
Soto del Real	Sierra Central	5.850	11.962
Torrelorones	Sierra Central	14.717	12.631
Valdemaqueda	Sierra Central	595	8.721
Valdemorillo	Sierra Central	6.482	11.310
Zarzalejo	Sierra Central	1.061	9.468
Acebeda (La)	Sierra Norte	56	8.552
Alameda del Valle	Sierra Norte	193	8.701
Algete	Sierra Norte	15.360	12.296
Atazar (El)	Sierra Norte	92	8.899
Berrueco (El)	Sierra Norte	334	9.487
Berzosa del Lozoya	Sierra Norte	131	9.605
Braojos	Sierra Norte	161	9.421
Buitrago del Lozoya	Sierra Norte	1.565	10.059
Bustarviejo	Sierra Norte	1.528	9.351
Cabanillas de la Sierra	Sierra Norte	473	10.342
Cabrera (La)	Sierra Norte	1.827	10.106
Canencia	Sierra Norte	404	9.564
Cervera de Buitrago	Sierra Norte	88	6.974
Fuente el Saz de Jarama	Sierra Norte	4.369	10.937
Garganta de los Montes	Sierra Norte	317	9.214
Gargantilla del Lozoya y Pinilla de Buitrago	Sierra Norte	245	9.613
Gascones	Sierra Norte	114	9.582
Guadalix de la Sierra	Sierra Norte	3.394	10.037
Hiruela (La)	Sierra Norte	77	8.275
Horcajo de la Sierra	Sierra Norte	146	9.999
Horcajuelo de la Sierra	Sierra Norte	108	9.807
Lozoya	Sierra Norte	459	9.784
Lozoyuela-Navas-Sieteiglesias	Sierra Norte	680	9.072
Madarcos	Sierra Norte	23	8.871
Molar (El)	Sierra Norte	3.898	10.299
Montejo de la Sierra	Sierra Norte	318	9.923
Navalafuente	Sierra Norte	588	10.136
Navarredonda y San Mamés	Sierra Norte	106	9.484
Patones	Sierra Norte	366	9.034
Pedrezuela	Sierra Norte	1.613	10.049
Pinilla del Valle	Sierra Norte	155	9.658
Piñuécar-Gandullas	Sierra Norte	140	8.827
Prádena del Rincón	Sierra Norte	103	6.974
Puebla de la Sierra	Sierra Norte	96	8.151

Municipio	Zonas	Población (1.1.01)	RFD pc (€)
Puentes Viejas	Sierra Norte	407	8.823
Rascafría	Sierra Norte	1.637	10.064
Redueña	Sierra Norte	188	9.479
Robledillo de la Jara	Sierra Norte	93	7.301
Robregordo	Sierra Norte	88	8.484
San Agustín de Guadalix	Sierra Norte	6.333	12.135
Serna del Monte (La)	Sierra Norte	115	8.379
Somosierra	Sierra Norte	98	9.819
Talamanca de Jarama	Sierra Norte	1.547	10.690
Torrelaguna	Sierra Norte	2.898	10.253
Torremocha de Jarama	Sierra Norte	351	9.836
Valdemanco	Sierra Norte	517	9.262
Valdeolmos-Alalpardo	Sierra Norte	1.804	12.942
Valdepiélagos	Sierra Norte	306	9.367
Valdetorres de Jarama	Sierra Norte	2.131	9.840
Vellón (El)	Sierra Norte	1.190	9.705
Venturada	Sierra Norte	919	12.717
Villavieja del Lozoya	Sierra Norte	185	9.388
Aranjuez	Sur	40.113	10.163
Batres	Sur	920	10.936
Casarrubuelos	Sur	875	9.825
Ciempozuelos	Sur	14.170	10.248
Cubas de la Sagra	Sur	1.793	10.364
Griñón	Sur	5.040	11.275
Humanes de Madrid	Sur	9.937	10.210
Moraleja de Enmedio	Sur	3.195	10.456
Serranillos del Valle	Sur	901	9.983
Torrejón de la Calzada	Sur	4.462	10.456
Torrejón de Velasco	Sur	2.085	9.619
Valdemoro	Sur	30.986	10.119
Ambite	Sureste	302	9.470
Arganda del Rey	Sureste	32.157	10.331
Belmonte de Tajo	Sureste	1.158	9.094
Brea de Tajo	Sureste	448	9.113
Campo Real	Sureste	2.839	9.824
Carabaña	Sureste	1.180	9.658
Chinchón	Sureste	4.270	10.088
Colmenar de Oreja	Sureste	5.548	9.396
Estremera	Sureste	1.049	9.582
Fuentidueña de Tajo	Sureste	1.471	8.997
Loeches	Sureste	3.176	10.344
Morata de Tajuña	Sureste	5.485	9.610
Nuevo Baztán	Sureste	3.677	11.976
Olmeda de las Fuentes	Sureste	146	11.166
Orusco de Tajuña	Sureste	644	9.079
Perales de Tajuña	Sureste	2.089	9.530
Pozuelo del Rey	Sureste	181	9.601

Municipio	Zonas	Población (1.1.01)	RFD pc (€)
San Martín de la Vega	Sureste	11.415	9.768
Tielmes	Sureste	2.003	9.257
Titulcia	Sureste	934	9.219
Valdaracete	Sureste	604	9.206
Valdelaguna	Sureste	573	9.296
Valdilecha	Sureste	1.930	9.604
Villaconejos	Sureste	2.946	9.003
Villamanrique de Tajo	Sureste	614	9.166
Villar del Olmo	Sureste	1.517	10.695
Villarejo de Salvanés	Sureste	5.753	9.464
Álamo (El)	Suroeste	4.932	10.274
Aldea del Fresno	Suroeste	1.503	9.697
Arroyomolinos	Suroeste	4.700	11.597
Cadalso de los Vidrios	Suroeste	2.311	9.141
Cenicientos	Suroeste	1.835	8.458
Chapinería	Suroeste	1.439	8.973
Colmenar del Arroyo	Suroeste	913	9.218
Fresnedillas de la Oliva	Suroeste	855	9.195
Navalagamella	Suroeste	1.200	10.287
Navalcarnero	Suroeste	14.256	9.792
Navas del Rey	Suroeste	1.746	9.547
Pelayos de la Presa	Suroeste	1.422	9.609
Quijorna	Suroeste	1.194	10.129
Rozas de Puerto Real	Suroeste	273	9.028
San Martín de Valdeiglesias	Suroeste	6.191	9.290
Sevilla la Nueva	Suroeste	4.054	11.372
Villa del Prado	Suroeste	4.106	9.398
Villamanta	Suroeste	1.663	9.493
Villamantilla	Suroeste	337	10.074
Villanueva de Perales	Suroeste	550	8.894

CAPÍTULO VI

Conclusiones y reflexión final

VI.1. PRINCIPALES CONCLUSIONES Y RESUMEN DE LA INVESTIGACIÓN

Una obra de estas características difícilmente podrá tener un número limitado de conclusiones que recojan con exactitud el amplio espectro de temas que han sido tratados en los Capítulos anteriores. Por este motivo, se ha optado por seleccionar aquellas ideas y sugerencias que creemos resultan de mayor interés para el lector interesado en el tema, ordenadas por Capítulos, para su más fácil identificación.

Previamente, podrían resumirse, tal como se hizo en la Introducción, los objetivos de carácter general planteados por esta tesis doctoral, en los siguientes cuatro puntos:

- El objetivo fundamental de esta tesis consiste en **proponer la predicción de datos espaciales como parte de la econometría espacial**, a partir de una metodología fundamentada en instrumentos exploratorios y confirmatorios propios de esta disciplina. En concreto, se desarrolla una subdisciplina de la predicción espacial, denominada **predicción-extrapolación espacial**, distinta del método geostadístico de interpolación espacial, entendida como el conjunto de métodos inferenciales destinados a obtener datos microterritoriales a partir de información espacial agregada.
- La consecución de este objetivo general exige una **presentación y sistematización del conjunto de técnicas de análisis exploratorio y confirmatorio de la econometría espacial**, realizada en el cuerpo central de esta obra, con la ayuda de un capítulo completo y actualizado de Bibliografía, así como de unos Anexos, que amplían algunas ideas sin romper el hilo de la exposición.

- **La metodología propuesta se evalúa a través de un ejercicio de predicción-extrapolación de la renta familiar disponible de los municipios de la Comunidad de Madrid**, en el que se hacen confluír tanto la utilización de técnicas de econometría espacial (a través de un “software” adecuado), como la experiencia de la autora en análisis microterritoriales, como Directora A. del Área de Economía Espacial Microterritorial del Instituto Lawrence R. Klein (Universidad Autónoma de Madrid), co-dirigido por D. José Vicéns Otero, director de esta tesis.
- Dado el interés y el cierto desconocimiento que aún pesa sobre el tema de la econometría espacial, esta tesis se plantea también, como objetivo, ser de utilidad para todos los investigadores sociales interesados en el análisis territorial, en cualquiera de sus formas, así como **contribuir, en la medida de lo posible, a la difusión de las técnicas de econometría espacial en nuestro país**, sobre todo en el ámbito de los estudios universitarios.

A continuación, se expone un elenco de conclusiones y sugerencias, desglosadas por capítulos.

Capítulo II: Econometría de datos espaciales

- La econometría espacial es una disciplina de la econometría general que incluye el conjunto de técnicas de especificación, estimación, contraste y predicción necesarias para el tratamiento de los datos espaciales. Es decir, **existe una actividad de predicción, parte del proceso econométrico espacial, que consiste en la estimación, para determinadas localizaciones, de valores no existentes.**
- **La econometría espacial se justifica por la necesidad de tratamiento propio de los datos espaciales**, que se diferencian sustancialmente de los datos temporales a causa de los llamados efectos de dependencia y heterogeneidad espacial, propios de la naturaleza georreferenciada y multidireccional de la información geográfica.
- **En la aún joven historia de la econometría espacial, podrían destacarse dos momentos clave, los años 1979 y 1988**, coincidentes con la publicación de “Spatial econometrics”, de Paelinck y Klaasen, y “Spatial econometrics: Methods and models”, de Anselin, respectivamente. La extrema dificultad de comprensión del primer libro ha motivado que, pese a que uno de sus dos autores, Jean Paelinck, sea unánimemente considerado como el padre de la econometría espacial, probablemente sea menos leído y citado que el segundo libro, de Luc Anselin, mucho más asequible en el contexto interdisciplinar de las ciencias sociales.
- **En la actualidad, la econometría espacial sigue siendo bastante ignorada en el campo económico y España no es una excepción en este aspecto: la difusión de la econometría espacial, tanto en el terreno teórico como aplicado, es también escasa.** La mayor parte de la literatura publicada en los últimos años sobre estos temas ha tenido poco eco entre la mayoría de los analistas de economía empírica, con la excepción de algunos investigadores que están dedicando sus esfuerzos a estos temas.

Mención especial cabría realizar de la labor de equipo desarrollada en este campo por el “**Grup d'Anàlisi Quantitativa Regional**”, de la Universidad de Barcelona. Además, hay también algunos investigadores interesados en el tema, por ejemplo, en las universidades Autónoma de Madrid, Complutense, Extremadura, Granada, Politécnica de Cartagena, Pontificia de Comillas y Zaragoza, así como en la Fundación Tomillo.

- **La naturaleza compleja y continua del espacio geográfico exige una alta tecnología informática** capaz de visualizar y tratar estadísticamente los datos procedentes del contexto espacial: los sistemas de información geográfica (GIS) que, hasta hace poco, resultaban extremadamente complejos de manejar. Los nuevos productos GIS no requieren de una gran especialización informática, pues son de fácil manejo para cualquier usuario que desee trabajar con datos espaciales.
- **La utilización del GIS, junto con el análisis de datos y modelos espaciales, es cada vez más común en los campos de la economía aplicada, política económica, economía urbana y medioambiental, y economía del desarrollo**, sobre todo en Estados Unidos. En los últimos años, se ha prestado especial atención a la integración del análisis espacial en el entorno GIS, tanto desde el campo académico como desde el comercial, planteándose diversos enfoques sobre el contenido concreto que debería tener el paquete de herramientas de análisis espacial de un sistema de información geográfica.
- **Los GIS se han revelado como extremadamente eficaces en el tratamiento de grandes bases de datos microterritoriales**, cuya demanda es creciente tanto desde organismos públicos como empresas privadas.
- **Debe advertirse, no obstante, que las bases de datos habitualmente disponibles para el ámbito microterritorial, no suelen estar convenientemente depuradas ni geocodificadas**, lo que dificulta especialmente el trabajo a este nivel. A esta dificultad deben añadirse otros problemas propios de los datos espaciales, como el sesgo de la agregación, motivado por un cambio de escala, errores de muestreo por la presencia de los efectos espaciales, errores derivados del proceso de informatización de los datos, gran tamaño de las bases de datos, etc.
- **La complejidad de los datos espaciales y su correcto tratamiento hacen necesaria una formación teórico-práctica adecuada en los investigadores**, que deberían profundizar en la literatura científica existente, así como manejar los programas informáticos de tratamiento de información espacial, dado que la gran dimensión de las bases de datos y la multitud de formas de obtenerlas dan lugar a un número de cuestiones que no siempre se resuelven satisfactoriamente y requieren de un mayor esfuerzo de investigación.
- **De un modo particular, debería impulsarse una formación interdisciplinar** que considerara el acervo científico acumulado por diversas ciencias sociales (economía, investigación de mercados, geografía, sociología, ciencia política, etc.) y técnicas, como la informática, para lograr una mayor integración de los investigadores en temas espaciales y geográficos.

Capítulo III: Análisis exploratorio de datos espaciales

- **El análisis exploratorio de datos (AED), que es especialmente adecuado cuando no se dispone de una hipótesis a contrastar claramente definida, consiste en un conjunto de técnicas específicas orientadas a encontrar, en una base de datos, alguna estructura o conglomerados (“clusters”).** Los métodos modernos de AED hacen hincapié en la interacción existente entre el conocimiento humano y la informática en forma de gráficos estadísticos dinámicos que permiten al usuario la manipulación directa de diversas “vistas” de los datos (histogramas, diagramas de caja, gráficos q-q, los diagramas de dispersión matricial, etc.).
- **El análisis exploratorio de datos espaciales (AEDE) surge como una disciplina del AED, especializada en el tratamiento específico de los datos geográficos.** Se trata de una técnica que combina el análisis estadístico con el gráfico, dando lugar a lo que podría denominarse una “visualización científica” que, a los contrastes estadísticos sobre los efectos espaciales de dependencia y heterogeneidad, une un amplio marco de gráficos o “vistas” múltiples y dinámicas sobre la información geográfica, que hacen de estos métodos algo más que simples mapas o gráficos estáticos de representación.
- **El AEDE puede ser abordado desde dos puntos de vista, geoestadístico y reticular (“lattice”), siendo este último el habitual en econometría espacial.** La perspectiva geoestadística contempla los datos espaciales como una muestra de datos puntuales procedentes de distribuciones espaciales continuas, mientras que para el punto de vista reticular, se trata de localizaciones geográficas discretas de puntos o polígonos. Los programas informáticos especializados en AEDE suelen estar marcados por alguna de estas tendencias.
- **El programa informático SpaceStat, desarrollado por Luc Anselin, está marcado por una orientación hacia el análisis econométrico espacial** (exploratorio y confirmatorio) y permite la estimación y contraste de los efectos espaciales en modelos de regresión lineal (modelos de dependencia y heterogeneidad espacial). Las extensiones de SpaceStat para el GIS ArcView tienen como objetivo facilitar el intercambio y visualización de datos de algunos resultados obtenidos por el paquete SpaceStat susceptibles de representación geográfica, haciendo especial hincapié en los métodos del AEDE. De esta forma, un mapa puede constituir una de las varias “vistas” posibles de una base de datos, junto con otras vistas (tabla, histograma, gráfico de caja u otro gráfico de carácter estadístico), que se encuentran conectadas de forma que cualquier observación señalada en una de ellas queda simultáneamente destacada en el resto.
- **Un buen método gráfico de AEDE es aquél que es capaz de analizar y representar las dos características fundamentales de toda distribución espacial: el elemento global y el local.** En las variables geográficas, el elemento global o alisado está referido a todo el mapa, como la autocorrelación espacial global, mientras que el elemento local o asperezas (“rough”) se refiere a aquellos datos localizados a cierta distancia del

elemento de alisado (mediana), como los puntos atípicos (“outliers”), situados bajo/sobre el primer/tercer cuartil de un diagrama de caja.

- **El objeto principal de análisis del AEDE en una variable, tanto desde un punto de vista global como local, es la detección y contrastación, en la misma, de los efectos de dependencia y heterogeneidad espacial.** La dependencia espacial se produce cuando una variable se distribuye de forma sistemática en el espacio, mientras que la heterogeneidad espacial es entendida como ausencia de estabilidad en el espacio de un fenómeno. A diferencia de lo que sucede con la dependencia, la heterogeneidad espacial es una cuestión que, en algunos casos, podría ser bien resuelta a través de procedimientos de la econometría clásica, aunque se recomienda el uso, también en estos casos, de técnicas de econometría espacial, debido a la complejidad de ciertas estructuras conjuntas de dependencia y heterogeneidad espacial, que hacen muy difícil su distinción.
- **Las interacciones espaciales que suelen producirse en muchos fenómenos humanos son la causa verdadera del efecto de dependencia espacial** y podrían expresarse matemáticamente como una relación funcional entre los valores que adopta una variable Y en la localización i (y_i) y los valores de dicha variable en un conjunto N de localizaciones situadas en el territorio en estudio. En la perspectiva reticular o “lattice”, que supone que las observaciones se encuentran organizadas en unidades espaciales discretas, **el conjunto de unidades vecinas a una dada suele representarse gráficamente como una estructura reticular y, de forma matemática, como una matriz de interacciones.**
- **En la matriz de interacciones, contigüidades o contactos espaciales, que es la más utilizada por su simplicidad, la situación de vecindad entre dos unidades espaciales se podría expresar mediante valores de tipo 0-1.** Es decir, si dos unidades espaciales tienen una frontera común de longitud no nula, se considera que son contiguas y se les asigna el valor 1 y, si no tienen ninguna frontera en común, se les asigna el valor 0. Se trata de una matriz simétrica (cuya diagonal principal está, por convenio, constituida por ceros), que no es capaz de incorporar influencias no recíprocas y que considera, como único determinante de las interdependencias regionales, la adyacencia física, descuidando con ello otro tipo de relaciones entre regiones alejadas, no contiguas (por ejemplo, estrechas relaciones comerciales).
- **La matriz de pesos espaciales, a diferencia de la matriz binaria de interacciones, permite al investigador la elección del conjunto de ponderaciones que se considere apropiado para cada fenómeno,** lo que supone una mayor flexibilidad en la definición de la estructura de interdependencias de un sistema regional y permite considerar cuestiones como las barreras naturales o el tamaño de las regiones. La especificación adecuada de los elementos de esta matriz (w_{ij}) es uno de los puntos metodológicos más difíciles y controvertidos en la econometría espacial.
- **El operador retardo espacial permite retardar o desplazar el valor de una variable en función de los valores que adopta en localizaciones vecinas,** simplificando las relaciones de interdependencia existentes entre distintas unidades espaciales. El retardo

espacial de una variable en una localización se obtiene como un promedio ponderado de los valores de dicha variable en localizaciones vecinas, a través de unos pesos o ponderaciones fijas.

- **Aunque la mera observación en un mapa de la distribución de una variable espacial permite captar ya, de forma intuitiva, la existencia o no de patrones de comportamiento en dicha variable, esta información será siempre subjetiva, por lo que resulta fundamental contar con medidas estadísticas capaces de detectar la presencia significativa de los efectos espaciales en dicha variable.** En concreto, los contrastes de autocorrelación espacial, en combinación con las herramientas del AEDE, ofrecen criterios estadísticos objetivos que permiten confirmar o rechazar la presencia de tendencias o estructuras espaciales en la distribución de una variable.
- **El fenómeno de autocorrelación espacial puede contrastarse estadísticamente desde la doble perspectiva, global y local.** Los estadísticos de autocorrelación espacial global son las primeras formulaciones propuestas en la literatura, con capacidad de resumir el esquema general de dependencia presente en una variable espacial en un único indicador. Recientemente, se han propuesto también los tests de autocorrelación espacial local, muy útiles para la detección de agrupaciones (“clusters”) de este fenómeno en el espacio geográfico general.
- **La perspectiva global del fenómeno de autocorrelación espacial tiene por objeto contrastar la hipótesis de que una variable se encuentre distribuida de forma totalmente aleatoria en un sistema espacial o, si por el contrario, existe algún tipo de asociación significativa de valores similares o distintos entre regiones vecinas.** Los estadísticos de dependencia espacial global más conocidos en la literatura son el test I de Moran, c de Geary y $G(d)$ de Getis y Ord.
- **Los contrastes de autocorrelación espacial local (familia G de Getis y LISA de Anselin) buscan la presencia de pequeños “clusters” espaciales en los que una variable, con dependencia global, experimenta una concentración/escasez importante de valores en alguna localización del espacio geográfico.** Los estadísticos LISA, además, permiten conocer la contribución de cada región al indicador global de dependencia, que no tiene por qué ser la misma.
- **De rechazarse la hipótesis de no autocorrelación espacial** en alguna de las variables integrantes de un modelo de regresión, este fenómeno de asociación espacial deberá ser considerado explícitamente en dicho modelo a través de especificaciones adecuadas.

Capítulo IV: Análisis confirmatorio: modelos de regresión espacial

- **Las econometría espacial hace posible la agrupación de múltiples especificaciones propias de los procesos espaciales** y realizar tipologías de modelos, como las propuestas por Anselin, y por Florax y Folmer.

- **En el contexto de los modelos de regresión lineal, tanto el efecto de dependencia como heterogeneidad espacial pueden estar presentes en la especificación de un modelo.** La dependencia espacial puede adoptar dos formas, sustantiva y residual, según que el efecto de autocorrelación espacial esté presente en variables sistemáticas (endógena y/o exógenas) o en el término de la perturbación aleatoria, respectivamente. Por su parte, el fenómeno de heterogeneidad espacial puede manifestarse como varianza no constante en los residuos de una regresión (heteroscedasticidad) o como variación, continua o discreta, en los coeficientes del modelo (inestabilidad paramétrica).
- **La especificación más sencilla del efecto de dependencia espacial es el modelo espacial autorregresivo de orden 1, SAR(1),** adecuada para expresar situaciones en las que los valores que adopta una variable determinada dependan sistemáticamente de la localización geográfica de la misma.
- **El modelo básico de regresión lineal (MBRL) será una especificación correcta en el ámbito espacial cuando no se produzca dependencia espacial en ninguna variable del mismo o cuando este efecto sea recogido correcta y totalmente por un grupo de regresores, a través de una relación estructural.** La presencia de dependencia espacial en la variable endógena de un modelo hace insuficiente la especificación del MBRL, debiendo utilizarse, en estos casos, otras especificaciones, como el modelo del error espacial y el modelo del retardo espacial.
- **El modelo del error espacial es, con mucho, la especificación más conocida.** En este modelo, la existencia de ciertos factores (variables no explícitamente consideradas en el modelo) trasladan hacia los términos del error la configuración de agrupación de valores (autocorrelación) presente en la variable endógena. Habitualmente, la distribución de la perturbación aleatoria suele especificarse como un proceso espacial autorregresivo de orden 1, SAR(1).
- **El modelo del retardo espacial incorpora la influencia de las variables omitidas a través de la variable dependiente espacialmente retardada,** es decir, de los valores que, para cada punto, adopta la variable endógena en un grupo de puntos vecinos.
- **El efecto de dependencia espacial en un modelo puede ser anulado a través de métodos de filtrado aplicados a las variables afectadas,** de forma que un modelo con problemas de autocorrelación espacial pueda expresarse correctamente a través de la especificación del MBRL. Esta transformación constituye también una solución a situaciones de predicción-extrapolación de datos espaciales en los que se necesite extrapolar, en concreto, un modelo del retardo espacial a un ámbito microterritorial, para el que no se dispone, obviamente, de la variable endógena, ni por tanto, de su correspondiente retardo espacial.
- **Anselin ha planteado recientemente una nueva taxonomía de modelos de regresión espacial capaces de expresar diversas formas de externalidades (“spillovers”),** fenómeno de gran actualidad en las ciencias sociales. Para ello, el autor parte de una forma reducida de los modelos de regresión caracterizada por incluir, en la parte derecha de la ecuación, únicamente variables exógenas, la perturbación aleatoria o

transformaciones espaciales de las mismas. En estos modelos, las externalidades se formulan como multiplicadores espaciales, que expresan el fenómeno de transmisión del impacto que, un cambio en el punto i , produce hacia el resto de localizaciones.

- **La gran innovación de la nueva taxonomía de externalidades de Anselin, estriba en la diferenciación que se hace, en el contexto de los modelos de regresión lineal, entre dependencia global y local**, dependiendo de que la estructura formal de dicha dependencia tenga la forma de procesos espaciales autorregresivos (SAR) o de medias móviles (SMA), respectivamente.
- **El efecto de heterogeneidad espacial está directamente relacionado con la localización geográfica, el área o cualquier otra característica de las unidades espaciales muestrales de un modelo de regresión.** Este fenómeno surge cuando se trabaja con unidades espaciales (países, regiones, municipios, secciones censales) en las que un fenómeno se distribuye de manera distinta sobre el espacio, lo que suele ocurrir en situaciones del tipo centro-periferia, norte-sur, este-oeste, etc., así como en ejercicios de predicción-extrapolación de datos espaciales.
- **La especificación más sencilla del efecto de heterogeneidad espacial en forma de varianza no constante en los residuos (heteroscedasticidad) es el modelo del error heteroscedástico**, que es un caso particular del modelo espacial con perturbaciones aleatorias no esféricas. En este modelo, la varianza no constante de la perturbación aleatoria adquiere una estructura que podría especificarse como heteroscedasticidad aditiva, de coeficientes aleatorios y de grupos (“groupwise”).
- **La heterogeneidad espacial puede también manifestarse como inestabilidad paramétrica, es decir, una variación, continua o discreta, en los parámetros del modelo de regresión**, de modo que tanto la forma funcional como los parámetros de una regresión puedan variar según la localización geográfica siendo no homogéneos en toda la muestra de datos.
- **La heterogeneidad espacial como inestabilidad paramétrica continua implica que los parámetros del modelo adoptan un valor diferente para cada observación muestral.** Las especificaciones propias de este enfoque son el modelo de parámetros aleatorios de Hildreth-Houck, modelo de expansión espacial de Caseti y el modelo de regresiones ponderadas geográficamente (RPG) de Fotheringham.
- **La heterogeneidad espacial como inestabilidad paramétrica discreta implica que los parámetros del modelo adoptan valores diferentes únicamente en un número limitado de regímenes o estructuras espaciales.** Las especificaciones propias de este enfoque son el modelo ANOVA espacial (SANOVA) y modelo espacial de regresiones cambiantes (“switching regressions”).
- **El análisis confirmatorio espacial puede llevarse a cabo en un proceso que comienza en la especificación del modelo propuesto como un MBRL, sin efectos espaciales, que se estima por MCO.** A continuación, sobre la perturbación aleatoria se aplican una serie de contrastes de autocorrelación y heterogeneidad espacial con el

objeto de aceptar o rechazar la hipótesis nula de ausencia de efectos espaciales que, de ser rechazada, requerirá de nuevas re-especificaciones del modelo que incluyan convenientemente los efectos espaciales, así como de otros métodos de estimación que produzcan buenas propiedades en los estimadores.

- **Los contrastes de dependencia espacial, sustantiva y residual, se dividen en dos grupos: los tests “ad-hoc”, sin hipótesis alternativa definida, y los contrastes basados en las propiedades óptimas del estimador MV, rigurosamente estructurados en términos de una hipótesis nula/alternativa específicas.** Entre los primeros, se encuentran los tests I de Moran y K-R de Kelejian y Robinson, y en el grupo de contrastes MV estarían los tests asintóticos de Wald, razón de verosimilitud (LR) y multiplicadores de Lagrange (LM-ERR, LM-EL, LM-LAG, LM-LE, SARMA y LM-LAG $_{\lambda}$).
- **Los contrastes de heteroscedasticidad más citados son el test del multiplicador de Lagrange de Breusch y Pagan (B-P) y el test de Koenker-Basset (KB),** que es la versión "estudentizada" del anterior. El que se utilicen uno u otro depende de la distribución de la perturbación aleatoria que, cuando no es normal, produce resultados sesgados en el test BP, que se muestra débil en modelos con muestras pequeñas, por lo que se aconseja la aplicación del test KB.
- **En presencia de dependencia espacial, los estimadores MCO son sesgados e inconsistentes,** independientemente del comportamiento de la perturbación aleatoria, por lo que deben ser sustituidos por otros estimadores con mejores propiedades que, dependiendo de la especificación, podrían ser estimadores MCO robustos, mínimos cuadrados generalizados factibles (MCGF), máximo-verosímiles (MV) y estimadores de variables instrumentales (VI), como los mínimos cuadrados espaciales en dos etapas (MC2E) o métodos de remuestreo “bootstrap”.

Capítulo V: Predicción-extrapolación espacial de datos microterritoriales

- **La predicción espacial es la parte de la econometría espacial que tiene como misión estimar, para determinadas localizaciones, valores no existentes.** El habitual proceso econométrico de especificación, estimación y contraste puede culminar, también en el ámbito espacial, en un ejercicio de predicción espacial, que suele realizarse tanto a través de métodos de interpolación como extrapolación espacial.
- **La predicción-extrapolación espacial consiste en obtener, a partir de datos y relaciones procedentes de un ámbito espacial agregado, información estadística de ámbito microterritorial.** Por ejemplo, a partir de la estructura de dependencia que liga, en un ámbito superior (por ejemplo, provincial) y en un momento dado, una magnitud económica (la renta) con un grupo de indicadores socio-económicos (instrucción, empleo, actividad), disponibles también para todas las unidades microterritoriales (municipios), será posible estimar dicha magnitud económica en la escala espacial desagregada.

- **Además de la econometría espacial, la predicción-extrapolación debe asumir los principios del Problema de la Unidad Espacial Modificable (MAUP) y las Inferencias Ecológicas (IE)**, considerados desde hace años por otros investigadores sociales, como geógrafos, sociólogos y políticos.
- **El MAUP pone de manifiesto el problema del sesgo de la agregación, según el cual los estadísticos básicos de una variable pueden variar sensiblemente con la agregación de datos microterritoriales (individuos, secciones, municipios) a una escala geográfica superior (provincias, regiones)**. Por ejemplo, la agregación de unidades geográficas produce, por un lado, una disminución de la varianza y el grado de autocorrelación espacial de una variable y, por otro, un incremento (y, a veces, cambio de signo) en el coeficiente de correlación lineal entre dos variables (así como en el coeficiente de determinación, R^2 , de una regresión lineal). Por tanto, las relaciones estadísticas entre variables obtenidas en un nivel agregado nunca pueden identificarse directamente con las relaciones existentes entre las mismas variables en ámbitos geográficos inferiores.
- **Una inferencia ecológica (IE) consiste en un proceso por el que, a partir de datos agregados (ecológicos), se estiman datos microterritoriales o individuales no disponibles, lo que relaciona profundamente este tema con la predicción-extrapolación espacial**. Uno de los problemas que plantean las IE sería la imposibilidad de verificación de la acuracidad de los valores de predicción, por lo que se aconseja disponer de información complementaria procedente, por ejemplo, de encuestas realizadas en otro momento o lugar. En cualquier caso, debe evitarse aceptar, sin más, cualquier resultado que sea técnicamente posible o lógicamente viable.
- **Los fenómenos de autocorrelación y heteroscedasticidad espacial son causa y efecto del sesgo de la agregación, por lo que deben ser tenidos en cuenta expresamente en toda IE, a través de instrumentos propios de la econometría espacial**. Ignorar la presencia de los efectos espaciales en un ejercicio de IE puede ocasionar importantes sesgos en los resultados.
- **Otro problema propio del proceso de predicción-extrapolación espacial sería la falta de correspondencia que puede producirse entre teorías, que siendo válidas en un ámbito agregado, pudieran no serlo en ámbitos microterritoriales**. Por eso, la bondad de una determinada relación económica, en cuanto destinada a ser inferida a un ámbito microterritorial, no siempre podrá ser evaluada a partir de los estadísticos básicos, como el coeficiente de correlación lineal, o ajustes de bondad como el coeficiente de determinación (R^2), pues éstos experimentan aumentos de tamaño con aumentos de escala.
- **Las soluciones que podrían proponerse a estos problemas son únicamente parciales**, en el sentido de que ayudan a paliar las consecuencias que el sesgo de la agregación produce en los fenómenos económicos y, por tanto, en cualquier ejercicio de predicción-extrapolación de datos microterritoriales a partir de inferencias ecológicas. Por ejemplo, se recomienda el contraste de resultados con estudios similares, búsqueda de información muestral complementaria sobre las unidades microterritoriales (como

ayuda para la verificación de relaciones económicas en este ámbito), elección cuidadosa de las variables independientes, aplicación de técnicas de AEDE sobre las variables del modelo (tanto en el ámbito agregado como microterritorial) y la especificación de modelos de econometría espacial.

- **Un caso particular de predicción-extrapolación espacial son las llamadas “estimaciones indirectas” de la renta familiar disponible de municipios y comarcas, en España.** En la mayor parte de estos trabajos, el sesgo de la agregación y los efectos espaciales han sido largamente ignorados, habiéndose concentrado los esfuerzos en la obtención de variables explicativas, disponibles en el ámbito microterritorial, así como en la formulación, más o menos compleja, de modelos econométricos que son trasladados directamente del campo temporal al espacial.
- **En España, el origen de la predicción-extrapolación microterritorial se encuentra en los trabajos de estimación de la renta familiar disponible publicados, con carácter periódico, en el Anuario del Mercado Español (AME), patrocinado por Banesto.** A partir de los datos de renta familiar disponible provincial del Banco de Bilbao y un grupo de variables explicativas, disponibles también para el ámbito municipal de España, se realizaba una inferencia ecológica que suponía la constancia paramétrica en todo el territorio nacional, ignorando posibles defectos de especificación por la presencia de dependencia espacial en el modelo.
- Dado que gran parte de la información municipal incluida en el AME estaba relacionada con la actividad económica (líneas telefónicas, parque de vehículos, oficinas de crédito, licencias comerciales, etc.), **los autores de esta obra, así como muchos otros, han aplicado, también al campo microterritorial, la teoría macroeconómica que relaciona la renta de las familias con la generación de rentas o producción económica.** Esta aplicación ha producido, a veces, sesgos que favorecen a municipios de tamaño mediano, que destacan por la presencia en los mismos de polígonos industriales o segunda vivienda, y perjudican a localidades residenciales en las que reside la población de alto “standing”.
- **El Instituto Lawrence R. Klein (Universidad Autónoma de Madrid) realiza, desde 1992, la predicción-extrapolación de los datos de renta bruta disponible municipal de España** que, agrupados en una escala del 1 al 10 y para los municipios de mil y más habitantes, publica con el patrocinio de “la Caixa” en el Anuario Económico de España. La metodología empleada ha ido experimentando importantes cambios, en un esfuerzo por superar los problemas que plantea el sesgo de la agregación y la incorporación de los efectos espaciales con técnicas de econometría espacial. Esta tesis doctoral supone (así lo esperamos) un avance en la consolidación de esta metodología, así como una propuesta para cualquier investigador interesado en este tema de la predicción-extrapolación de datos microterritoriales.
- **Todo ejercicio de predicción-extrapolación de datos microterritoriales es un proceso que debe constar de las siguientes 6 etapas:** fundamentos teóricos, selección y tratamiento de la información disponible, análisis exploratorio espacial, análisis

confirmatorio espacial, predicción-extrapolación de datos microterritoriales y análisis de resultados.

- **Los resultados obtenidos para la RFD per capita de los municipios de la CM, en 2000, siguen un esquema de heterogeneidad espacial del tipo Centro-Periferia**, registrándose los mayores valores (más de 14.000 € unos 2,3 millones de pesetas) en Madrid capital, Corona Metropolitanas Oeste y Norte, y parte de la Sierra Central y Henares. En el otro extremo, los municipios situados en las “esquinas” de la CM (Sierra Norte, Suroeste y Sureste) son los que registran menores valores de RFD per capita, con menos de 7.000 € (1,2 millones de pesetas).

VI.2. LÍNEAS FUTURAS DE INVESTIGACIÓN

Esperamos que esta obra, así como el esfuerzo de investigación realizado, constituya un impulso que permita difundir más extensamente las técnicas de econometría espacial en nuestro país, tanto en el ámbito de los estudios universitarios como fuera de él. Pese a los muchos avances de la ciencia en este sentido, quedan aún aspectos pendientes, que podrían constituir líneas de investigación en el futuro, de las que se indican algunas a continuación. Con ello, ponemos punto final a la obra.

- Realización de un mayor esfuerzo de **trabajo interdisciplinar** entre investigadores de datos espaciales procedentes de distintos campos de las ciencias sociales y técnicas, como economistas, geógrafos, sociólogos, políticos, informáticos, biólogos, médicos, geólogos, físicos, etc.
- Desarrollo de **modelos de predicción-interpolación espacial**, que permitan la integración de datos procedentes de distintos dominios (estadísticas sociales, datos físicos o geológicos) a través de especificaciones que traten correctamente la presencia de los efectos espaciales de autocorrelación y heterogeneidad espacial.
- **Desarrollo de métodos de validación de resultados en el ejercicio de la predicción-extrapolación de datos microterritoriales**, que reduzcan la cierta carga de subjetividad que, en algunos aspectos, padece esta disciplina.
- **Desarrollo de tests estadísticos** que permitan discernir entre especificaciones de tipo autorregresivo o de medias móviles en la estructura de dependencia espacial propia de las variables de un modelo de regresión.
- Especificación de **métodos estadísticos de muestreo espacial** que tengan en cuenta la presencia del fenómeno de autocorrelación espacial en muchas variables geográficas; por ejemplo, mediante una selección de observaciones “suficientemente alejadas” unas de otras (o incorrelacionadas), que sean consideradas como representativas de una población.

- **Consolidación de la disciplina del geomarketing** incorporando a la misma los conceptos y herramientas propios de los análisis espacial exploratorio (AEDE) y confirmatorio de la econometría espacial.
- **Implementación de nuevas especificaciones de modelos y contrastes sobre los efectos espaciales en paquetes informáticos** capaces de interaccionar, de forma dinámica, en un entorno GIS “amigable”.
- **Nuevas especificaciones de las matrices de pesos espaciales en el campo de la economía regional**, que tengan en cuenta conceptos de distancia que superen la tradicional separación física entre regiones para considerar cuestiones, más propias del mundo desarrollado, como la dotación en infraestructuras de comunicaciones.
- **Análisis de nuevas formas de econometría espacial** en el campo de los modelos de ecuaciones espaciales simultáneas, especificaciones no lineales, modelos “borrosos” de interacción espacial, modelos espaciales de elección discreta, inferencias bayesianas espaciales, modelos espaciales de cointegración de series, etc.

Anexos

ANEXO I. APLICACIONES DE MODELOS ESPACIALES DE REGRESIÓN LINEAL

Contenido del Anexo I

- A.I.1. Proceso a seguir en un ejercicio básico de modelización en econometría espacial
- A.I.2. Modelos de dependencia y/o heterogeneidad espacial

A.I.1. Proceso a seguir en un ejercicio básico de modelización en econometría espacial

En este punto, se presentan dos ejemplos del proceso a seguir en la modelización econométrica espacial, partiendo del modelo básico de regresión lineal sobre el que se contrasta la presencia de los efectos espaciales de dependencia y heterogeneidad espacial. Se trata, en concreto, de un modelo propuesto por Anselin (1988A) para una función de delitos comunes, y de la función clásica de convergencia- β , sobre la que Moreno y Vayá (2000) contrastan los efectos espaciales.

➤ Anselin (1988A)

El autor presenta un ejemplo del proceso a seguir en cualquier ejercicio básico de modelización en econometría espacial, por lo que advierte de la sencillez de algunas de las especificaciones dado que fueron propuestas únicamente con fines ilustrativos. Para ello, utiliza información correspondiente a los 49 barrios de Columbus (Ohio).

Como primera etapa a cubrir, se propone la estimación del modelo básico de regresión lineal por el método de mínimos cuadrados ordinarios (MCO), para el estudio del número de delitos comunes en los barrios de Columbus, del modo siguiente:

$$\begin{aligned} Crime_i &= \beta_0 + \beta_1 Inc_i + \beta_2 House_i + u_i \\ u &\approx N(0, \sigma^2) \end{aligned} \quad A.VII.1.$$

donde Crime: tasa de delitos comunes (robos en viviendas y automóviles) por mil habitantes, correspondiente al barrio i , para $i = 1, 2, \dots, 49$.
Inc: renta disponible de los hogares por habitante
House: precio de la vivienda

Aunque los coeficientes estimados eran significativos y de signo negativo para las dos variables exógenas, como era de esperar, los contrastes de dependencia y heterogeneidad espacial (basados en el multiplicador de Lagrange) pusieron de manifiesto la existencia de una mala especificación en el modelo, sobre todo, debido a la presencia de una fuerte autocorrelación espacial tanto en la variable endógena como en la perturbación aleatoria.

Por eso, se propone, en primer lugar, la estimación de un **modelo mixto autorregresivo de regresión espacial**, o modelo del retardo espacial, por el método de máxima-verosimilitud (MV), en el que se incluye la variable endógena espacialmente retardada como explicativa en el modelo:

$$\begin{aligned} Crime_i &= \beta_0 + \rho \cdot WCrime_i + \beta_1 Inc_i + \beta_2 House_i + u_i \\ u &\approx N(0, \sigma^2) \end{aligned} \quad A.VII.2.$$

donde WCrime: tasa de delitos comunes espacialmente retardada a partir de una matriz W de interacciones espaciales, obtenida como contigüidad de primer orden entre los barrios de Columbus y estandarizada por filas.
 ρ : coeficiente autorregresivo espacial.

En este caso también todos los coeficientes estimados fueron muy significativos. Por otro lado, los contrastes específicos demostraron asimismo que la consideración del retardo espacial de la variable endógena en el modelo anulaba la presencia de autocorrelación espacial en la perturbación aleatoria.

Pese a todo, dado que en el modelo A.VII.1. se puso de manifiesto la potencial existencia de dependencia espacial en la perturbación aleatoria, el autor propone la estimación de dos modelos adecuados para este problema: el modelo del error espacial y el modelo Durbin espacial.

En primer lugar, se estima, por el método MV, un **modelo de regresión con perturbaciones autorregresivas espaciales**, o **modelo del error espacial** del modo siguiente:

$$\begin{aligned} Crime_i &= \beta_0 + \beta_1 Inc_i + \beta_2 House_i + u_i \\ u_i &= \lambda Wu_i + \varepsilon_i \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.3.$$

donde λ es el parámetro autorregresivo asociado al retardo espacial de la perturbación aleatoria del modelo (u), obtenido a partir de la misma matriz de interacciones espaciales utilizada en el modelo anterior, W. La fuerte significatividad de este parámetro, así como los resultados arrojados por el

test de Breush-Pagan, pusieron de manifiesto la fuerte presencia en la perturbación aleatoria de dependencia y heteroscedasticidad espacial, respectivamente.

Además, el autor estima un **modelo Durbin espacial**, también adecuado en modelos con problemas de dependencia espacial en la perturbación aleatoria, con la siguiente especificación:

$$\begin{aligned} Crime_i &= \beta_0 + \rho \cdot WCrime_i + \beta_1 Inc_i + \beta_2 House_i + \beta_3 WInc_i + \beta_4 WHouse_i + u_i \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.4.$$

En este modelo, que constituye también lo que se llama el contraste COMFAC, el coeficiente autorregresivo ρ volvió a resultar estadísticamente significativo, pero no sucedió lo mismo con los coeficientes β_3, β_4 de las variables exógenas espacialmente retardadas, siendo incluso incorrecto el signo negativo obtenido por el estimador de la variable $WInc$ (según la denominada hipótesis del factor común, el coeficiente $\beta_3 = \rho \cdot \beta_1$, por lo que el signo esperado para β_3 era positivo). Estos resultados pusieron de manifiesto, con mayor claridad, que la causa principal de dependencia espacial en el modelo se encontraba, no tanto en la perturbación aleatoria, cuanto en la omisión de la variable endógena espacialmente retardada como explicativa en el mismo. Por ese motivo, se propone la especificación del modelo del retardo espacial (modelo mixto autorregresivo de regresión espacial) como óptima.

Sin embargo, la detección del efecto de heterogeneidad espacial en forma de heteroscedasticidad en el modelo (indicada por el test de Breusch-Pagan, en la expresión *A.VII.3.*), aconsejaba la estimación de un modelo de heterogeneidad espacial que tuviera en cuenta explícitamente este problema. Para ello, se proponen dos especificaciones: el modelo de expansión espacial y el modelo de coeficientes aleatorios.

La especificación escogida para el **modelo de expansión espacial** fue la más sencilla, dado que el objetivo de la aplicación, como ya se indicó al principio, era la simple ilustración. Para ello, en el modelo se consideraron cuatro nuevas variables explicativas (variables de expansión), que se contruyeron como productos de las exógenas originales (Inc y $House$) por las coordenadas terrestres de longitud y latitud correspondientes a los centroides de los barrios de Columbus:

$$\begin{aligned} Crime_i &= \beta_0 + \beta_1 Inc_i + \beta_2 House_i + \beta_3 m \cdot Inc_i + \beta_4 p \cdot Inc_i + \\ &\quad + \beta_5 m \cdot House_i + \beta_6 p \cdot House_i + u_i \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.5.$$

donde m : coordenada terrestre de longitud (tendencia Este-Oeste)
 p : coordenada terrestre de latitud (tendencia Norte-Sur)

La introducción de cuatro variables explicativas más no afectó en gran medida al ajuste MCO del modelo (en términos del coeficiente de determinación R^2), aunque los estimadores de los parámetros (signos y valores) y su significatividad estadística sí sufrieron cambios importantes. Así, mientras los coeficientes correspondientes a las variables de longitud (m) fueron muy significativos, no sucedió lo mismo ni con las variables de latitud (p) ni con el coeficiente de la variable no expandida del precio de la vivienda ($House$), que incluso obtuvo un signo incorrecto. Por otro lado, la variable no expandida de renta (Inc) fue también mucho menos significativa que en estimaciones anteriores.

Dado que el problema de heteroscedasticidad en un modelo puede estar ocasionado por errores en la construcción de las variables de expansión, se llevaron a cabo también tres estimaciones robustas de los estimadores MCO utilizando los métodos de White, White adaptado (a pequeñas muestras) y el método de la navaja (“jackknife”). Los resultados obtenidos para los coeficientes por cada uno de estos métodos fueron contradictorios. Además, los estadísticos del multiplicador de Lagrange pusieron también de manifiesto la presencia en este modelo de dependencia espacial, tanto en la variable endógena como en la perturbación aleatoria.

Por eso, el autor opta por abordar el citado problema de heteroscedasticidad utilizando la especificación de un **modelo de coeficientes aleatorios** (que es un caso particular del modelo del error heteroscedástico) del modo siguiente:

$$\begin{aligned} Crime_i &= \beta_0 + \beta_1 Inc_i + \beta_2 House_i + u_i \\ Var(u_i) &= \gamma_0 + \gamma_1 Inc_i^2 + \gamma_2 House_i^2 \\ E[uu'] &= \Omega \end{aligned} \quad A.VII.6.$$

donde la varianza no constante de la perturbación aleatoria se distribuye según el esquema común de heteroscedasticidad aditiva (el propio del test de Breush y Pagan), siendo función lineal del cuadrado de las variables explicativas del modelo, a través de los coeficientes $\gamma_0, \gamma_1, \gamma_2$. En este modelo, la matriz Ω de covarianzas de la perturbación aleatoria es diagonal.

Los coeficientes aleatorios de la varianza de la perturbación aleatoria pueden ser obtenidos en la estimación del modelo por el método MV. En este caso, el coeficiente correspondiente a la renta (Inc), resultó muy significativo, pero fue desechado al obtener un signo negativo, incompatible con el modelo de coeficientes aleatorios. Por este motivo, se estimó también un **modelo del error heteroscedástico**, similar al anterior, en el que se consideraba el precio de la vivienda (House) como única variable causante de heteroscedasticidad, de forma que $Var(u) = \gamma_0 + \gamma_2 House^2$. En este modelo, el coeficiente aleatorio correspondiente a la variable precio de la vivienda (House) fue escasamente significativo.

Por último, los contrastes de dependencia espacial en ambas especificaciones ponían de manifiesto la presencia persistente de este efecto en el modelo.

Otra forma de abordar el problema de heterogeneidad en un modelo es a través de la inestabilidad o cambio estructural espacial, mediante un **modelo de estructuras espaciales (“spatial regimes”)** que, en el campo temporal, se corresponde con el modelo de regresiones cambiantes o “switching regressions”. De este modo, el autor propone la consideración de dos regresiones distintas para los barrios del este y oeste de la ciudad de Columbus, apoyado en la significatividad de las variables correspondientes a la coordenada de longitud (X) en el modelo de expansión espacial (ver expresión A.VII.5.). De este modo, se especificó un modelo sin restricciones, con coeficientes diferentes para cada zona de la ciudad (este/oeste), en el que se consideró también la existencia de dependencia espacial en la perturbación aleatoria. Es decir, se trataba de la combinación de un modelo del error espacial con un modelo de estructuras espaciales, del modo siguiente:

$$\begin{aligned} Crime_{iE} &= \beta_{0E} + \beta_{1E} Inc_{iE} + \beta_{2E} House_{iE} + u_i \quad ; \quad \text{para E=Este} \\ Crime_{iO} &= \beta_{0O} + \beta_{1O} Inc_{iO} + \beta_{2O} House_{iO} + u_i \quad ; \quad \text{para O=Oeste} \end{aligned} \quad A.VII.7.$$

$$u_i = \lambda W u_i + \varepsilon_i$$

$$\varepsilon \approx N(0, \sigma^2 I)$$

La estimación de este modelo por el método MV dio lugar a resultados muy diferentes de los obtenidos en el modelo restringido (con coeficientes únicos para cada variable exógena) del error espacial. Efectivamente, mientras que en el modelo del error espacial (A.VII.5.) las dos variables exógenas (Inc, House) resultaron claramente significativas, en el modelo actual (A.VII.7.), la variable de renta per capita (Inc) sólo resulta significativa en la ecuación correspondiente a los barrios del este de la ciudad (E), mientras que la variable precio de la vivienda (House) lo es sólo para la ecuación correspondiente a los barrios del oeste (O), siendo además los coeficientes no significativos (β_{1O} , β_{2E}) de signo positivo (incorrecto). Además, el coeficiente autorregresivo λ , de la perturbación aleatoria, resultó en el modelo sin restricciones muy significativo, y mayor, que en el modelo restringido.

Por otro lado, aunque el test de Chow espacial (que sólo puede obtenerse en modelos de estructuras espaciales, sin término autorregresivo en la perturbación, estimados por el método MCO) aceptó la hipótesis nula de ausencia de cambio estructural, los contrastes espaciales asintóticos rechazaron la hipótesis nula, proporcionando evidencias de inestabilidad estructural en los parámetros del modelo.

Aunque la especificación más adecuada parecía ser la proporcionada por el modelo de estructuras espaciales y error espacial (A.VII.7.), se propuso como última expresión a contrastar, el **modelo general de regresión espacial** sin el término del retardo espacial, que considera la existencia de ambos efectos de dependencia y heterogeneidad espacial en la perturbación aleatoria. Es decir, se trataría de la combinación del modelo del error espacial (A.VII.3.) y el modelo (A.VII.6.) del error heteroscedástico, del que se excluyó la variable de renta per capita (Inc), como explicativa de la varianza de la perturbación aleatoria, por su valor negativo (incompatible).

$$Crime_i = \beta_0 + \beta_1 Inc_i + \beta_2 House_i + u_i$$

$$u_i = \lambda W u_i + \varepsilon_i \quad A.VII.8.$$

$$\varepsilon \approx N(0, \Omega) \quad ; \quad \sigma_\varepsilon^2 = \gamma_0 + \gamma_2 House_i^2$$

La estimación por el método MV produjo unos coeficientes muy significativos, tanto para las variables exógenas del modelo como para el término autorregresivo de la perturbación aleatoria, aunque no para el coeficiente aleatorio de la varianza del error, asociado a la variable House (precio de la vivienda). Por ese motivo, se rechazó también esta última especificación en favor de la anterior.

➤ **Moreno y Vayá (2000)**⁷⁶

Las autoras proponen la estimación de la ecuación de convergencia- β tradicional, para una muestra de 108 regiones europeas, según la cual el crecimiento del VAB por ocupado de una región, durante

⁷⁶ Ver también otros estudios previos sobre el tema, por parte de las autoras, en Vayá (1996), Vayá y Moreno (2000), así como el resumen realizado en López-Bazo *et al.* (2001). También han estimado modelos similares para las regiones europeas Baumont *et al.* (2000) y Toral (2000), para los estados brasileños, Magalhães *et al.* (2000) y para las provincias españolas, Chasco y Vicéns (2000), todos ellos basados en un famoso artículo escrito por Rey y Montouri (1999), en el que se analiza el proceso de convergencia entre los estados de EEUU desde una perspectiva de econometría espacial.

un período dado, está en función del nivel de VAB por ocupado en el período inicial considerado, del modo siguiente:

$$gy_i = a - (1 - e^{-\beta t}) \ln y_i + u_i$$

$$u \approx N(0, \sigma^2)$$

A.VII.9.

donde gy : tasa de crecimiento del VAB por ocupado entre los años 1975 y 1992, de la región i
 $\ln y$: logaritmo neperiano del VAB por ocupado en el período inicial (1975)
 a : coeficiente que recoge el nivel del estado estacionario o nivel de equilibrio a largo plazo del VAB por ocupado
 β : coeficiente que refleja la velocidad de convergencia hacia dicho estado estacionario,
 de forma que la hipótesis de convergencia se aceptará siempre que β sea positivo y significativo
 u : perturbación aleatoria.

Al no disponer de un modelo teórico previo que incluya la existencia de un esquema de dependencia espacial en este modelo, se propone la estrategia modelizadora de Folmer y Florax (1992) denominada como EEV2 (expansión espacial de variables 2), con el objetivo de contrastar la posibilidad de errores de especificación en el modelo clásico de convergencia- β (*A.VII.9.*), por la existencia en el mismo del efecto de autocorrelación espacial. Además, las autoras llevaron a cabo un análisis exploratorio (AEDE) previo de la variable endógena (crecimiento del VAB por ocupado en 1975-1992) y la exógena (logaritmo neperiano del VAB por ocupado en 1975), en el que se ponía de manifiesto claramente la existencia de un esquema de autocorrelación espacial positiva en ambas variables, así como problemas de heterogeneidad espacial, debido a las diferencias mostradas por dichas variables en las regiones centrales de Europa y el resto de regiones.

En primer lugar, la estimación del modelo por el método MCO produce un coeficiente significativo β del 2,7%, aunque los contrastes LM (multiplicador de Lagrange) obligan a rechazar la hipótesis nula de ausencia de dependencia espacial en el modelo, poniéndose de manifiesto así las relaciones de interdependencia regionales ya constatadas en el AEDE. En concreto, el mayor valor del test LM-ERR sobre el test LM-LAG sugiere la reespecificación del modelo incluyendo un término autorregresivo en la perturbación aleatoria. El mapa de los residuos de la estimación MCO, por su parte, también pone de manifiesto la existencia, en esta variable, de una continuidad geográfica.

Por eso, se estima por MV el siguiente modelo del error espacial:

$$gy_i = a - (1 - e^{-\beta t}) \ln y_i + u_i$$

$$u_i = \lambda W u_i + \varepsilon_i$$

$$\varepsilon \approx N(0, \sigma^2)$$

A.VII.10.

donde λ : parámetro autorregresivo espacial que mide la intensidad de la interdependencia entre los residuos.
 W : matriz de interacciones espaciales, definida como contigüidad de primer orden entre regiones, estandarizada por filas.

En este modelo, los dos coeficientes resultaron significativos, incluido el parámetro autorregresivo λ , siendo el valor del parámetro de convergencia, β (ligeramente inferior al estimador MCO) del

2,5%. Esta especificación, en principio correcta, fue también ratificada a través del contraste COMFAC o de la estimación del modelo Durbin espacial siguiente:

$$\begin{aligned} gy_i &= a - (1 - e^{-\beta T}) \ln y_i + \lambda_1 W gy_i + \lambda_2 W \ln y_i + u_i \\ u &\approx N(0, \sigma^2) \end{aligned} \quad A.VII.11.$$

para el que se cumple la restricción del factor común: $H_0(\beta \cdot \lambda_1 = \lambda_2)$. Por este motivo, se llegó a la conclusión de que la especificación más adecuada del modelo de convergencia- β podría ser el modelo del error espacial. Pese a confirmarse la superioridad de esta especificación, las autoras estimaron también un modelo del retardo espacial (por MV) que, pese a producir estimadores significativos, obtuvo peores resultados en términos de ajuste (medido a través del logaritmo de verosimilitud), por lo que fue rechazado.

Por otro lado, el problema de heterogeneidad espacial detectado en las variables del modelo por el AEDE, resultó confirmado en el modelo del error espacial seleccionado (A.VII.10), en forma de heteroscedasticidad, a través del contraste de Breush-Pagan. Por este motivo, se estimó también un modelo de expansión espacial, en el que las variables de expansión fueron las coordenadas terrestres de latitud o tendencia norte-sur (p) y longitud o tendencia este-oeste (m), obtenidas para cada una de las regiones europeas consideradas en la muestra. Para ello, se transformó el modelo inicial:

$$\begin{aligned} gy_i &= a + b \ln y_i + u_i \\ u &\approx N(0, \sigma^2) \end{aligned} \quad A.VII.12.$$

donde $b = -(1 - e^{-\beta T})$ y se definieron las variables de expansión tanto para el parámetro (b) de la variable exógena como para el término independiente (a), de forma que el modelo de expansión espacial definido fue el siguiente:

$$\begin{aligned} gy_i &= a_0 + a_1 m_i + a_2 p_i + b_0 \ln y_i + b_1 m_i \ln y_i + b_2 p_i \ln y_i + u_i \\ u &\approx N(0, \sigma^2) \end{aligned} \quad A.VII.13.$$

para
$$\begin{aligned} a_i &= a_0 + a_1 m_i + a_2 p_i \\ b_i &= b_0 + b_1 m_i + b_2 p_i \end{aligned}$$

En este modelo, que fue estimado por el método MCO, el valor de los parámetros (a_i , b_i) varía para cada región i de la muestra pudiéndose representar su distribución en sendos mapas donde quedaba de manifiesto la continuidad geográfica de ambos parámetros. De un lado, el estado estacionario (a) al que tienden las regiones europeas es cada vez menor a medida que se produce un desplazamiento en dirección al sudeste y, de otro, la velocidad de convergencia (b) sigue un proceso contrario, pues parece reducirse en dirección noroeste.

Pese a la alta significatividad de los coeficientes estimados en este modelo, los contrastes de dependencia espacial pusieron de manifiesto la presencia de este efecto, sustantivo y/o residual. Por otro lado, la hipótesis de una expansión lineal en la inestabilidad del proceso de convergencia parece simplista y probablemente equivocada.

Por todo esto, dado que ya se había contrastado la presencia de dependencia espacial en el término de error, se reestimó el modelo anterior incluyendo en el mismo una estructura autorregresiva de primer orden en la perturbación aleatoria, obteniéndose así un modelo de expansión espacial con perturbaciones aleatorias autorregresivas espaciales (modelo del error espacial):

$$\begin{aligned} gy_i &= a_0 + a_1 m_i + a_2 p_i + b_0 \ln y_i + b_1 m_i \ln y_i + b_2 p_i \ln y_i + u_i \\ u_i &= \lambda W u_i + \varepsilon_i \\ \varepsilon &\approx N(0, \sigma^2) \end{aligned} \quad A.VII.14.$$

Este modelo corrige el problema de dependencia espacial residual, dada la alta significatividad del coeficiente λ autorregresivo y el rechazo de dependencia sustantiva. Pero, los resultados obtenidos para los coeficientes a_2 y b_2 , correspondientes a la tendencia norte-sur, no fueron positivos, por lo que las autoras concluyen con el propósito de buscar, en un futuro, una forma más precisa de especificación de la variación espacial existente en la ecuación de convergencia- β que corrija los problemas de heterogeneidad espacial.

A.I.2. Modelos de dependencia y/o heterogeneidad espacial

➤ **Bloomstein (1985)**

Propone este autor un modelo Durbin espacial o modelo mixto autorregresivo regresivo cruzado de segundo orden, SAR(2), del VAB agrícola en las 26 provincias irlandesas.

$$\begin{aligned} y_i &= \delta_0 t + \delta_1 X_{1i} + \delta_2 W^1 X_{1i} + \delta_3 W^2 X_{1i} + \delta_4 W^1 y_i + \delta_5 W^2 y_i + u_i \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.15.$$

donde y_i : vector (N,1) del porcentaje de variación del VAB agrícola, para $i = 1, 2, \dots, N=26$
 X_{1i} : vector (N,1) del índice de accesibilidad por carretera en cada provincia
 W^1, W^2 : matrices de pesos espaciales de primer y segundo orden, respectivamente
 δ : parámetros autorregresivos (excepto δ_0, δ_1).

El esfuerzo del autor se centra en obtener una matriz de pesos espaciales de segundo orden, W^2 , sin problemas de circularidad o redundancia.

➤ **Basu y Reinsel (1994)**

Los autores proponen la estimación del modelo de producción de cebada, presentado por Student como ejemplo, especificado como un modelo ARMA(1,1) espacial o modelo de regresión con perturbación aleatoria autorregresiva y de medias móviles de primer orden. Las observaciones proceden de 160 parcelas, de aproximadamente 4 pies cuadrados, situadas en una cuadrícula de 5 x 32, y están referidas a un experimento realizado en el año 1913. El modelo fue estimado, en esta ocasión, por el método de MCG, con la siguiente especificación:

$$y_i = \beta_0 + \sum_{k=1}^7 \beta_k x_{ik} + u_i$$

$$u = \phi Wu + \varepsilon + \theta W\varepsilon$$

$$\varepsilon \approx N(0, \sigma^2 I)$$

A.VII.16.

donde y_i : cosecha de cebada en la parcela i
 k : subíndice que indica las 7 clases de cebada cultivadas en diversas parcelas
 x_{ik} : variable ficticia, que toma el valor 1 en la variable x correspondiente al tipo de cebada cultivada en la parcela i considerada
 W : matriz de interacciones (criterio de la torre).

➤ Haining (1994)

Se trata de un modelo del error espacial, SAR(1), también temporalmente autorregresivo, AR(1), de los precios de la gasolina en 61 gasolineras de la zona suroeste de Sheffield. Según el autor, el precio de la gasolina presenta autocorrelación temporal, pues las gasolineras que la venden más cara un mes tenderán a seguir vendiéndola más cara en los meses siguientes, así como espacial, dado que las gasolineras situadas en la misma zona suelen vender la gasolina a precios similares. Es decir, el modelo tendría la siguiente especificación:

$$y_{mar,i} = \beta_0 + \beta_1 y_{feb,i} + u_i$$

$$u_i = \lambda Wu_i + \varepsilon_i$$

$$\varepsilon \approx N(0, \sigma^2 I)$$

A.VII.17.

donde $y_{mar,i}$: precio medio de la gasolina en el mes de marzo, en la gasolinera i
 $y_{feb,i}$: precio medio de la gasolina en el mes de febrero, en la gasolinera i
 β_0, β_1 : parámetros correspondientes al término independiente y variable exógena (que es la variable endógena temporalmente desplazada un mes).
 λ : parámetro autorregresivo espacial correspondiente a la perturbación aleatoria.

La estimación del modelo por el método MV resultó significativa en los parámetros y aceptable en términos de ajuste, cuando se depuraron de la muestra ciertos valores atípicos (normalmente correspondientes a gasolineras situadas en lugares aislados) cuya influencia sobre los contrastes de dependencia espacial suele ser importante.

➤ Bolduc, Laferrière y Santarossa (1995)

Se propone un modelo de componentes del error con perturbación aleatoria espacial autorregresiva de primer orden o MCE-SAR(1) para la explicación de los desplazamientos, por motivo de trabajo, de una muestra del 20% de los residentes en Winnipeg (Canadá), en las primeras horas punta de los días del verano de 1976. La información se agregó a 147 zonas de la ciudad y la estimación se realizó con un total de 211 observaciones (flujos origen-destino) para el siguiente modelo:

$$\ln\left(\frac{p_B}{p_A}\right)_{ij} = \beta_0 + \beta_1 TB_{ij} + \beta_2 TA_{ij} + \beta_3 CA_{ij} + \beta_4 Rpc_i + \beta_5 A_i + \beta_6 H_{ij} + u_{ij}$$

$$u_{ij} = \alpha_i + \lambda_j + \varepsilon_{ij}$$

$$\alpha_i = \rho_1 W \alpha_i + \xi_1 \quad ; \quad \xi_1 \approx N(0, \sigma_{\xi_1}^2)$$

$$\lambda_j = \rho_2 W \lambda_j + \xi_2 \quad ; \quad \xi_2 \approx N(0, \sigma_{\xi_2}^2)$$

$$\varepsilon_{ij} = \rho_3 W \varepsilon_{ij} + v_{ij} \quad ; \quad v \approx N(0, \sigma_v^2)$$

A.VII.18.

donde p_B, p_A : población que se desplaza al trabajo en autobús (B) y automóvil particular (A).
 TB_{ij} : tiempo de viaje (minutos) del trayecto en autobús entre i,j (suma de la espera, acceso, estancia en el vehículo y bajada del mismo)
 TA_{ij} : tiempo de viaje (minutos) del trayecto en automóvil entre i,j
 CA_{ij} : coste del desplazamiento en automóvil (incluye también el coste de aparcamiento)
 Rpc_i : renta per cápita de los hogares de la zona i de origen del trayecto
 A_i : tasa del parque de automóviles por hogar existente en el origen i
 H_{ij} : tasa de hombres varones que se desplazan entre i,j, dado que existe una mayor predisposición en los hombres a utilizar el coche para ir a trabajar.
 W : matriz de pesos espaciales en la que $w_{ij} = d_{ij}^{-\theta}$, siendo d_{ij} la distancia entre i,j.

El coste del desplazamiento en autobús, al ser el mismo para todos en la ciudad, está incluido en el término independiente de forma implícita, con un coeficiente que se supone será igual a β_3 .

La estimación MV del modelo produjo coeficientes significativos y de signo correcto. Así, por ejemplo, las variables relativas al origen (tasa de vehículos y la renta per cápita de los hogares del origen) fueron negativas o inversamente proporcionales a la tasa de utilización del autobús frente al coche en los desplazamientos por motivos de trabajo.

➤ Getis (1995)

Este autor cita dos estudios en los que la presencia contrastada de autocorrelación espacial en todas las variables (endógena y exógenas) del modelo lleva al autor a someterlas a un proceso de filtrado espacial que elimine de ellas dicho problema y permita la estimación de un MBRL por MCO.

En primer lugar, propone un modelo de disparidades regionales llevado a cabo en las 16 regiones de Turquía, en el que se especificaba el siguiente modelo de producción tipo Cobb-Douglas:

$$\log y_i^* = \beta_0 + \beta_1 \log P_i^* + \beta_2 \log I_i^* + \beta_3 \log S_i^* + \beta_4 \log A_i^* + u_i$$

$$u \approx N(0, \sigma^2 I)$$

A.VII.19.

donde y_i : VAB per capita de la región i
 P : tasa de población activa
 I : tasa de población no analfabeta
 S : tasa de ocupación
 A : tasa de concentración de la población en grandes núcleos de población

El asterisco (*) sobre las variables indica que éstas han sido previamente filtradas o corregidas del efecto de autocorrelación espacial mediante la siguiente transformación, por ejemplo, para y :

$$y_i^* = \frac{y_i \left(\frac{W_i}{N-1} \right)}{G_i(d)} \quad A.VII.20.$$

Como puede observarse, este método de filtrado espacial se basa en el estadístico $G_i(d)$, de la familia de estadísticos de autocorrelación local G de Getis (ver Apartado III.2.3.2.):

$$G_i(d) = \frac{\sum_{j=1}^N w_{ij}(d) y_j}{\sum_{j=1}^N y_j} \quad ; \quad \text{para } j \neq i \quad A.VII.21.$$

donde $[w_{ij}(d)]$: matriz de interacciones espaciales (0-1), tal que: $w_{ij} = 1$ para todos los vínculos existentes entre aquellos puntos j situados en un radio d del punto i y $w_{ij} = 0$ para el resto de vínculos, incluida la relación entre el punto i consigo mismo.

W_i : $W_i = \sum_j w_{ij}(d)$, por lo que el cociente $W_i/N-1$ es la media de G_i .

En otro estudio, el autor propone un modelo de gasto para los estados de EEUU en el año 1970:

$$PCEXP_i^* = \beta_0 + \beta_1 PCAID_i^* + \beta_2 POP_i^* + \beta_3 PCINC_i^* + u_i \quad A.VII.22.$$

$$u \approx N(0, \sigma^2 I)$$

donde $PCEXP_i$: gasto público total (nacional y local) por habitante, en el estado i
 $PCAID$: transferencias públicas realizadas por el estado federal a cada estado por habitante
 POP : población total
 $PCINC$: renta disponible per capita

Todas las variables fueron filtradas utilizando el estadístico $G_i(d)$ de autocorrelación local, de la misma manera que en el caso anterior, pudiéndose estimar por MCO el modelo final transformado.

➤ Haining (1995)

Se especifica un modelo del error espacial SAR(1) para las tasas de mortalidad por cáncer, con datos extraídos de las 87 áreas sanitarias de la ciudad de Glasgow, en función de un índice sintético de pobreza. Previamente, el autor demuestra que la estimación MCO del modelo básico de regresión es incorrecta debido a errores de especificación, por lo que se introduce una estructura autorregresiva de primer orden de dependencia espacial en las perturbaciones aleatorias. El modelo finalmente especificado fue, por tanto, el siguiente:

$$\begin{aligned}
 y_i &= \beta_0 + \beta_1 x_i + u_i \\
 u_i &= \lambda W u_i + \varepsilon_i \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{A.VII.23}$$

- donde y_i : tasa de mortalidad por cáncer en el área sanitaria i
 x : indicador sintético de pobreza obtenido a partir de las siguientes variables: tasa de hogares de inferior clase social, densidad de población, tasa de hogares sin automóvil, tasa de ocupación.
 W : matriz de pesos espaciales estandarizada por filas.

➤ **Fotheringham, Charlton y Brunson (1997)**

Los autores proponen un modelo de regresiones geográficamente ponderadas (RGP) aplicado a la variable LLTI ("Limiting Long-Term Illness"), que incluye varias enfermedades graves (males respiratorios, infartos, artritis graves, incapacidades físicas, etc.) que incapacitan para el trabajo laboral. Para ello, se ha tomado una muestra de 605 secciones censales correspondientes a cuatro condados nororientales de Inglaterra (Tyne and Wear, Durham, Cleveland y North Yorkshire), con datos procedentes del Censo de 1991.

$$LLTI_i = \beta_{0i} + \beta_{1i} UNEM_i + \beta_{2i} CROW_i + \beta_{3i} SPF_i + \beta_{4i} SC1_i + \beta_{5i} DENS_i + u_i \tag{A.VII.24}$$

- donde LLTI_{*i*}: porcentaje de individuos, de 45-65 años, residentes en un hogar en el que se ha registrado algún caso de enfermedades LLTI, en la sección censal i (en cada sección residen 200 hogares, por término medio)
 UNEM: tasa de paro
 CROW: tasa de hogares con más de una persona por habitación
 SC1: tasa de hogares cuyo cabeza de familia tiene una categoría profesional de 1 (puestos no directivos)
 DENS: densidad de población
 $\beta_{0i}, \dots, \beta_{5i}$: parámetros correspondientes a las variables explicativas, diferentes para cada sección censal i de la muestra, de forma que $\hat{\beta}_i = [X'W_i X]^{-1} X'W_i y$, siendo W_i una matriz diagonal de orden (n, n) , propia para cada sección i , siendo los elementos de la diagonal principal una función de la distancia del resto de observaciones a dicha sección i , en concreto: $w_{ij} = \exp(-\beta \cdot d_{ij}^2)$.

➤ **Pace y Barry (1997)**

Los autores diseñan un modelo mixto autorregresivo regresivo cruzado, que se estimará por MV, para el porcentaje de votos válidos obtenidos en las elecciones presidenciales norteamericanas del año 1980, en los 3.107 condados continentales de este país.

$$\begin{aligned} \ln(PVt)_i &= \beta_0 + \rho W \ln(PVt)_i + \beta_1 \ln(P18)_i + \beta_2 \ln(Ins)_i + \beta_3 \ln(Viv)_i + \\ &+ \beta_4 \ln(Rta)_i + \beta_5 W \ln(P18)_i + \beta_6 W \ln(Ins)_i + \beta_7 W \ln(Viv)_i + \\ &+ \beta_8 W \ln(Rta)_i + u_i \end{aligned} \quad A.VII.25.$$

$$u \approx N(0, \sigma^2 I)$$

donde $\ln(PVt)_i$: logaritmo neperiano del porcentaje de votos válidos por la población de 18 y más años, en el condado i

W : matriz de pesos espaciales función de la distancia entre los centroides de cada condado

$\ln(P18)$: logaritmo neperiano de la población de 18 y más años

$\ln(Ins)$: logaritmo neperiano de la población con alto nivel de instrucción (grado 12° o superior)

$\ln(Viv)$: logaritmo neperiano del número de viviendas en propiedad

$\ln(Rta)$: logaritmo neperiano del total de renta

Como puede observarse, el porcentaje de votos válidos obtenido en un determinado condado viene explicado, no sólo por factores internos (población mayor de edad, instrucción, etc.), sino también por el comportamiento de las variables (endógena y exógenas) en los condados más cercanos.

➤ Kristensen (1998)

Este autor especifica un modelo de heterogeneidad espacial multinivel para la variable precio del suelo urbano en la capital de Dinamarca, Copenhague (aunque también se analizan otras ciudades danesas de tamaños diversos, como Aarhus y Odense). Los datos, correspondientes al año 1977, se refieren a diversas localizaciones situadas en un radio de 5 km. del centro urbano, de forma que cada localización se encuentre, una de otra, a una distancia mínima de 250 m. Para ello, se especifica un modelo de expansión espacial en el que, como se verá a continuación, la perturbación aleatoria presenta un esquema de autocorrelación espacial que se explicitará más adelante.

$$LP_i = \beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \beta_3 d_i^3 + u_i \quad A.VII.26.$$

donde LP_i : logaritmo neperiano del precio del suelo urbano para la localización i de la ciudad de Copenhague (también se realizó para las ciudades de Aarhus y Odense).

d_i : distancia entre cada localización de la muestra, i , y el centro urbano

La expansión de los coeficientes β se produce a partir de las siguientes expresiones topológicas:

$$\begin{cases} \beta_0 = \gamma_{00} + \gamma_{10} \cos v_i + \gamma_{20} \sin v_i \\ \beta_1 = \gamma_{01} + \gamma_{11} \cos v_i + \gamma_{21} \sin v_i \\ \beta_2 = \gamma_{02} + \gamma_{12} \cos v_i + \gamma_{22} \sin v_i \\ \beta_3 = \gamma_{03} + \gamma_{13} \cos v_i + \gamma_{23} \sin v_i \end{cases} \quad A.VII.27.$$

donde $\cos v_i$: coseno del ángulo que forma la línea de longitud este-oeste que pasa por el centro urbano y la línea que une dicho centro urbano con el punto i

$\sin v_i$: seno del ángulo que forma la línea de longitud este-oeste que pasa por el centro

urbano y la línea que une dicho centro urbano con el punto i ; es decir, si un punto i se encuentra sobre la línea norte-sur, el $\cos v = 0$ y $\sin v = 1$.

Por tanto, el modelo expandido completo tendrá la forma siguiente:

$$\begin{aligned}
 LP = & (\gamma_{00} + \gamma_{10} \cos v_i + \gamma_{20} \sin v_i) + \\
 & + (\gamma_{01} + \gamma_{11} \cos v_i + \gamma_{21} \sin v_i) d_i + \\
 & + (\gamma_{02} + \gamma_{12} \cos v_i + \gamma_{22} \sin v_i) d_i^2 + \\
 & + (\gamma_{03} + \gamma_{13} \cos v_i + \gamma_{23} \sin v_i) d_i^3 + u_i
 \end{aligned}
 \tag{A.VII.28}$$

El autor propone la estimación de este modelo de expansión espacial por MCO de forma que se obtengan los residuos (e_i) del mismo que, con toda seguridad, afirma estarán espacialmente autocorrelacionados debido a la ausencia en el modelo de una variable explicativa referida al total del área urbana considerada. Esta variable omitida, en cada localización i , tiene un impacto sobre los precios de la vivienda de localizaciones vecinas y , por tanto, produce autocorrelación espacial, que se traslada al término de error. Por eso, se propone una especificación de la estructura de dependencia espacial de la perturbación aleatoria mediante la descomposición de los residuos obtenidos de esta primera estimación, por ejemplo, en cuatro componentes (que posteriormente serán reducidos a dos) de la manera siguiente:

$$e_i = \xi_1 EC_i + \xi_2 EP_i + \xi_3 ED_i + \xi_4 EI_i \tag{A.VII.29}$$

donde e_i : residuo obtenido de la estimación del modelo anterior en la localización i
 EC: residuo correspondiente a la localización vecina a una dada i sobre la línea radial en dirección al centro urbano
 EP: residuo de la localización vecina a una dada sobre la línea radial en dirección a la periferia
 ED: residuo de la localización vecina a una dada sobre el anillo de circunvalación en dirección a la derecha
 EI: residuo de la localización vecina a una dada sobre el anillo de circunvalación en dirección a la izquierda

Estos componentes del residuo podrían reducirse a dos considerando, por un lado, que EC y EP determinan la estructura de dependencia espacial producida sobre las líneas radiales que unen el centro con la periferia de las ciudades, mientras que ED y EI determinarían la dependencia sobre los anillos de circunvalación, de la siguiente manera:

$$e_i = \lambda_1 ECP_i + \lambda_2 EDI_i \tag{A.VII.30}$$

donde ECP: dirección radial centro-periferia definida como $ECP = (EC+EP)/2$
 EDI: dirección de circunvalación derecha-izquierda definida como $EDI = (ED+EI)/2$.

Por tanto, el modelo de expansión espacial con perturbación aleatoria espacialmente dependiente quedaría finalmente definido como una estructura multinivel, de la manera siguiente:

$$\tag{A.VII.31}$$

$$\begin{aligned}
LP = & (\gamma_{00} + \gamma_{10} \cos v_i + \gamma_{20} \sin v_i) + \\
& + (\gamma_{01} + \gamma_{11} \cos v_i + \gamma_{21} \sin v_i) d_i + \\
& + (\gamma_{02} + \gamma_{12} \cos v_i + \gamma_{22} \sin v_i) d_i^2 + \\
& + (\gamma_{03} + \gamma_{13} \cos v_i + \gamma_{23} \sin v_i) d_i^3 + \\
& + \left[\begin{aligned} & (\gamma_{00,cp} + \gamma_{10,cp} \cos v_i + \gamma_{20,cp} \sin v_i) + \\ & + (\gamma_{01,cp} + \gamma_{11,cp} \cos v_i + \gamma_{21,cp} \sin v_i) d_i + \\ & + (\gamma_{02,cp} + \gamma_{12,cp} \cos v_i + \gamma_{22,cp} \sin v_i) d_i^2 + \\ & + (\gamma_{03,cp} + \gamma_{13,cp} \cos v_i + \gamma_{23,cp} \sin v_i) d_i^3 \end{aligned} \right] * ECP_i + \\
& + \left[\begin{aligned} & (\gamma_{00,di} + \gamma_{10,di} \cos v_i + \gamma_{20,di} \sin v_i) + \\ & + (\gamma_{01,di} + \gamma_{11,di} \cos v_i + \gamma_{21,di} \sin v_i) d_i + \\ & + (\gamma_{02,di} + \gamma_{12,di} \cos v_i + \gamma_{22,di} \sin v_i) d_i^2 + \\ & + (\gamma_{03,di} + \gamma_{13,di} \cos v_i + \gamma_{23,di} \sin v_i) d_i^3 \end{aligned} \right] * EDI_i + \\
& + \varepsilon_i \\
\varepsilon \approx & N(0, \sigma^2 I)
\end{aligned}$$

Como puede observarse, en este modelo no se incluye ninguna variable explicativa concreta, dado que todas ellas se encontrarían representadas por la estructura de dependencia espacial (radial y de circunvalación) de los residuos, en la que se incluiría un grupo heterogéneo de instituciones locales tales como centros de negocios, estaciones de ferrocarril, parques, distritos urbanos, etc. Estas instituciones son las que suelen ejercer gran influencia sobre el precio del suelo urbano de localizaciones circundantes de una forma que, incluso para una misma institución, el precio de las viviendas circundantes puede depender de que se localicen junto a su fachada delantera o trasera. La estimación del modelo puede realizarse por MCO por contar, así especificado, con perturbaciones esféricas.

➤ **Stough, Lall y Trice (1998)**

Se plantea un modelo explicativo de la producción del sector tecnológico en los condados del área metropolitana de Washington DC en función de una serie de variables explicativas. En este modelo, que se estimó tanto para el año 1988 como para 1995, se detectó la presencia de autocorrelación espacial en tres de las variables (endógena y dos exógenas), por lo que previamente se les aplicó a estas tres variables el método de filtrado espacial G_i de Getis, para poder abordar la estimación de este modelo por MCO, como en el MBRL:

$$\begin{aligned}
y_i^* &= \beta_0 + \beta_1 GA_i + \beta_2 Rpc_i^* + \beta_3 DA_i^* + \beta_4 IU_i + u_i \\
u &\approx N(0, \sigma^2 I)
\end{aligned}
\tag{A.VII.32}$$

donde y_i : tasa de establecimientos del sector tecnológico (sobre el total de establecimientos de actividad económica) del condado i , variable “proxi” de la demanda industrial y del flujo de capital

- GA: gasto en autopistas por habitante
 Rpc: renta disponible per capita
 DA: densidad de autopistas
 IU: tasa de instrucción universitaria (sobre el total de la población adulta).

Las variables marcadas con el asterisco son las afectadas por el efecto de dependencia espacial, por lo que no se encuentran en su estado original, sino transformadas por el método de filtrado espacial G_i , del modo siguiente:

$$y_i^* = \frac{y_i \left(\frac{W_i}{N-1} \right)}{G_i(d)} \quad A.VII.33.$$

donde $W_i = \sum_j w_{ij}(d)$.

En esta aplicación, se tomó una matriz W binaria de interacciones espaciales de contigüidad de primer orden, para un radio de distancia (d) de 27 millas entre los centroides de los condados.

Los resultados pusieron de manifiesto, entre otras cosas, que el papel de ciertas variables explicativas, como la tasa de instrucción universitaria y la renta per capita, sobre la producción del sector tecnológico es menor en un modelo sin efectos espaciales, aunque el papel explicativo del espacio (medido como diferencia entre la varianza de los modelos con y sin efectos espaciales) es menor en 1995 que en 1988.

➤ Anselin, Varga y Acs (2000)

Los autores se proponen analizar las externalidades producidas por el conocimiento académico a través de la función de producción del conocimiento de Griliches-Jaffe (tipo Cobb-Douglas), tomando como unidad de investigación las áreas metropolitanas de EEUU, por ser la escala más adecuada para la medición de las economías de aglomeración. Para ello, se especifican varios modelos en los que se consideran los efectos de autocorrelación y heterogeneidad espacial, este último en forma de distintas estructuras o regímenes espaciales, que permitan verificar los conceptos teóricos sobre los que se basan las economías de aglomeración.

Las observaciones proceden de las 125 áreas metropolitanas de los EEUU en las que se produjeron innovaciones durante el año 1982 (fecha de referencia). Además, se consideraron 2 estructuras o regímenes, según que las áreas metropolitanas estuvieran o no "conectadas", es decir, a una distancia inferior o superior a 50 millas, respectivamente.

Las innovaciones acaecidas en el sector farmacéutico y químico fueron explicadas a través de un modelo básico de regresión de la función de producción, que se especifica como lineal mediante transformación logarítmica, en el que la única variable explicativa significativa fue el gasto de las industrias en I+D, descartándose cualquier tipo de externalidad en el mismo, siendo la muestra las 48 áreas metropolitanas en las que se produjeron innovaciones en este sector, durante 1982:

$$\ln K_i = \beta_0 + \beta_1 \ln R_i + u_i$$

$$u \approx N(0, \sigma^2 I) \quad A.VII.34.$$

- donde K_i : número de innovaciones registradas en el sector farmacéutico y químico en la base de datos de la Asociación de PYMES de los EEUU, como innovaciones introducidas en el mercado americano, en 1982.
- R_i : variable “proxy” de la actividad en I+D de las industrias: personal especializado ocupado en laboratorios industriales de I+D, en i .

Además, estimaron un modelo para las innovaciones del sector de la electrónica, estimado por MV, debido a la existencia de autocorrelación espacial en la perturbación aleatoria, aunque no tan importante como cabría pensar a priori:

$$\begin{aligned} \ln K_i &= \beta_0 + \beta_1 \ln R_i + \beta_2 \ln U_i + \beta_3 \ln O_i + \beta_4 \ln E_i + u_i \\ u_i &= \lambda W u_i + \varepsilon \\ \varepsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.35.$$

- donde K_i : número de innovaciones registradas en el sector de la electrónica, durante 1982, en 70 áreas metropolitanas
- R_i : variable “proxy” de la actividad en I+D de las industrias: personal especializado ocupado en laboratorios industriales de I+D, en i .
- U_i : gasto realizado en las universidades de i en actividades de investigación (dólares)
- O_i : variable “proxy” de la influencia que el entorno económico del área i ejerce sobre las innovaciones: ocupados en el sector de la electrónica en el área i
- E_i : variable “proxy” de la influencia que el entorno económico del área i ejerce sobre las innovaciones: ocupados en empresas del área i .
- W : matriz de pesos en la que $w_{ij} = 1$ para observaciones (áreas metropolitanas) situadas a una distancia igual o inferior a 50 millas (también se consideró para 75 millas).

Para el sector de maquinaria industrial, se estimó un modelo del retardo espacial (modelo mixto autorregresivo de regresión espacial), para las 89 áreas metropolitanas con innovaciones de este tipo, en 1982, en el que resultó asimismo significativa la consideración de 2 estructuras o regímenes espaciales (A y B), según que dichas áreas metropolitanas estuvieran o no “conectadas” (a una distancia menor o igual a 50 millas). Es decir, en este sector se produjeron externalidades en el conocimiento que trascendían la escala geográfica de las áreas metropolitanas, tal como era de esperar, aunque en contra de las expectativas iniciales, no resultó significativa la contribución de las actividades de investigación (I+D) realizadas por la industria y la universidad.

$$\begin{aligned} \ln K_i &= \rho^A W \ln K_i + \beta_0^A + \beta_1^A \ln O_i + \beta_2^A \ln E_i + \\ &\quad + \rho^B W \ln K_i + \beta_0^B + \beta_1^B \ln O_i + \beta_2^B \ln E_i + u_i \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.36.$$

- donde K_i : número de innovaciones registradas en el sector de maquinaria industrial, en 1982, en el área i
- O_i : variable “proxy” de la influencia que el entorno económico del área i ejerce sobre las innovaciones: ocupados en el sector de la electrónica en el área i
- E_i : variable “proxy” de la influencia que el entorno económico del área i ejerce sobre el resto
- W : matriz de pesos en la que $w_{ij} = 1$ para observaciones (áreas metropolitanas) situadas a una distancia igual o inferior a 50 millas (también se consideró para 75 millas).

Por último, para el sector de instrumental médico, se propone también un modelo de estructuras espaciales con ausencia de autocorrelación espacial, para 63 áreas metropolitanas con innovaciones en este sector. En este caso, las únicas variables significativas en ambas estructuras fueron actividad de I+D de las industrias y ocupados en empresas, ya que mientras que la investigación universitaria no resultó significativa en la estructura de áreas metropolitanas conectadas, la ocupación en este sector no lo fue en las áreas no conectadas. Este resultado podría estar poniendo de manifiesto que las economías clásicas de aglomeración funcionan mejor en áreas metropolitanas conectadas (A), debido a la existencia de factores no considerados en el modelo (y presentes en un residuo con problemas de heteroscedasticidad espacial, pero sin autocorrelación), mientras que el papel jugado por estos elementos es asumido por la investigación universitaria en la estructura (B) de áreas metropolitanas no conectadas.

$$\ln K_i = \beta_0^A + \beta_1^A \ln R_i + \beta_2^A \ln E_i + \beta_1^B \ln R_i + \beta_2^B \ln E_i + u_i$$

$$u \approx N(0, \sigma^2 I) \quad A.VII.37.$$

- donde K_i : número de innovaciones registradas en el sector de maquinaria industrial, en 1982, en el área i
 R_i : variable “proxy” de la actividad en I+D de las industrias: personal especializado ocupado en laboratorios industriales de I+D, en i .
 E_i : variable “proxy” de la influencia que el entorno económico del área i ejerce sobre el resto
 W : matriz de pesos en la que $w_{ij} = 1$ para observaciones (áreas metropolitanas) situadas a una distancia igual o inferior a 50 millas (también se consideró para 75 millas).

➤ **López y Palacios (2000), López, Palacios y Ruiz (2001)**

Los autores proponen la aplicación de tres modelos sencillos de superficie tendencial, considerando como endógenas tres variables económicas (IPC, renta familiar disponible per capita y paro registrado), en las 47 provincias españolas peninsulares, del modo siguiente:

$$y_i = \alpha + \beta_1 m_i + \beta_2 p_i + u_i \quad A.VII.38.$$

- donde z_i : variable analizada (IPC, renta per capita, tasa de paro) en la provincia i
 m : coordenada de longitud (tendencia este-oeste) de la capital de provincia
 p : coordenada de latitud (tendencia norte-sur) de la capital de provincia

Previamente, a partir de una matriz de interacciones espaciales, de contigüidad de primer orden entre provincias, se había contrastado la existencia de autocorrelación espacial en dichas variables, en un AEDE en el que se aplicó el test I de Moran de dependencia global. El objetivo de la aplicación era comprobar si esta especificación era capaz de absorber la dependencia espacial residual del modelo, aplicando el test I de Moran a los residuos. El resultado negativo indica que deben buscarse nuevos modelos o nuevas especificaciones de la superficie tendencial planteada.

En una nueva aplicación (López *et al.*, 2001) se estima el mismo modelo anterior, para el mismo corte transversal de provincias y la variable tasa de paro, en tres años (1990, 1995 y 2000), con el objeto de comprobar la capacidad explicativa de la superficie tendencial, que se demostró superior a

lo esperado, sobre todo en los últimos años. Aunque volvió a comprobarse que la tendencia del proceso no es, por sí sola, capaz de absorber toda la dependencia espacial de esta variable.

➤ **Pace y LeSage (2000)**

Los autores presentan la llamada especificación MESS (matrix exponential spatial specification) como una forma alternativa de transformación espacial de los datos, especialmente interesante porque estima más rápidamente que los procedimientos habituales el logaritmo de verosimilitud, por lo que se recomienda especialmente para el tratamiento de grandes bases de datos. En esta línea, proponen un modelo de estimación de precios de la vivienda en las 57.641 secciones censales continentales de los EEUU, especificado de la siguiente manera (todas las variables se encuentran expresadas en logaritmos):

$$S(\alpha) \ln P_i = \beta_0 + \beta_1 \ln A_i + \beta_2 \ln P_i + \beta_3 \ln I_i + \beta_4 \ln E_i + \beta_5 D \ln A_i + \beta_6 D \ln P_i + \beta_7 D \ln I_i + \beta_8 \ln E_i + u_i \quad A.VII.39.$$

$$u \approx N(0, \sigma^2 I)$$

donde P_i : precio de la vivienda en la sección censal i
 A : extensión territorial de la sección censal i
 P : población de i
 I : renta per capita de i
 E : edad media de la vivienda construida
 S : matriz exponencia, de orden (n,n) , positiva, definida como:

$$S = e^{\alpha D} = \sum_{i=0}^{\infty} \frac{\alpha^i D^i}{i!}, \text{ siendo } D_i \text{ la matriz de contigüidades espaciales de orden } i \text{ y}$$

α un escalar real, de forma que asigna un peso relativamente menor a los valores de la variable conforme aumente la potencia u orden de vecindad i .

D : matriz (n,n) positiva de pesos espaciales estocástica por filas, definida como:

$$D = \frac{\sum_{i=1}^m \rho^i D_i}{\sum_{i=1}^m \rho^i}, \text{ siendo } D_i \text{ la matriz de contigüidades espaciales de orden } i \text{ y } \rho \text{ el}$$

parámetro autorregresivo espacial que, teniendo el orden i como exponente, expresa el efecto la matriz de vecindad individual de orden i tiene sobre la matriz final D .

En este modelo, las variables endógena y exógenas se encuentran espacialmente transformadas, siendo posible su estimación y contraste a través de un programa diseñado por los autores en el programa Matlab.

➤ **Walker, Moran y Anselin (2000)**

Los autores realizan un análisis inferencial de la deforestación tropical originada por los ranchos de ganado y la producción agrícola en la mayoría de los países latinoamericanos, a partir de datos de destrucción de bosques obtenidos vía satélite. La especificación elegida fue el modelo básico de regresión lineal estimado por MCO, tras aceptar la hipótesis nula de ausencia de autocorrelación

espacial (sustantiva y residual). La muestra de datos procede de una encuesta realizada, en 1993, a una muestra de 132 granjas, propiedad de pequeños productores agrícolas, muchas de ellas no contiguas, localizadas en el estado brasileño de Pará.

$$y_i = \beta_0 + \sum_{k=1}^8 \beta_k x_{ki} + u_i \quad A.VII.40.$$

$$u \approx N(0, \sigma^2 I)$$

- y_i : superficie (ha.) deforestada, obtenida a partir de imágenes vía satélite, en un punto i
 x_k : variable explicativa k -ésima relacionada con las cosechas y la intensidad del uso del suelo: ganado, producción de arroz, judías y maíz, número de cocoteros, plantas de la pimienta y de café, tiempo de ocupación en la actividad agrícola.
 β_0 : término independiente que expresa el resto de causas de la deforestación no atribuibles a la actividad agrícola.

En todos los modelos analizados, los contrastes de autocorrelación espacial (sustantiva y residual) pusieron de manifiesto la ausencia de este efecto, considerando tres tipos de matrices de pesos espaciales, para distancias de 7.000, 20.000 y 40.000 metros entre observaciones. Pese a que la actividad agrícola suele presentar asociación espacial, los autores explican que estos resultados podrían deberse a la no contigüidad entre las observaciones consideradas, localizadas a una distancia media de más de 20.000 metros unas de otras.

➤ **Acs, Anselin y Varga (2001)**

Los autores pretenden demostrar la fiabilidad de dos indicadores muy similares, patentes e innovaciones, como “proxys” de actividad regional de cambio tecnológico, continuando el trabajo expuesto por ellos mismos en Anselin *et al.* (2000). Para ello, construyen un modelo, a partir de una función de producción de conocimiento tipo Cobb-Douglas, en el que dichas variables serán la variable endógena en sendas estimaciones. La información que se maneja es muy extensa, 8.074 innovaciones (nuevo producto, proceso o servicio resultante de un invento) producidas en 125 áreas metropolitanas de los EEUU en el año 1982. Una de las especificaciones propuestas, que se presenta a continuación, es un modelo mixto regresivo cruzado, con perturbación aleatoria autorregresiva, estimado por MV:

$$\ln K_i = \beta_0 + \beta_1 \ln R_i + \beta_2 \ln U_i + \beta_3 \ln Z_i + \beta_5 W \ln R_i + \beta_6 W \ln U_i + u_i$$

$$u_i = \lambda W u_i + \varepsilon \quad A.VII.41.$$

$$\varepsilon \approx N(0, \sigma^2 I)$$

- donde K_i : variable “proxy” del conocimiento (número de innovaciones o patentes) del área i
 R_i : variable “proxy” de la actividad en I+D de las industrias (ocupados en laboratorios industriales de I+D)
 U_i : gasto realizado en las universidades en actividades de investigación (dólares)
 Z_i : variables “proxy” de la influencia que el entorno económico del área i ejerce sobre las innovaciones: ocupados en el sector TIC, empleo empresarial y tasa de grandes empresas (de más de 500 empleados)
 W : matriz de pesos en la que $w_{ij} = 1$ para observaciones (áreas metropolitanas) situadas a una distancia igual o inferior a 50 millas (también se consideró para 75 millas).

➤ **Baller, Anselin, Messner, Deane y Hawkings (2001)**

Los autores analizan la tasa de homicidios cometidos en los condados de EEUU en los años 1960, 1970, 1980 y 1990, a través de cuatro modelos de regresión en los que la tasa media de homicidios se expresa en función de una serie de variables frecuentemente citadas en la literatura como explicativas: un indicador sintético de pobreza (tasa de población de raza negra, renta familiar media, índice de desigualdad de la renta de Gini, tasa de familias con ingresos mínimos y tasa de familias cuyo cabeza de familia es una mujer) y otro de población (total población y densidad de población, ambas en logaritmos). Además, se consideraron otras variables, como la edad media, tasa de paro, tasa de divorcios.

Tal como era de esperar, la tasa de homicidios presenta una fuerte autocorrelación espacial que persiste en el modelo incluso con la introducción en el mismo de variables explicativas clásicas en este tipo de análisis. Por otro lado, se detectaron también dos regímenes o estructuras diferentes entre los condados del Sur (y el distrito de Columbia) y el resto del país, debido a las altas tasas de homicidios (superiores a la media) registradas en el Sur. Asimismo, la perturbación aleatoria también se encuentra afectada por problemas de dependencia espacial, por lo que pone de manifiesto la existencia de otros factores explicativos de la tasa de homicidios no incluidos en el modelo. Por último, las citadas variables explicativas presentaban sensibilidades distintas en las dos estructuras detectadas, siendo el modelo menos preciso en los condados del Sur que en el resto.

Por este motivo, se optó por estimar modelos diferentes según la pertenencia de los condados al Sur (S) o al resto del país norteamericano (R). En concreto, para los estados del Sur (S), en los cuatro años analizados, la más correcta especificación fue la correspondiente al modelo del retardo espacial (modelo mixto autorregresivo de regresión espacial):

$$\text{Sur (S): } \begin{aligned} H_i &= \rho W H_i + \beta_0 + \beta_1 D_i + \beta_2 P_i + \beta_3 V_i + \beta_5 U_i + \beta_6 E_i + u_i \\ u &\approx N(0, \sigma^2 I) \end{aligned} \quad A.VII.42.$$

siendo H_i : tasa de homicidios en el condado i del Sur
 D : indicador sintético de pobreza
 P : indicador sintético de población
 V : tasa de divorcios
 U : tasa de paro
 E : edad media
 W : matriz de contigüidades de primer orden.

Los autores estimaron este modelo para los 4 años en los que disponían de información (1960, 70, 80 y 90) advirtiendo, entre otras cosas, que la capacidad explicativa del coeficiente correspondiente al retardo espacial, ρ , es cada vez menor con el paso de los años.

En cuanto al modelo correspondiente al resto de estados no sureños (R), la mejor especificación fue la correspondiente al modelo del error espacial, con perturbación aleatoria autorregresiva (excepto en el año 1960, para el que hubo que estimarse un modelo del retardo espacial):

$$\begin{aligned}
 H_i &= \beta_0 + \beta_1 D_i + \beta_2 P_i + \beta_3 V_i + \beta_6 E_i + u_i \\
 \text{Resto (R): } u_i &= \lambda W u_i + \varepsilon_i \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{A.VII.43}$$

siendo H_i : tasa de homicidios en el condado i no sureño

La variable tasa de desempleo no resultó significativa en ninguno de los años analizados

➤ **Fischer y Varga (2001)**

Los autores proponen un modelo explicativo de los efectos de desbordamiento (“spillover”), a nivel intra e inter-regional, del conocimiento que produce la actividad investigadora de las universidades sobre la producción del sector TIC en Austria. Se trata, en concreto, de una función de producción de conocimiento, tipo Cobb-Douglas, que explica el output resultante del proceso de producción del conocimiento a partir de una serie de inputs, para los distritos austríacos (provincias). El modelo fue especificado como un modelo del error espacial, estimado por MV, que en su forma lineal se expresa del modo siguiente:

$$\begin{aligned}
 \ln P_i &= \beta_0 + \beta_1 \ln \Omega_i + \beta_2 \ln \Phi_i + u_i \\
 u_i &= \lambda W u_i + \varepsilon \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{A.VII.44}$$

- donde P_i : patentes registradas en el distrito i (variable “proxy” del nuevo conocimiento corporativo generado por las empresas de alta tecnología).
- Ω_i : indicador de gasto en actividad I+D realizado, no sólo por las industrias radicadas en i , sino también por las situadas en distritos vecinos, convenientemente corregidas por una variable de distancia, del modo siguiente: $\log \Omega_i = \log R_i + \log A_i^R$, siendo R_i el gasto realizado por las industrias radicadas en i , en actividad I+D, y A_i^R un indicador de accesibilidad que busca captar potenciales efectos desbordamiento del conocimiento entre empresas de I+D, que, a su vez, viene expresado, para $i \neq j$ como: $A_i^R = \sum_j R_j d_{ji}^{-\beta}$, siendo d_{ji} la distancia por carretera (en kilómetros) entre j e i , y β un parámetro de fricción.
- Φ_i : indicador de gasto en actividad de investigación realizado por las universidades radicadas en i , así como por las situadas en distritos vecinos, del modo siguiente: $\log \Phi_i = \log U_i + \log A_i^U$, siendo U_i el gasto realizado por las universidades de i en investigación y A_i^U un indicador de potenciales efectos desbordamiento de la investigación universitaria interregional, que, a su vez, viene expresado, para $i \neq j$ como: $A_i^U = \sum_j U_j d_{ji}^{-\beta}$.
- W : matriz de pesos espaciales definida como matriz de contigüidades de primer orden, matriz de distancias inversas (unitaria y cuadrática y matrices de intervalos (50 km, 75 km y 100 km) entre los centros administrativos de los distritos).

➤ **Sandberg y Johansson (2001)**

Los autores formulan un modelo de regresión espacial en el que se estiman los precios hedónicos de las viviendas en régimen de cooperativa existentes en la ciudad sueca de Umeå, en el período 1998-99. Los precios hedónicos recogen, no sólo el precio de un bien, sino también el valor añadido que genera cualquier cambio cualitativo (susceptible de ser apreciado por el consumidor). En línea con este tipo de estudios, se definen tres tipos de variables referidas, por un lado, a cualidades propias (cuota mensual, superficie de la vivienda y número de habitaciones), y por otro, a características del entorno (densidad de población, densidad de viviendas individuales y tasa de retorno de las viviendas en régimen de cooperativa) y, por último, variables de accesibilidad (distancia a la universidad, al centro urbano de negocios y al centro comercial más cercano).

El modelo, inicialmente estimado por MCO, presentaba problemas de autocorrelación espacial y no normalidad en la perturbación aleatoria. Por ese motivo, los autores se decantaron por la estimación por el método general de los momentos⁷⁷ (MGM) del siguiente modelo del error espacial:

$$\begin{aligned}
 P_i &= \beta_0 + \beta_1 F_i + \beta_2 R_i + \beta_3 D_i + \beta_4 S_i + \beta_5 T_i + \beta_6 U_i + \beta_7 B_i + \beta_8 C_i + u_i \\
 u_i &= \lambda W u_i + \varepsilon_i \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{A.VII.45}$$

donde P_i : precio de venta por metro cuadrado de vivienda, en i
 F : cuota mensual (a pagar por el inquilino a la cooperativa) por metro cuadrado
 R : número de habitaciones de la vivienda
 D : densidad de población
 S : densidad de viviendas individuales
 T : tasa de retorno de las vivienda en régimen de cooperativa
 U : distancia a la universidad
 B : distancia al centro urbano de negocios
 C : distancia al centro comercial más cercano
 u : perturbación aleatoria con problemas de autocorrelación espacial de primer orden
 ε : perturbación aleatoria esférica
 W : matriz de pesos espaciales

La matriz de pesos espaciales recibió 5 especificaciones basadas en diferentes intervalos de distancia entre observaciones: 1 km, 1,5 km., 2 km., 2,5 km y 3 km.

Finalmente, los autores aplicaron un método de alisado espacial para conocer los precios de la vivienda en régimen de cooperativa en todos los puntos de la ciudad, especialmente en nuevas zonas de expansión de la ciudad, en las que existen proyectos de construcción de nuevas viviendas.

➤ **Thomas (2001)**

Se propone un modelo de crecimiento económico de las 214 ciudades francesas más importantes, en el período 1982-1994, especificado como un modelo del error espacial de la siguiente forma:

⁷⁷ Una solución alternativa vendría constituida por el método de variables instrumentales (VI) que fue desechado al no existir en la literatura propuestas adecuadas sobre los instrumentos a utilizar en este tipo de aplicaciones.

$$\begin{aligned}
 P_i &= \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 G_i + \beta_4 E_i + \beta_5 D_i + u_i \\
 u_i &= \lambda W u_i + \varepsilon_i \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{A.VII.46}$$

- donde y_i : impuesto de actividad económica desarrollado por profesionales de i en 1994
 I : gasto realizado por el ayuntamiento en inversión en 1996
 P : productividad del factor trabajo, expresada como VAB por ocupado (1996)
 G : gasto previsto en los presupuestos municipales (1994)
 E : variable ficticia que adopta el valor 1 en ciudades en las que existen planes estratégicos de desarrollo urbano
 D : distancia (en kilómetros) entre cada ciudad y París
 R : variable cuyos valores coinciden con el puesto ocupado por cada ciudad en el ranking de VAB generado por las mismas (1996)
 W : matriz de interacciones o contigüidades de primer orden

➤ **Van der Kruk (2001)**

En este caso, se propone también una aplicación de precios hedónicos, en concreto, para conocer el impacto que las tierras bajas (húmedas) holandesas producen sobre el precio de las viviendas cercanas. Las tierras bajas conllevan un importante valor añadido, no sólo por su diversidad biológica, sino por aportar a los residentes cercanos zonas de espacio abierto, agradables vistas, así como un aislamiento de ruidos y otras formas de polución atmosférica. El autor, en un contexto aún teórico, propone un modelo mixto autorregresivo regresivo cruzado con perturbación autorregresiva espacial, denominado modelo SARMA, que explica el precio de mercado de la vivienda, en un punto del espacio geográfico y un momento del tiempo, en función de ciertas características propias de la vivienda así como del precio de mercado y características de viviendas situadas en localizaciones cercanas del modo siguiente:

$$\begin{aligned}
 y &= \rho W_1 y + X \beta_1 + W_2 R \beta_2 + u \\
 u &= \lambda W_3 u + \varepsilon \\
 \varepsilon &\approx N(0, \sigma^2 I)
 \end{aligned}
 \tag{A.VII.47}$$

- donde y : precio de mercado de una vivienda dada
 W_1, W_2, W_3 : matrices de pesos espaciales
 X : matriz del total considerado de características de la vivienda consideradas
 R : matriz de un determinado número de características de viviendas vecinas
 ρ, β_1, β_2 : vectores de parámetros
 u : vector $(N, 1)$ de perturbaciones aleatorias autorregresivas (de primer orden)
 λ : parámetro autorregresivo asociado al retardo espacial $W_3 u$
 ε : vector de perturbaciones aleatorias, ruido blanco.

El autor propone la descomposición de la matriz W de pesos espaciales en tres matrices con orden de contigüidad 1, 2 y 3, convenientemente ponderadas (ver con más detalle en el Anexo V, Apartado A.V.I), del modo siguiente:

$$W = \sum_{d=1}^D W_d \quad A.VII.48.$$

siendo W: matriz final de pesos espaciales
 d: orden de vecindad
 D: diámetro del sistema espacial o número máximo de órdenes de vecindad existentes en el mismo.

La primera matriz de pesos correspondiente al primer orden de vecindad, W_1 , es una matriz binaria de interacciones espaciales cuyos elementos valen 1 ó 0 si las observaciones que representan están o no unidas por algún tipo de vecindad, respectivamente. En cuanto a la segunda matriz de pesos, W_2 , correspondiente al segundo orden de vecindad, la existencia de relaciones entre observaciones es ponderada por $\frac{1}{2}$, y así sucesivamente. De este modo, se incorporan a la matriz final W todas las relaciones de vecindad existentes para cada observación, convenientemente ponderadas según el orden con el que se producen.

➤ Ceccato, Haining y Signoreta (2002)

Los autores proponen la especificación de varios modelos explicativos de distintas formas de inseguridad ciudadana (delincuencia) existente en la ciudad de Estocolmo, a partir de datos procedentes de 119 divisiones territoriales de dicha localidad, registrados en 1998. Estas divisiones fueron generadas (a partir de las 360 secciones oficiales) con el criterio de obtención de un volumen de población similar en cada una de ellas. Como variables explicativas, se seleccionaron varios indicadores socioeconómicos a partir de la literatura existente sobre el tema.

En primer lugar, se estimó un modelo básico de regresión lineal, por MCO, del número de actos vandálicos de la manera siguiente:

$$y_i = \beta_0 + \beta_1 x_{3i} + \beta_2 x_{10i} + u_i \quad A.VII.49.$$

$$u \approx N(0, \sigma^2 I)$$

donde y: tasa de vandalismo contra la propiedad privada y los espacios públicos (destrucción ilegal de la propiedad en calles, plazas y transporte público), por hectárea
 x_3 : tasa de paro de los residentes de edades comprendidas entre 25-64 años
 x_{10} : variable ficticia que discrimina entre casos de vandalismo cometidos en el centro urbano o los barrios de la periferia

En segundo lugar, se especificó un modelo del error espacial, estimado por MV, de robos de vehículos:

$$y_i = \beta_0 + \beta_1 x_{7i} + \beta_2 x_{10i} + u_i \quad A.VII.50.$$

$$u_i = \lambda W u_i + \varepsilon_i$$

$$\varepsilon \approx N(0, \sigma^2 I)$$

donde y: tasa de robos en i de vehículos, por hectárea
 x_7 : tasa de hogares multifamiliares

- x_{10} : variable ficticia que discrimina entre casos de vandalismo cometidos en el centro urbano o los barrios de la periferia
- W : matriz de interacciones espaciales cuyos elementos w_{ij} adoptan el valor uno cuando poseen una frontera común o se encuentran en un radio de contigüidad de primer a tercer orden, y cero en caso contrario.

En último lugar, se presenta un modelo básico de regresión lineal, estimado por MCO, de delitos cometidos en residencias privadas:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{4i} + \beta_3 x_{5i} + \beta_4 x_{7i} + u_i$$

$$u \approx N(0, \sigma^2 I)$$

A.VII.51.

- donde y : tasa de delitos en residencias privadas (con o sin robo), por número de hogares.
- x_1 : tasa de aloctonía (residentes nacidos en otro lugar)
- x_4 : tasa de ingresos por población potencialmente activa (16-64 años)
- x_5 : tasa de inmigración
- x_7 : tasa de hogares multifamiliares

ANEXO II. APLICACIONES DE PREDICCIÓN-EXTRAPOLACIÓN DE MAGNITUDES ECONÓMICAS MICROTERRITORIALES

En los últimos años, algunos organismos públicos y privados de países como España o Portugal, están realizando estimaciones de magnitudes económicas, sobre todo, la renta familiar disponible (muchas veces identificada con la renta bruta disponible de los hogares), que debe su interés a que se trata de la macromagnitud más aproximada al poder adquisitivo. Esto justifica el esfuerzo de los analistas por estimar este indicador para ámbitos cada vez más microterritoriales, no sólo las provincias, comarcas o municipios, sino también en unidades como los distritos, barrios y secciones censales. A continuación, se presentan algunas propuestas interesantes al respecto.

➤ **José Ayuso, José Antonio Fernández y José Luis Plaza (Banesto)**

El Banco Español de Crédito, Banesto, fue el pionero en la elaboración de Anuarios con información microterritorial para el conjunto de España. Desde 1967 hasta 1993, publicó el Anuario del Mercado Español (A.M.E.) en el que, entre otros indicadores, se incluía el “nivel de renta” o renta familiar disponible por habitante (escalada en 10 niveles) para el total de municipios españoles, siendo la última actualización correspondiente al año 1991.

Los directores técnicos de la obra optaron por una estimación indirecta de la renta bruta disponible de los hogares, de ámbito municipal, basada en un modelo de regresión lineal, estimado por mínimos cuadrados ordinarios (MCO), a partir de la renta provincial publicada por el Banco Bilbao Vizcaya (actualmente, BBVA) y de un conjunto de variables explicativas para las que se disponía también de información municipal. Por ejemplo, en la última edición del AME 1993, la variable endógena del modelo fue la renta familiar disponible por habitante provincial (dato de avance publicado por el BBV referido al año 1991) explicada a partir de las siguientes variables exógenas, también de ámbito provincial (todas ellas relativizadas por la población de hecho del Censo de Población de 1991):

- Líneas telefónicas en funcionamiento y pendientes de instalación (1992): de esta variable se destaca su carácter explicativo, siempre, de la renta de las familias.
- Segunda vivienda (Censo de Viviendas, 1991): entendiéndose que los municipios en los que se produce una alta concentración de este tipo de vivienda, suelen realizar actividades económicas derivadas de las exigencias de servicios a las familias que las ocupan con frecuencia.
- Oficinas de bancos y cajas de ahorro (1991): considerada como la variable que mejor refleja la actividad productiva de un municipio.

Los autores, que también utilizaron otras variables exógenas (por ejemplo, en el AME 1988, las licencias comerciales), indican que la selección de variables explicativas en este tipo de modelos debe realizarse según los siguientes criterios: existencia de información para todos los municipios, fiabilidad de los datos y carácter explicativo de la renta disponible de las familias.

Por tanto, el modelo de regresión especificado fue el siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad A.IX.1.$$

siendo y : renta familiar disponible por habitante de la provincia i
 x_1 : teléfonos por habitante
 x_2 : segunda vivienda por habitante
 x_3 : bancos y cajas por habitante
 β : parámetro a estimar por MCO
 u : perturbación aleatoria.

Los principales resultados de la estimación proporcionados por los autores son los siguientes:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \quad R^2 = 0,91 \quad A.IX.2.$$

(10,95)
(-2,94)
(5,63)

Una vez estimados, los parámetros $\beta_0, \beta_1, \beta_2, \beta_3$ son aplicados a los valores de las variables x_1, x_2, x_3 correspondientes a cada municipio obteniendo, de esta manera, el dato de renta familiar disponible por habitante, que es posteriormente escalado en 10 niveles, de la manera siguiente:

Niveles	Renta por hab. (pts)	Niveles	Renta por hab. (pts)
1	Hasta 440.000	6	880.000-1.100.000
2	440.000-515.000	7	1.100.000-1.320.000
3	515.000-585.000	8	1.320.000-1.615.000
4	585.000-700.000	9	1.615.000-1.900.000
5	700.000-880.000	10	Más de 1.900.000

A continuación, se presentan algunos de los resultados obtenidos, en concreto, para los municipios de la Comunidad de Madrid, presentando las puntuaciones extremas, los municipios con nivel de renta 1 y 10. Como puede observarse, los mejores resultados en renta familiar disponible tienen lugar en municipios de pequeño o mediano tamaño que concentran alguna importante actividad económica (polígonos industriales o segunda vivienda), como Ajalvir y Nuevo Baztán, pero no en localidades residenciales en las que reside la población de alto "standing", como Boadilla del Monte, Majadahonda o Pozuelo de Alarcón. Estos resultados ponen de manifiesto la debilidad de las variables habitualmente consideradas como buenas explicativas de la renta familiar.

Mayor nivel de renta (1991)		Menor nivel de renta (1991)	
Municipio	Nivel de renta	Municipio	Nivel de renta
Ajalvir	10	Braojos	1
Aldea del Fresno	10	Fresno de Torote	1
Nuevo Baztán	10	Hiruela (La)	1
Pozuelo del Rey	10	Horcajo de la Sierra	1
Serranillos del Valle	10	Horcajuelo de la Sierra	1
Valverde de Alcalá	10	Madarcos	1
		Prádena del Rincón	1
		Puebla de la Sierra	1
		Santa María de la Alameda	1

Los principales problemas que plantea este modelo podrían resumirse en los siguientes puntos:

- Se identifica producción con renta de las familias, al considerar como óptimos los indicadores de actividad económica (segunda vivienda y entidades de crédito).

- No se explica el sentido del signo negativo (incorrecto) del coeficiente propio de la variable segunda vivienda, probablemente debido a la presencia de efectos espaciales de autocorrelación y/o heterogeneidad espacial en el modelo.

➤ SADEI

La Sociedad Asturiana de Estudios Económicos e Industriales (SADEI) realiza, desde 1976, una estimación bienal de la renta familiar disponible de los concejos (municipios) asturianos, basada en un método directo de estimación que se contrasta posteriormente con otro indirecto. En concreto, la estimación de la renta de los municipios asturianos, para el año 1994, constaría de dos etapas:

1. Estimación de la renta familiar disponible municipal por un método directo a partir de una matriz municipal de empleo (población legalmente empleada en Asturias, que incluye el empleo asalariado y autónomo) que consta de 78 filas (municipios) y 47 columnas (ramas de producción). Para ello, además de estimaciones propias, se ha contado con diversas fuentes de excepción, como la Estadística de Convenios Colectivos del Ministerio de Trabajo y Seguridad Social, los índices de producción industrial de Asturias procedentes de la Encuesta Industrial (INE), la explotación desagregada por actividades de la Central de Balances del Banco de España (para las empresas con establecimientos en Asturias), memorias de empresas y liquidaciones presupuestarias de organismos públicos.

De esta manera, es posible calcular el ratio de remuneración por asalariado, los excedentes de explotación, el consumo de capital fijo (amortizaciones) y el valor de la producción por empleo (asalariado y no asalariado) para cada una de las 47 ramas de actividad en que se encuentra dividida la economía asturiana. La aplicación de los ratios regionales a los datos municipales de empleo permite la obtención del producto interior bruto municipal. La conversión del producto interior a producto municipal (correspondiente únicamente a los residentes) es posible a través del saldo de pagos a los factores de producción no residentes (salarios y excedentes netos de explotación).

Una vez obtenido el producto o renta municipal, el cálculo de la renta familiar disponible se realiza en una doble operación: por un lado, se deducen del producto municipal las rentas no percibidas por las familias (impuesto sobre renta y patrimonio de las empresas, ahorro de las empresas y rentas del estado) y las deducciones de la renta personal (impuesto sobre la renta y el patrimonio de las familias, y cotizaciones sociales); por otro lado, al resultado anterior deben añadirse las transferencias monetarias realizadas a las familias (prestaciones sociales y otras) y en especie (enseñanza y sanidad). Los autores ponen de manifiesto el interés del cálculo no sólo de la renta familiar disponible monetaria, sino también en especie que conjuntamente dan lugar a la renta familiar disponible total.

2. Esta estimación de renta familiar disponible municipal, para 1994, es contrastada con otra estimación, de carácter indirecto, de la misma a partir de un modelo de regresión de la renta familiar disponible municipal, estimada por SADEI para 1992, y una batería de cien variables explicativas, referidas también a 1992 (aunque disponibles también para 1994), previamente agrupadas en 20 factores (por componentes principales y rotación Varimax).

El grupo de variables explicativas es muy exhaustivo y comprende indicadores de variación de población, tasas de actividad, ocupación, empleo y paro por sectores, cabaña de ganado vacuno según razas (por hectárea de SAU o superficie agraria utilizada), potencia del parque de maquinaria agrícola, según clase de máquinas (por ha. de SAU); tasa de licencias

comerciales, según tipos de productos; tasa de parque de vehículos, según tipos; tasa de servicio telefónico; tasa de oficinas bancarias; tasa de matriculación, según nivel de enseñanza; tasa de servicios sanitarios; tasa de presupuestos municipales.

A partir de aquí, la renta familiar disponible municipal de 1994 se calcula aplicando los coeficientes estimados por MCO en la ecuación de regresión a unos componentes idénticos, referidos a las mismas 100 variables explicativas correspondientes al año 1994. El vector diferencia entre los resultados obtenidos por los dos métodos, directo e indirecto, para la renta familiar disponible municipal de Asturias, en 1994, recoge por un lado, los pagos netos intermunicipales de rentas de la propiedad y de la empresa, así como también los márgenes de error propios de los procedimientos de estimación empleados.

A continuación, se presentan las puntuaciones extremas de renta familiar disponible por habitante de 1994 de los municipios asturianos.

Mayor renta disponible por hab. (1994)		Menor renta disponible por hab. (1994)	
Municipio	Renta (pta)	Municipio	Renta (pta)
Oviedo	1.282.565	Amieva	859.544
Noreña	1.271.314	Somiedo	866.294
Castrillón	1.260.064	Onís	868.544
Siero	1.237.563	Pesoz	869.668
Gijón	1.203.811	Illano	871.919
Degaña	1.192.560	Santo Adriano	874.169
Llanera	1.181.310	Villayón	875.294
Ibias	1.170.059	Santa Eulalia de Oscos	877.544
Avilés	1.147.558	San Tirso de Abres	878.670
Mieres	1.141.933	Caravia	879.795

El problema más importante que puede plantear un método directo de estimación de renta familiar disponible, además de no controlar todas las fuentes de ingresos, es su incapacidad para detectar la aportación de la economía sumergida a los ingresos de muchos hogares. Quizá por este motivo, municipios asturianos de fuerte actividad turística y/o comercial, como Cangas de Onís, El Franco, Llanes, Rivadesella y Vegadeo obtengan, con este método, menor nivel de renta per capita del que posiblemente les correspondería.

➤ **Enrique Urcola, Pascual Matesanz, Carlos Molpeceres y Miguel Ángel Braña (CIISA)**

CIISA (Consulting e Ingeniería Int., S.A.), con el patrocinio de la Fundación Caixa Galicia, publicó, en 1988, una primera aproximación a la estimación de la renta familiar disponible de los municipios gallegos, para los años 1982, 1984 y 1986. Para ello, estima tres modelos de regresión lineal, por el método MCO, en los que la variable endógena es la renta familiar disponible publicada por el BBV para los años 1981, 1983 y 1985, respectivamente, y las variables exógenas diferentes en cada caso.

Según los autores, las variables explicativas del modelo deben cumplir la condición de estar muy correlacionadas con la variable endógena, aunque escasamente relacionadas matemáticamente entre sí, con el objeto de poder individualizar la contribución marginal de cada variable a la explicación del modelo. Además, dichas variables explicativas deben estar también disponibles, para el ámbito municipal, en el período de predicción, 1982, 1984 y 1986 que, como puede observarse, presentan un desfase temporal de un año respecto del período de estimación. Además, debido a este desfase, las series de renta provincial fueron también deflactadas conforme al índice de precios al consumo

(IPC) provincial para cada año de estimación. Los resultados de la estimación de los tres modelos se presentan a continuación:

1. Modelo de regresión múltiple provincial, para 1981, estimado por MCO:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + u_i \quad R^2 = 0,91 \quad A.IX.3.$$

siendo y_i : renta familiar disponible por habitante (BBV) para la provincia i
 x_1 : empleo asalariado en el sector secundario por mil habitantes
 x_2 : media tipificada de turismos y teléfonos
 x_3 : parque de motocicletas por mil habitantes
 x_4 : licencias comerciales alimentarias por mil habitantes
 x_5 : licencias comerciales no alimentarias por mil habitantes

2. Modelo de regresión múltiple provincial, para 1984, estimado por MCO:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + u_i \quad R^2 = 0,91 \quad A.IX.4.$$

siendo x_6 : parque de camiones por mil habitantes

3. Modelo de regresión múltiple provincial, para 1986, estimado por MCO:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_6 x_{6i} + \beta_7 x_{7i} + u_i \quad R^2 = 0,91 \quad A.IX.5.$$

siendo x_7 : licencias comerciales totales por mil habitantes

A continuación, se presentan las puntuaciones extremas de renta familiar disponible por habitante de 1986 de los municipios gallegos.

Mayor renta disponible por hab. (1994)		Menor renta disponible por hab. (1994)	
Municipio	Renta (pta)	Municipio	Renta (pta)
Cervo (LG)	736.560	Dozón (PO)	267.601
Vigo (PO)	726.488	O Covelo (PO)	274.826
O Grove (PO)	725.600	Baltar (OU)	275.154
A Coruña (CO)	723.118	Castrelo de Miño (OU)	276.103
Ourense (OU)	710.348	Toques (CO)	277.652
As Pontes (CO)	710.339	Rairiz de Veiga (OU)	280.369
San Cibrao das Viñas (OU)	707.857	Os Blancos (OU)	281.392
Carballada (OU)	700.186	Lobeira (OU)	283.591
Lugo (LU)	690.189	Verea (OU)	285.681
A Rúa (OU)	684.952	Muiños (OU)	287.846

Al igual que en las estimaciones de Banesto, los mejores resultados en renta familiar disponible tienen lugar en municipios de pequeño o mediano tamaño que concentran alguna importante actividad económica, no quedando siempre bien situadas algunas localidades residenciales, como sería el caso de Negreira y Ames (A Coruña), situadas en puestos muy retrasados en el ranking de renta disponible per capita de Galicia.

Los principales problemas que plantea este modelo podrían resumirse en los siguientes puntos:

- Se identifica producción con renta de las familias.
- En lugar de tres modelos de regresión diferentes, podría haberse estimado un modelo de datos de panel con un corte temporal de tres años y datos provinciales.

➤ CAZAR

La Caja de Ahorros de Zaragoza, Aragón y Rioja, CAZAR (actualmente Ibercaja), publicó en 1984 la estimación de la renta municipal disponible de Aragón para el año 1981. Al igual que Sadei, estima por un método directo la renta municipal, teniendo que recurrir a un método indirecto para obtener la renta municipal disponible. Este último método se basa en un modelo de regresión lineal, estimado por MCO, a partir de la renta familiar disponible provincial del Banco de Bilbao-Vizcaya (BBVA hoy) en 1981 en función de una serie de variables explicativas (todas ellas relativizadas por la población), como turismos, activos en la industria y saldos de imposición en cajas de ahorros (en una estimación posterior, CAZAR, 1986, considera como explicativas los depósitos de ahorro, oficinas bancarias, flujos intermunicipales de personas y tasa de ocupación sectorial). Se comprobó también la capacidad explicativa de dichas variables en años anteriores (1971, 1973, 1975, 1977 y 1979).

A continuación, se clasificaron las provincias españolas en 4 grupos, para lo que, en primer lugar, se realizó un análisis de componentes principales con las tres variables explicativas seleccionadas, obteniéndose dos factores que conjuntamente representaban casi el 90% de la varianza total. En segundo lugar, tras representar las 50 provincias españolas sobre un eje cartesiano con los dos factores obtenidos, se establecieron 4 grupos provinciales por afinidad en las distancias euclídeas, construyendo, a su vez, grupos similares entre los municipios. Por último, para comprobar que la tipología establecida no hacía variar las estructuras que describen la relación entre la renta y las variables explicativas, se realizó también un análisis de la covarianza.

Finalmente, para cada grupo se estimaron las 4 relaciones definitivas, a partir de las cuales, conocidos los valores municipales de las 3 variables explicativas y sustituyendo estos valores en la expresión que a cada municipio le correspondía, se obtuvieron los valores de renta municipal disponible de Aragón, para 1981.

Entre los problemas que plantea este método cabe destacar que, aunque intenta adaptarse más a la realidad regional estableciendo una tipología de municipios, sigue considerando que existen las mismas elasticidades entre variables para el ámbito provincial que municipal. Por ejemplo, mientras que la variable de activos en la industria suele mostrar una clara relación positiva con la renta disponible en el ámbito provincial, puede ser poco significativa de la renta familiar en el nivel municipal, dado que presenta altos valores en localidades de carácter industrial-obrero en las que la renta disponible de los residentes no suele ser muy elevada, y viceversa en el caso de núcleos de carácter residencial.

➤ Remírez

Remírez (1991) presenta los resultados de la estimación de la renta familiar disponible de 1985 para los municipios navarros, a partir de un modelo de regresión lineal que relaciona la renta familiar disponible por habitante de Navarra, en el período 1967-1985 (a partir de la serie del BBV), en función de los presupuestos municipales por habitante, del modo siguiente:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 f_{67} + \beta_3 f_{68} + \beta_4 f_{69} + \beta_5 f_{70} + \beta_6 f_{71} + \beta_7 f_{76} + u_t \quad A.IX.6.$$

siendo y: renta familiar disponible por habitante de Navarra (BBV), t = 1967-1985
 x₁: presupuestos de gastos municipales por habitante
 f: variable ficticia que adopta el valor 1 en el período 1967 a 1976
 u: perturbación aleatoria

Este modelo pretende superar las limitaciones que suele plantear la aplicación de relaciones obtenidas para el ámbito provincial global de España, en el ámbito municipal de una determinada comunidad autónoma, pero incurre en otros problemas que se resumen a continuación:

- En el caso español, es muy difícil obtener una información suficientemente amplia de series estadísticas homogéneas para cada provincia.
- Se intenta superar el problema de ausencia de estabilidad estructural espacial entre los ámbitos provincial-municipal a cambio de suponer la estabilidad temporal, sin contrastar este supuesto que podría no ser cierto.
- Los coeficientes estimados en un contexto temporal, para una provincia determinada, se utilizan en otro contexto de corte transversal (municipios de dicha provincia), para estimar la renta per capita municipal, siendo el significado de los coeficientes también diferente (y quizá más) entre los ámbitos temporal y espacial.
- Por último, debe también indicarse que la variable de presupuestos de gastos municipales por habitante no es buena explicativa de la renta familiar disponible, resultando muy elevada en municipios de tamaño pequeño o mediano con alguna actividad económica importante (industria, comercio, turismo, etc.).

➤ **Fernández (Universidad de Navarra) y Sierra (Universidad de Zaragoza)**

Estos autores presentaron una comunicación en la VI Reunión de Asepelt (Granada) en la que proponían un método de estimación de renta familiar disponible municipal de La Rioja, para 1985. Esta propuesta metodológica, inspirada en el modelo de Remírez (1991), consiste en la estimación de un modelo de regresión lineal temporal de la renta familiar disponible de La Rioja (procedente del BBV), para el período 1967-1985, en función del consumo de energía eléctrica para el uso doméstico (variable también disponible para los municipios riojanos), habiendo relativizado ambos indicadores por la población, del modo siguiente:

$$y_t = \beta_0 + \beta_1 x_t + u_t \quad R^2 = 0,97 \quad DW = 1,96 \quad A.IX.7.$$

siendo y: renta familiar disponible por habitante de La Rioja (BBV), en el período 67-85
 x: consumo de energía eléctrica para uso doméstico por habitante
 u: perturbación aleatoria
 t: período 1967-1985

Este modelo fue estimado por el método Prais-Whinsten (Johnston, 1984), al presentar los residuos una alta correlación temporal en la estimación MCO. Además, los autores se decantaron finalmente por la variable consumo de energía eléctrica, tras un proceso de selección en el que se consideraron otros indicadores (siempre relativizados por la población), como líneas telefónicas, consumo de

energía eléctrica total, parque de camiones, potencia del parque de maquinaria agrícola, licencias comerciales mayoristas y licencias comerciales minoristas, en los cuales se detectó una fuerte multicolinealidad que, aunque motivó, en un primer intento, la aplicación de componentes principales sobre los mismos, finalmente fue también rechazada por el principio de parsimonia, derivado de las consideraciones siguientes.

Por un lado, aunque la mayoría de los indicadores presentó alta correlación con la renta familiar disponible en el período considerado, las variables correspondientes a las licencias comerciales se revelaron poco significativas, por lo que fueron rechazadas. Por otro lado, los autores constataron también la problemática existente en los municipios muy pequeños en los que algunas variables, como potencia del parque de maquinaria agrícola o parque de camiones, presentan valor cero, por carecer de dicho equipamiento, lo que revelaba la incapacidad de estos indicadores para explicar la renta de los residentes en dichas localidades. En cuanto a la variable líneas telefónicas, aunque sólo tres municipios carecían de este servicio, se comprobó que un número considerable de ellos sólo disponían de una línea, en la mayoría de los casos correspondiente a la centralita pública, lo que ponía de manifiesto la dependencia que presentaba este indicador de la infraestructura desarrollada por la compañía telefónica más que de la renta disponible de las familias, por lo que fue también desechada. Por último, entre las variables de consumo de energía eléctrica, total y uso doméstico, se optó por la última de ellas, más cercana a la renta de las familias que la primera, que considera las rentas procedentes de otros sectores.

Una vez obtenidos los valores de la variable consumo de energía eléctrica para uso doméstico por habitante, la aplicación de los coeficientes estimados por el modelo sobre dichos valores proporciona los datos de renta familiar disponible por habitante municipal de La Rioja, en 1985, cuyos principales resultados se exponen a continuación:

Mayor renta disponible por hab. (1985)		Menor renta disponible por hab. (1985)	
Municipio	Renta (pta)	Municipio	Renta (pta)
Fuenmayor	675.091	Pazuengos	394.560
Baños de Río Tobía	673.369	Cabezón	395.396
Rasillo (El)	671.728	Santa Engracia	468.486
Casalarreina	671.539	Villarroya	471.019
Cenicero	668.301	Cidamón	472.894
Briñas	660.743	Ajamil	478.863
Nalda	660.571	Sojuela	484.508
Nájera	659.854	Lumbreras	489.571
Haro	655.357	Cornago	488.995
Calahorra	654.833	Bezares	490.586

Los autores ponen de manifiesto algunas de las limitaciones del modelo propuesto, a las que se añaden otras que apuntamos a continuación:

- Hay municipios de pequeño tamaño (población de derecho inferior a 100 habitantes) cuya población real es bastante menor (de hecho no residen allí) en los que el modelo subvalora la renta familiar disponible debido a la relativización de las variables por una población superior a la real. Éste sería el caso de Cellorigo, Cidamón, Pazuengos, Castroviejo, Bezares y Cabezón.
- Por otro lado, existen municipios con una población de hecho muy superior a la oficialmente empadronada, dado que acogen gran afluencia de turistas con segunda vivienda en los mismos, en los que el modelo produce estimaciones de renta muy superiores a la realidad, como el caso de El Rasillo.

- Además, algunos municipios presentan también valores atípicos de renta per capita alta debido a la presencia en los mismos de una importante actividad industrial, como Baños de Río Tobía y Nájera.
- A estas consideraciones cabría añadir la misma crítica realizada para Remírez (1991) sobre el peligro de inferir relaciones estimadas en un contexto temporal a un ámbito espacial y la ausencia de contrastes de cambio estructural temporal.
- Por último, debe indicarse también que, en el caso español, hoy por hoy, resulta del todo imposible obtener la variable consumo de energía eléctrica de uso doméstico para todos los municipios de España, debido a que no todas las empresas suministradoras están dispuestas a facilitar dicha información.

➤ **José Pérez Blanco (Diputación de Sevilla)**

En 1995, la Diputación de Sevilla publicó una estimación de renta personal de los municipios de Sevilla, en los años 1986 a 1994, mediante la estimación por MCO de un modelo de regresión potencial de la renta personal por habitante de la provincia, en el período 1961-1994, en función del consumo de energía eléctrica de uso doméstico.

$$y_t = e^{\alpha} x_t^{\beta} \quad R^2 = 0,99 \quad A.IX.8.$$

donde y: renta personal por habitante de la provincia de Sevilla, en pesetas
 constantes de 1990 (1961-93: BBV; 1994: Anuario de El País)
 x: consumo de electricidad para usos domésticos, kwh/hab/año (Sevillana)
 α, β : parámetros a estimar
 t: período 1961-1994

En un estudio anterior, el autor había considerado, junto con el consumo de energía eléctrica, la presencia en el modelo de otras variables explicativas, como el parque de turismos y camiones o las líneas telefónicas, pero fueron desechadas debido a la fuerte multicolinealidad presentada entre ellas. Consideradas de forma individual, la variable consumo de energía eléctrica de uso doméstico por habitante resultó ser la más eficaz como explicativa de la renta per capita, por lo que fue elegida como única exógena en el modelo.

La renta personal municipal para los años 1986 a 1994 se calcula aplicando esta ecuación de regresión a los valores de consumo de electricidad para usos domésticos de los municipios sevillanos en el citado período. Por ejemplo, en 1994, los municipios con valores extremos de renta per capita serían los siguientes:

Mayor renta personal por hab. (1994)		Menor renta personal por hab. (1994)	
Municipio	Renta (pta)	Municipio	Renta (pta)
Valencina de la Concepción	1.729.087	Villanueva de San Juan	669.942
Palomares del Río	1.720.695	Saucejo (El)	828.862
Espartinas	1.871.582	Molares (Los)	829.939
Umbrete	1.603.489	Huévar	835.316
Algarrobo (El)	1.579.358	Corrales (Los)	845.792
Tomares	1.489.793	Martín de la Jara	868.464
Mairena del Aljarafe	1.474.327	Marinaleda	869.582

Mayor renta personal por hab. (1994)		Menor renta personal por hab. (1994)	
Carmona	1.441.104	Cañada Rosal	888.148
Gines	1.436.177	Pruna	896.641
Mairena de Alcor	1.377.187	Alcolea del Río	899.838

Uno de los problemas que plantea este método estriba en el concepto mismo de renta que maneja, renta personal o ingresos de las familias, que no coincide exactamente con la noción de renta familiar disponible (que es la que se desea explicar mediante el consumo eléctrico de uso doméstico). Efectivamente, la principal diferencia entre ambas consiste en que la renta personal incluye las deducciones de la renta personal (cotizaciones sociales) y no considera las transferencias que reciben las familias del estado (prestaciones por desempleo y seguridad social) y del exterior.

El propio autor añade, además, otras deficiencias del modelo:

- Por un lado, afirma que no es totalmente correcto aplicar el mismo modelo a distintos tipos de municipios, con tipología económica diversa (ciudades “versus” grandes o pequeños pueblos; ámbitos urbanos o rurales; economías agrarias, industriales o de servicios, etc).
- Un modelo de renta disponible per capita basado en el consumo doméstico de electricidad tiende a sobrevalorar el dato estimado de renta en núcleos residenciales con alta tasa de segunda vivienda de recreo, en las que se produce un alto consumo de electricidad como consecuencia del funcionamiento de piscinas, depuradoras, sistemas de riego de jardines privados, etc. Efectivamente, en el caso de disponer de esta variable, resulta imposible separar el consumo de los empadronados del realizado por los no residentes.

➤ **J. Bachero, Jesús Esteban, Isabel López y Félix Ruiz (Universidad de Valencia)**

Este trabajo fue presentado en la XXIII Reunión de Estudios Regionales de la Asociación Española de Ciencia Regional, celebrado en Valencia, en el que se estimaba la renta familiar disponible de los municipios de la Comunidad Valenciana, en 1995. Para ello, parten de los datos provinciales publicados por el Banco de Bilbao-Vizcaya (actual BBVA). Su principal aportación consiste en la consideración de modelos de regresión diferentes para tres grupos de municipios, previamente clasificados a través de técnicas multivariantes (análisis cluster y discriminante), a partir de una serie de indicadores, fundamentalmente de actividad económica, procedentes del Anuario Comercial de España publicado por “la Caixa”. Se trata de las siguientes variables: población, Índice turístico, líneas telefónicas, oficinas de crédito, parque del total de vehículos y actividades comerciales.

Los autores pretenden recoger, de este modo, la heterogeneidad espacial propia de la distribución de la renta municipal, aunque no se especifica la forma funcional del/los modelo/s de regresión utilizados ni las variables explicativas que, de ser las mismas anteriormente expuestas, presentan una alta multicolinealidad.

A continuación, se presentan los principales resultados obtenidos por los autores:

Mayor renta personal por hab. (1995)		Menor renta personal por hab. (1995)	
Municipio	Renta (pta)	Municipio	Renta (pta)
Benicasim/Benicàssim (CA)	2.652.553	San Isidro (AL)	778.563
Peñíscola (CA)	2.536.193	Fortaleny (VA)	793.340
Oropesa del Mar (CA)	2.438.770	Alpuente (VA)	838.802
Teulada (AL)	2.399.492	Granja de Rocamora (AL)	843.904
Calpe/Calp (AL)	2.171.417	Rafelguaraf (VA)	887.691
Benissa (AL)	2.135.414	Tous (VA)	923.214
Altea (AL)	2.058.569	Benimodo (VA)	938.613
Denia (AL)	1.915.338	Llutxent (VA)	946.966
Poblets (els) (AL)	1.954.435	Montroy (VA)	952.652
Santa Pola (AL)	1.911.469	Xeresa (VA)	971.310

Como puede observarse, se producen las consencuencias lógicas de las estimaciones municipales de renta familiar disponible basadas en variables de actividad económica. Por un lado, como las diferencias intermunicipales en términos de actividad económica per capita son superiores a las diferencias en términos de riqueza de las familias (debido al poder redistributivo del Estado), la variable experimenta una gran dispersión, probablemente no correcta, que distancia excesivamente a los municipios con valores extremos de renta disponible. Efectivamente, la distancia entre los municipios de mayor y menor renta media supera, en la Comunidad Valenciana, el millón y medio de pesetas. Por otro lado, esta estimación beneficia excesivamente a municipios de mediano tamaño con importante actividad turística, como Benissa (Alicante/Alacant), situado por delante de núcleos residenciales como l'Elia (Valencia/València), probablemente con una población residente de mayor poder adquisitivo.

➤ **Antonio de las Heras, Carmen Murillo y Juan M. Rodríguez-Poo (Universidad de Cantabria)**

Estos investigadores presentaron una comunicación en el I Congreso de Economía Aplicada de Barcelona en la que se exponían dos métodos de estimación indirecta de la renta familiar disponible de todos los municipios españoles (excepto País Vasco y Navarra), que son continuación de la propuesta metodológica realizada por el primero de ellos, sólo para los municipios cántabros, en su tesis doctoral (Heras, 1992), cuya principal novedad estriba en la utilización, en esta ocasión, de información fiscal.

En concreto, de cara a la obtención de la renta familiar disponible per capita municipal de 1991, los autores proponen dos especificaciones, un modelo biecuacional y otro uniecuacional, a partir de dos variables fundamentalmente: las tasas de ocupación (sector primario, secundario y terciario) y el impuesto sobre la renta de las personas físicas (IRPF). Mientras que el primer modelo sólo permite la estimación de renta cada 5 años, coincidiendo con la publicación en censos y padrones del dato de población ocupada sectorial por municipios, el segundo modelo hace posible una estimación anual, a partir únicamente de la variable de impuestos.

1. Modelo biecuacional, en el que no se conoce, a priori, el tipo de relación existente en la primera ecuación entre la renta bruta disponible y la población ocupada:

$$\begin{cases} y_i = f(x_{1i}, x_{2i}, x_{3i}) + u_{1i} \\ I_i = \gamma_0 + \gamma_1 y_i + u_{2i} \end{cases} ; \quad (u_{1i}; u_{2i}) \approx N(0, \Sigma) \quad A.IX.9.$$

donde y_i : renta familiar disponible de la provincia i , excepto las provincias vascas y Navarra

- (Contabilidad Regional, INE)
- x_1 : % ocupados en el sector primario sobre total población (Censo, 1991)
- x_2 : % ocupados en el sector secundario sobre total población (Censo, 1991)
- x_3 : % ocupados en el sector terciario sobre total población (Censo, 1991)
- β, γ : parámetros a estimar
- I : IRPF per capita = (base imponible gravada media x nº declarantes x tipo efectivo medio cotización) / población residente
- u_{1i}, u_{2i} : perturbaciones aleatorias de las ecuaciones 1 y 2, para la provincia i .

A partir de aquí, los autores proponen un proceso de estimación de la renta disponible municipal a partir de un método de estimación no paramétrico que permita la estimación de la renta per capita en un municipio a partir de la renta per capita de la provincia con una distribución de la población ocupada (variable explicativa) lo más parecida posible a la de dicho municipio. Es decir, la renta bruta disponible de un municipio se obtiene como suma ponderada de las rentas de todas las provincias españolas, con un peso inversamente proporcional a la divergencia entre las distribuciones de la población ocupada en los sectores primario, secundario y terciario de ese municipio, y la de las diferentes provincias.

El resultado final permite establecer, para cada valor de renta municipal un intervalo de confianza, bajo el supuesto de igualdad entre la función de distribución de la renta en los ámbitos provincial y municipal.

2. Modelo uniecuacional:

$$y_i = h(b_i - I_i) + u_i \quad A.IX.10.$$

donde b_i : (base imponible gravada media x nº declarantes) / población residente

La estimación no paramétrica de este modelo permite la obtención de la renta disponible de los hogares en un determinado municipio como suma ponderada de las rentas de todas las provincias españolas, con ponderaciones que son inversamente proporcionales a la distancia existente entre la base imponible (b) menos el IRPF (I) del municipio cuya renta se desea estimar y la base menos el IRPF de cada una de las provincias, de la forma siguiente:

$$\hat{y}_j = \frac{\sum_{i=1}^N K\left(\frac{z_j - z_i}{h}\right) y_i}{\sum_{i=1}^N K\left(\frac{z_j - z_i}{h}\right)} \quad A.IX.11.$$

- donde \hat{y}_j : renta bruta disponible por habitante del municipio j
- z : base imponible menos IRPF, ($b - I$), para cada municipio j o provincia i
- h : parámetro de suavizado fijado mediante validación cruzada
- K : función de Kernel gaussiana, $K(u) = (1/\sqrt{2}) \cdot \exp[-(1/2)u^2]$, $-\infty < u < +\infty$
- y_i : renta bruta disponible por habitante de la provincia i

Una forma de validación del método consistiría en sumar los datos de renta de todos los municipios de una provincia y contrastar el resultado con el dato real de la renta provincial considerado en el modelo. De este modo, el orden de provincias, según el valor de la suma de renta bruta disponible estimada de cada uno de los municipios correspondientes, sería el siguiente:

Mayor renta disponible por hab. (1991)		Menor renta disponible por hab. (1991)	
Provincia	Renta (pta)	Provincia	Renta (pta)
Madrid	1.296.343	Badajoz	816.036
Barcelona	1.203.240	Cádiz	821.898
Zaragoza	1.136.536	Jaén	847.504
Girona	1.114.916	Almería	848.709
Balears (Illes)	1.112.401	Huelva	854.195
Tarragona	1.068.154	Cáceres	861.424
Burgos	1.066.338	Córdoba	862.655
Lleida	1.062.230	Ciudad Real	867.793
Rioja (La)	1.046.298	Málaga	869.357
Huesca	1.045.921	Granada	869.891

Los resultados estimados para las provincias difieren bastante de los datos reales ofrecidos por el INE, sobre todo en la ordenación de las provincias situadas en los primeros puestos (no tanto en los últimos), observándose una sobrevaloración de la renta municipal de provincias que, como Madrid, Barcelona y Zaragoza, son importantes centros generadores de actividad económica, dejando atrás a otras provincias, como Girona o Lleida que, por su alto nivel de renta disponible e inferior volumen de población, suelen ocupar los primeros puestos del ranking de la renta per capita.

Los resultados, en el ámbito municipal, pusieron de manifiesto una concentración de la renta bruta disponible por habitante en las capitales de provincia y municipios densamente poblados, así como en las zonas residenciales próximas, destacando, en cada provincia, estos municipios sobre el resto. Una crítica importante a este método sería su especial dificultad para detectar, a través de las variables explicativas, el peso de la economía sumergida en algunos municipios.

Por otro lado, en un nivel puramente metodológico, podría también afirmarse que el modelo bicuacional planteado adolece de problemas relacionados con una deficiente especificación, no resultando aceptable utilizar como única variable explicativa de la renta familiar disponible variables de ocupación, puesto que el nivel de empleo no es la única fuente de ingresos que perciben los hogares.

➤ **Jordi Arcarons, Joan Luria, Anna Tarrach y Carme Poveda (Departament d'Economia i Finances, Generalitat de Catalunya)**

En 1998, el Departamento de Economía y Finanzas (Generalitat de Catalunya) publica la estimación de la renta familiar disponible de las comarcas, capitales de comarca y municipios de más de 5.000 habitantes de Cataluña, para todos los años del período 1989-1995. Se trata de una propuesta alternativa a la realizada por el propio Instituto de Estadística de Cataluña, que estima por métodos directos similares al utilizado por SADEI en Asturias, la renta familiar disponible de las comarcas y municipios catalanes de más de 10.000 habitantes (Costa y Rovira, 1994).

En este caso, los autores defienden la utilización de variables explicativas derivadas del IRPF (como en el caso de Cantabria, aunque a través de un modelo diferente), debido fundamentalmente a la capacidad demostrada por este impuesto para identificar los diversos componentes de la riqueza

de los hogares, así como por la facilidad de obtención periódica anual de estos datos, que se producen con un desfase aproximado de año y medio de la realidad de las familias que controla.

Asimismo, los directores técnicos de la obra hacen también hincapié en la necesidad de superar las limitaciones que muestra la capacidad explicativa del IRPF: de un lado, se trata de un indicador más cercano, aunque no idéntico, a la renta personal (concepto que incluye las cotizaciones sociales y excluye las transferencias a los hogares) que a la renta bruta disponible o renta familiar disponible y, de otro, la incapacidad del impuesto de recoger algunas rentas de hogares agrarios o el fraude fiscal, así como las rentas mínimas exentas del impuesto. Para paliar estas deficiencias, se proponen las siguientes variables derivadas del IRPF:

- Rendimiento total del IRPF (RT), que recoge el nivel de renta agregada de los declarantes.
- Porcentaje del rendimiento total del IRPF representado por los rendimientos empresariales y profesionales (PRT), considerado como un buen indicador del fraude fiscal.
- Porcentaje de declarantes sobre la población de derecho (PDEC), que permite sacar a la luz el número de personas exentas de declaración del IRPF.

La disponibilidad de información espacio-temporal de la renta familiar disponible regional, en el período 1989-1993, permite la especificación de un modelo exponencial de datos de panel de efectos aleatorios, fácilmente linealizables por transformación logarítmica, que permite, a su vez, la interpretación de los estimadores como elasticidades. En concreto, se trata del siguiente modelo:

$$\log y_{it} = \beta_0 + \beta_1 \log RT_{it} + \beta_2 \log PRT_{it} + \beta_3 \log PDEC_{it} + v_{it} \quad A.IX.12.$$

$$\bar{R}^2 = 0,99$$

- siendo y_{it} : renta familiar disponible (BBV) de la comunidad autónoma i en el año t
 i : comunidad autónoma: del total de comunidades, se excluyen País Vasco, Navarra, Ceuta y Melilla, para las que no se dispone de las variables del IRPF, por lo que $i = 1, \dots, 14$
 t : año: se excluye el período anterior al año 1989 debido a diferencias significativas en el cálculo del IRPF, fundamentalmente la posibilidad para los cónyuges de declarar individual o conjuntamente, por lo que $t = 5$
RT: rendimiento total
PRT: % rendimientos empresariales y profesionales sobre el rendimiento total
PDEC: % declarantes sobre población total
 u_{it} : perturbación aleatoria de la comunidad autónoma i y año t .
 $u_{it} = v_{it} + \mu_i$, donde v_{it} es una perturbación aleatoria esférica y μ_i un término que representa los efectos individuales de cada comunidad autónoma

Los autores contrastan también otros modelos de datos de panel, sin efectos y con efectos fijos, demostrando la mayor fiabilidad de la estimación por efectos aleatorios (por el método de mínimos cuadrados generalizados), a partir de contrastes habitualmente propuestos en la literatura econométrica (ver, por ejemplo, Baltagi, 1995; Breusch-Pagan, 1979; White, 1980; Hausman, 1978; etc.). La aplicación de los coeficientes estimados a través de este modelo a los valores de las variables explicativas correspondientes a las comarcas y municipios de más de 5.000 habitantes catalanes, de 1989 a 1995, permite conocer la renta familiar disponible de estas unidades. Los

principales resultados en los municipios catalanes de más de 5.000 habitantes, para el último año disponible, 1995, fueron los siguientes:

Mayor renta disponible por hab. (1995)		Menor renta disponible por hab. (1995)	
Municipio	Rta (mil.pta)	Municipio	Rta (mil.pta)
Matadepera (BA)	3.141	Constantí (TA)	1.035
Alella (BA)	2.577	Santa Margarida de Montbui (BA)	1.056
Sant Cugat del Vallès (BA)	2.446	Santa Coloma de Gramenet (BA)	1.101
Sant Just Desvern (BA)	2.442	Sant Vicenç de Castellet (BA)	1.130
Sant Andreu de Llavaneres (BA)	2.386	Cornellá del Vallès (BA)	1.156
Tiana (BA)	2.234	Vilanova del Camí (BA)	1.165
Premià de Dalt (BA)	2.142	Navàs (BA)	1.189
Sitges (BA)	2.037	Sant Boi de Llobregat (BA)	1.172
Sant Quirze del Vallès (BA)	2.002	Hospitalet de Llobregat (L')	1.173
Vilassar de Mar (BA)	1.906	Salt (GI)	1.174

Los autores también analizan la serie dinámica de renta per capita municipal, poniéndose de manifiesto un proceso de convergencia en términos de renta al experimentar los municipios más retrasados en 1989 (Cervera, Alcanar, Sant Joan Despí, Flix, Gadesa, Deltebre y Gironella), mayores crecimientos en el período considerado, que el experimentado por los más adelantados por su riqueza (Matadepera, Alella, Sant Andreu de Llavaneres, Sant Just Desvern, Corbera de Llobregat y Vielha i Mijaran).

➤ Consejo Económico y Social de la Región de Murcia

Los autores de este estudio estiman la renta familiar disponible municipal de la Región de Murcia referida a los años 1987 y 1996, realizando también un análisis de evolución de renta. Para ello, realizan una inferencia ecológica de la renta disponible provincial, publicada por el Servicio de Estudios del Banco de Bilbao-Vizcaya (hoy BBVA), a partir de indicadores, disponible en el ámbito municipal, procedentes del Anuario del Mercado Español (AME), de Banesto (para 1987) y el Anuario Comercial de España, de “la Caixa” (para 1996). La especificación del modelo, supone parámetros constantes en toda la muestra provincial, es la siguiente:

$$Rta_i = \hat{\beta}_1 \underset{(6,05)}{Telef_i} + \hat{\beta}_2 \underset{(3,31)}{Banco_i} + \hat{\beta}_3 \underset{(3,55)}{IVA_i} + \hat{\beta}_4 \underset{(4,34)}{Ocup_i} \quad \bar{R}^2 = 0,86 \quad A.IX.13.$$

donde Rta: renta bruta disponible por habitante provincial (BBV)
 Telef: líneas telefónicas por habitante, corregidas del peso de la segunda vivienda
 Banco: oficinas de crédito por habitante
 IVA: valor del IVA repercutido respecto al número de declarantes
 Ocup: tasa de ocupación

Los autores confían en esta especificación porque los coeficientes estimados fueron significativos (pese al problema de multicolinealidad presente en la matriz de regresores) y el coeficiente de determinación resulta aceptable. También hacen mención de la bondad del test de autocorrelación temporal Durbin-Watson que, en el contexto espacial de la estimación, no tiene validez.

Los resultados obtenidos ponen de manifiesto, una vez más, el sesgo a favor de localidades de mediano tamaño con alguna importante actividad económica, el turismo, en perjuicio de grandes ciudades, como Murcia y Cartagena. Pese a la corrección efectuada sobre la variable teléfonos, este sesgo es debido a la influencia en el modelo de otros indicadores, como oficinas bancarias y ocupados, especialmente elevados en localidades turísticas, donde se concentran las entidades de crédito y existe una mayor ocupación de carácter temporal, no bien remunerado.

Mayor renta disponible por hab. (1996)		Menor renta disponible por hab. (1996)	
Municipio	Renta (pta)	Municipio	Renta (pta)
San Javier	1.518.511	Ricote	810.319
San Pedro del Pinatar	1.333.445	Blanca	834.564
Alcázares (Los)	1.239.329	Ceutí	871.115
Yecla	1.205.277	Cieza	871.711
Torre Pacheco	1.194.002	Villanueva	875.690
Fuente Álamo	1.183.796	Calasparra	887.775
Ulea	1.163.386	Unión (La)	897.473
Mazarrón	1.167.427	Fortuna	898.613
Molina de Segura	1.165.860	Pliego	946.175
Cartagena	1.164.771	Albudeite	977.077

➤ Instituto Lawrence R. Klein (Universidad Autónoma de Madrid)

El Instituto Lawrence R. Klein elabora, desde 1992, la variable de renta disponible de los hogares para la totalidad de municipios españoles, que ha sido publicada en diversas obras (Atlas Comercial de España 1994, Atlas Comerciales para diversas comunidades autónomas y, desde 1996, Anuario Económico de España). La última actualización se encuentra publicada, escalada en 10 niveles y solo para los municipios de más de mil habitantes, en el Anuario Económico de España 2001, publicado por “la Caixa”, y se refiere al año 1999.

El proceso de estimación parte de un modelo de datos de panel, que considera como variable a explicar las series de renta bruta disponible de los hogares, publicadas en la Contabilidad Regional del INE, correspondientes a las provincias españolas durante el período 1986-1995 (en el momento de la estimación, los autores no disponían de los nuevos datos de renta provincial obtenidos a partir del Sistema Europeo de Cuentas, SEC-95). Se trata, en concreto, de un modelo dinámico de datos de panel de efectos fijos, estimado por de mínimos cuadrados generalizados (MCG), al haberse contrastado la existencia de heteroscedasticidad contemporánea en la perturbación aleatoria, que permitirá la estimación-predicción de la renta bruta disponible de las provincias en el año 1999, a partir de un conjunto de 16 variables explicativas, previamente agrupadas, a través de la técnica de análisis factorial, en 4 factores (con una información del 77% de la varianza total):

$$y_{it} = \alpha_i + \rho y_{it-1} + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + u_{it} \quad \bar{R}^2 = 0,99 \quad A.IX.14.$$

siendo y_{it} : renta bruta disponible por habitante provincial (INE)
 x : factores explicativos: 1) empleo, 2) ahorro, 3) consumo y turismo, 4) industria.
 β , ρ : parámetros de la regresión ; u : perturbación aleat.
 i : provincias españolas, excepto Ceuta y Melilla; t : período 1986-1995

Una vez estimado el valor de renta bruta disponible provincial, se ajusta el mismo al valor nacional de dicha variable, en 1999, publicado por el INE en la Contabilidad Nacional, de forma que la suma

de la renta de las provincias dé como resultado el dato de renta bruta disponible de los hogares del total nacional, siendo el coeficiente de ajuste aplicado igual a 0,03 (3% de desviación).

A continuación, se plantea un modelo de estructuras espaciales (tipo “switching regression”), estimado por MCO robustos, debido a la existencia contrastada (a partir del test de Chow espacial) del efecto de heterogeneidad espacial en el mismo, debido al diferente comportamiento de las provincias con un nivel inferior en renta disponible de los hogares (zona A, con 20 provincias del Sur-Oeste retrasado) y aquéllas con mayor riqueza (zona B, 30 provincias del Norte-Este próspero).

Este modelo, que considera coeficientes diferentes en cada zona de renta, explica la renta per capita provincial de 1999 a través de un grupo de 20 variables explicativas, cuyo comportamiento en el ámbito municipal se ha comprobado como eficiente, previamente agrupadas en 4 factores a partir de un análisis factorial (81% de la varianza total), de la forma siguiente:

$$y_i = \beta_{0A} + \beta_{1A} x_{1i} + \beta_{2A} x_{2i} + \beta_{3A} x_{3i} + \beta_{4A} x_{4i} + \beta_{0B} + \beta_{1B} x_{1i} + \beta_{2B} x_{2i} + \beta_{3B} x_{3i} + \beta_{4B} x_{4i} + w_i \quad \bar{R}^2 = 0,87 \quad A.IX.15.$$

siendo y_i : renta bruta disponible por habitante provincial (Klein, 1999)
 x_i : factores explicativos a partir de 20 variables (81% de la varianza total):
 1) paro, 2) consumo y turismo, 3) ahorro y empleo cualificado,
 4) incultura y ruralidad.
 $\beta_{0A}, \dots, \beta_{4B}$: coeficientes a estimar
 w_i perturbación aleatoria, para las “i” provincias: $w_i = u_{iA} + u_{iB}$.
 A: 20 provincias del Sur-Oeste retrasado; B: 30 provincias del Norte-Este próspero

La predicción-extrapolación de la renta bruta disponible de los municipios españoles se lleva a cabo a través de la aplicación, sobre los valores municipales de las variables explicativas, de los coeficientes estimados, según la pertenencia de los municipios a la zona A o B. A continuación, se presentan algunos de los resultados obtenidos, en concreto, para los municipios de la Comunidad de Madrid, presentando los casos extremos, los municipios con mayor y menor nivel de renta per capita. Estos resultados son también previamente ajustados de forma que la suma de la renta bruta municipal dé como resultado el dato provincial, estimado para 1999, de dicha renta. Los coeficientes de ajuste son, en este caso, diferentes para cada provincia, oscilando, en términos absolutos, entre el 2% y el 20%.

Mayor renta disponible por hab. (1999)		Menor renta disponible por hab. (1999)	
Municipio	Rta (mil.pta)	Municipio	Rta (mil.pta)
Majadahonda	2.301	Prádena del Rincón	1.200
Boadilla del Monte	2.300	Madarcos	1.265
Rozas de Madrid (Las)	2.279	Cenicientos	1.295
Villanueva de la Cañada	2.272	Hiruela (La)	1.312
Hoyo de Manzanares	2.261	Torremocha de Jarama	1.317
Pozuelo de Alarcón	2.250	Valdaracete	1.343
Galapagar	2.100	Braojos	1.364
Tres Cantos	2.100	Garganta de los Montes	1.369
Collado Mediano	2.099	Villamanrique de Tajo	1.372
Colmenarejo	2.066	Puebla de la Sierra	1.380

El método seguido por el Instituto L.R. Klein corrige, por un lado, el efecto de algunas variables que, además de ser indicadoras de renta de los hogares también recogen rentas propias de actividad

económica. El caso más significativo, por su peso en el modelo, es teléfonos por habitante, que se corrige, a partir de información disponible en el Instituto, para obtener una aproximación de las líneas de uso doméstico exclusivamente. Además, se aplica otro coeficiente correcto para evitar el desmesurado peso que adquiere esta variable en municipios turísticos en los que existe una gran proporción de segunda vivienda, entendiéndose que el impacto que ejerce esta actividad sobre la renta de los residentes, nunca es proporcional al recogido por la variable. Por último, se introduce en la estimación el ratio que expresa la proporción de actividad económica profesional (profesiones liberales) sobre el total de actividad económica, para detectar los municipios residenciales en los que no existe normalmente una gran actividad económica per capita, pero que concentran residentes con alto nivel de vida, muchos de ellos ejerciendo profesiones liberales (abogados, arquitectos, economistas, médicos, etc.).

➤ **Pedro Nogueira Ramos (Instituto Nacional de Estadística de Portugal)**

El Instituto de Nacional de Estadística de Portugal ha publicado, desde 1993, cuatro estimaciones de la renta per capita municipal de este país, con el nombre Índice del poder de compra, utilizando el análisis factorial como herramienta estadística principal. La última estimación, realizada el año 2000 y referida a 1998, parte de una selección de 18 variables (tras el análisis de un total de 70), explicativas del poder de compra de la población, más desde una óptica de producción o generación de rentas que de utilización de las mismas. Se trata de los siguientes indicadores: impuesto de vehículos, consumo doméstico de electricidad, teléfonos, accesos RDSI, ocupados en comercio, precio vivienda urbana, depósitos y créditos bancarios, IRPF, hipotecas, densidad de población, contribución municipal, empresas y ocupados de hostelería.

Estas variables son empleadas en un análisis factorial por componentes principales, seleccionándose el primer factor como Índice del poder de compra, que recoge un 52,5% de la varianza total. El segundo factor es también utilizado como Factor de dinamismo relativo, entendido como la dinámica comercial y económica que subyace y hace posible el primero de los factores, el cual está fuertemente relacionado con la actividad turística (absorbe un 23,1% de la varianza total). El cálculo del indicador del poder de compra municipal se realiza finalmente atribuyendo el valor 100 al valor obtenido por el primer factor para todo el país.

Además, se propone por primera vez, en esta edición, una primera estimación de la variación del índice de poder de compra municipal en el período 1995-1998, tomando como referencia los datos estimados para este índice en una publicación anterior. Los autores piden que sea interpretado con prudencia, sobre todo por encontrarse medidas algunas variables explicativas en términos nominales (escudos corrientes de cada año) y no poder separar del índice el efecto de los precios.

Los principales resultados del Índice de poder de compra de los municipios portugueses, en 1998, son los siguientes:

Mayor índice de poder de compra, base=100 (1998)		Menor índice de poder de compra, base=100 (1998)	
Municipio	Índice	Municipio	Índice
Lisboa (Lisboa y V.Tajo-Grande Lisboa)	305	Calheta (Madeira)	33
Porto (Norte-Grande Porto)	239	Celorico de Basto (Norte-Tâmega)	34
Oeiras (Lisboa y V.Tajo-Grande Lisboa)	164	Sernancelhe (Norte-Douro)	34
Cascais (Lisboa y V.Tajo-Grande Lisboa)	149	Santana (Madeira)	34
São João da Madeira (Norte-Douro e V.)	149	Ribeira de Pena (Norte-Tâmega)	34
Faro (Algarbe)	139	Terras de Bouro (Norte-Cávado)	34
Coimbra (Centro-Baixo Mondego)	137	São Vicente (Madeira)	35

Mayor índice de poder de compra, base=100 (1998)		Menor índice de poder de compra, base=100 (1998)	
Municipio	Índice	Municipio	Índice
Amadora (Lisboa y V.Tajo-Grand.Lisboa)	131	Boticas (Norte-Alto Tras-os-Montes)	35
Aveiro (Centro-Baixo Vouga)	127	Cinfães (Norte-Tâmega)	35
Almada (Lisboa y V.Tajo-Penins.Setúbal)	124	Alcoutim (Algarbe)	36

➤ **José María Otero, Antonio Fernández y Fernando Isla (Instituto de Estadística de Andalucía).**

El Instituto de Estadística de Andalucía publica (en 1999) la estimación de renta familiar disponible para las comarcas y municipios andaluces de más de 10.000 habitantes, referida al año 1994, mediante un modelo que parte de la renta bruta disponible de los hogares procedente de la Contabilidad Regional de España (INE). Para el resto de municipios de esta comunidad autónoma, los directores técnicos de la obra ofrecen el estrato de renta familiar en el que se encuentran, dividiendo la distribución de la variable en 7 niveles.

La novedad que constituye este modelo consiste en el proceso econométrico riguroso que siguen los autores para superar problemas propios de toda estimación y, en concreto, de la predicción espacial: elección de forma funcional, multicolinealidad, heteroscedasticidad, estabilidad estructural, orden de causalidad y tratamiento de la información extramuestral (predicción espacial). En cuanto a la forma funcional, se decantan por la función lineal debido a su simplicidad, al no haberse aún demostrado otra forma más eficaz. La presencia de multicolinealidad no constituye un problema grave cuando el modelo se estima con fines de predicción, siempre y cuando las pautas de correlación entre los regresores en el ámbito de predicción no difieran significativamente de las que se presentan en la muestra espacial agregada, lo que difícilmente se cumple en la práctica, por lo que los autores recurren al método de análisis factorial de las variables explicativas. Por su parte, el problema de heteroscedasticidad es solucionado a través de la estimación por el método de mínimos cuadrados ponderados siempre y cuando se ponga de manifiesto su presencia, a través de contrastes como el test de White. Además, se aconseja el contraste de estabilidad estructural en los parámetros del modelo mediante un test de Chow aplicado a tres grupos de provincias según su nivel de renta per capita (alta, media, baja). En cuanto al orden de causalidad de las variables, los autores ponen de manifiesto que suele estar invertido al considerar variables de consumo como explicativas de la renta cuando debería ser al contrario. Este último problema se suaviza un poco en modelos en los que sólo interviene una variable explicativa, siempre y cuando el coeficiente de determinación esté próximo a la unidad, pues la relación existente entre las predicciones realizadas por los modelos Y respecto de X y X respecto de Y (en desviaciones a la media) es igual al cuadrado del coeficiente de correlación entre ambas variables. Esta solución, que implica la estimación del modelo con variables en desviaciones a la media, se propone también como solución al problema de heterogeneidad espacial que se produce en toda estimación indirecta.

El modelo de regresión lineal de la renta per capita provincial, en desviaciones a la media, fue finalmente estimado por MCO, al haber sido contrastado homoscedástico, de la siguiente forma:

$$\hat{y}_i = \beta_0 + \beta_1 F_{1i} + u_i \quad \bar{R}^2 = 0,78 \quad A.IX.16.$$

siendo \hat{y}_i : renta familiar disponible (INE) por habitante de la provincia i , de 1994, en desviaciones a la media (valor nacional de renta familiar disponible por habitante)
 F_1 : primer factor del análisis de componentes principales (49% de variación) de las siguientes variables (relativizadas por la población de hecho de 1994): oficinas

bancarias, base imponible IRPF, oficinas de cajas de ahorros, camiones, consumo de energía eléctrica, licencias IAE de actividades empresariales, líneas telefónicas, tractores, turismos.

i: provincias españolas ($i=1, \dots, 50$)

u_i: perturbación aleatoria de la provincia i.

Las predicciones de la renta familiar disponible municipal se obtienen en un proceso que se inicia con la reparametrización del modelo provincial inicial, sustituyendo el factor explicativo por su expresión correspondiente como función lineal de las 9 variables explicativas estandarizadas (en desviaciones a la media), de forma que, para cada provincia i el modelo tendría la forma siguiente:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^9 \alpha_j \hat{x}_{ji} + u_i \quad A.IX.17.$$

A continuación, para cada provincia i se obtiene el dato de renta disponible por habitante de cada uno de sus correspondientes municipios, j(i), a partir del valor estimado de la variable endógena, en desviaciones a la media, que es el valor de renta disponible per capita provincial del INE (y_i), del modo siguiente:

$$\hat{y}_{j(i)} = \sum_{k=1}^9 \alpha_k \hat{x}_{kj(i)} \Rightarrow \hat{y}_{j(i)} = \hat{y}_{j(i)} + y_i \quad A.IX.18.$$

De este modo, es posible considerar tantos modelos como provincias se consideren inicialmente, todos ellos con los mismos coeficientes de regresión pero con distintas ordenadas en el origen, una por cada valor de renta per capita provincial. Esta forma de operar permite considerar toda la información disponible eficientemente, según los autores.

Los resultados más importantes correspondientes a la renta familiar disponible por habitante de 1994, correspondiente a los municipios andaluces de más de 10.000 habitantes son:

Mayor renta disponible por hab. (1994)		Menor renta disponible por hab. (1994)	
Municipio	Rta (mil.pta)	Municipio	Rta (mil.pta)
Torremolinos (MA)	1.328	Álora (MA)	512
Fuengirola (MA)	1.239	Archidona (MA)	549
Benalmádena (MA)	1.222	Barbate de la Frontera (CA)	637
Marbella (MA)	1.178	Puebla del Río (SE)	639
San Roque (CA)	1.019	Arcos de la Frontera (CA)	643
Nerja (MA)	1.155	Cártama (MA)	651
Rota (CA)	1.151	Coín (MA)	660
Roquetas de Mar (AL)	1.146	Alhaurín el Grande (MA)	664
Alcalá de Guadaira (SE)	1.102	Íllora (GR)	669
Albolote (GR)	1.082	Medina-Sidonia (CA)	674

Pese al buen análisis realizado, los autores no contrastan de forma adecuada (a través de técnicas de econometría espacial) algunos problemas como la inestabilidad paramétrica (heterogeneidad espacial), ni consideran especificaciones basadas en cuestiones geográficas, como la continuidad. Obviamente, tampoco se contrastan las consecuencias de autocorrelación espacial en el modelo inicial, aunque la solución aportada de estimación en desviaciones a la media resulta interesante y en la línea de soluciones que se han propuesto para las inferencias ecológicas en la literatura.

Por último, señalar que la mezcla realizada de variables explicativas de renta (IRPF, automóviles, entidades de crédito) y producción económica (energía eléctrica, actividades empresariales y licencias comerciales) puede distorsionar los resultados a favor de municipios de tamaño mediano con alguna importante actividad (por ejemplo, el turismo en municipios de la Costa del Sol), en perjuicio de las capitales de provincia, cuyo valor de renta per capita resultaría posiblemente infravalorado.

➤ **Fernando del Castillo, Ángel de Santos, Javier Horcajo, Teresa Fernández, Agustín Martín, Susana Gómez y Javier Mateo (Instituto de Estadística de la Comunidad de Madrid)**

El Instituto de Estadística de la Comunidad de Madrid (IECAM), publica en el año 2001, un CD-ROM en el que se incluyen los valores de renta familiar disponible por habitante de los municipios de esta Comunidad, en el período 1994-1997 (ya publicados en ediciones anteriores), así como una novedad: el dato de renta per capita para las secciones censales y, por tanto, por agregación, para los distritos y barrios de Madrid capital. Para ello, se ha escogido un modelo único basado en la explotación del IRPF municipal, como en Arcarons *et al.* (1998) y, en esta ocasión también, por secciones censales. En concreto, la base imponible del IRPF ha sido estimada en un 85% del total de la renta familiar disponible en la Comunidad de Madrid, por lo que se considera adecuada como indicador. Sin embargo, aunque las relaciones entre ambas variables parecen muy estrechas, los mismos autores advierten del peligro de utilizar estos datos sin una corrección adecuada, debido a los siguientes problemas:

- En el IRPF hay rentas que están mejor recogidas que otras. Así, en el primer caso se encontrarían las rentas del trabajo, rentas estables o rentas industriales, frente a las rentas mixtas o de capital, rentas esporádicas o rentas agrarias.
- Con independencia del fraude fiscal, algunas rentas no se contabilizan, como las rentas mínimas (inferiores a cierto nivel).
- La base imponible no es un concepto totalmente equivalente a la renta disponible.
- La asignación municipal del titular de las rentas en el IRPF no coincide, en todos los casos, con el concepto de residente que se utiliza en estadística (a partir de los datos del Padrón municipal).

Por eso, los directores de la obra proponen, para la estimación municipal, la corrección de los datos procedentes del impuesto con otros indicadores. Para ello, se estima un indicador sintético a partir del siguiente grupo de variables:

1. Rendimiento medio del IRPF por declarante.
2. Porcentaje de declarantes sobre la población total, para ajustar el rendimiento medio a la población residente y corregir la asignación territorial.
3. Porcentaje de declarantes sobre la población de 15 y más años, para aproximar los declarantes a la población susceptible de declaración y considerar las rentas mínimas exentas.

4. Porcentaje de rentas declaradas por conceptos distintos del trabajo, para aproximar las rentas peor recogidas por el IRPF.
5. Indicador sintético sobre la condición socioeconómica de los municipios de la Comunidad de Madrid, aplicando análisis factorial a varios indicadores indirectos, como las tasas de universitarios, directivos y técnicos, turismos por habitante, etc.

Sin explicar con detalle el modelo, el indicador sintético final se relaciona matemáticamente con el dato de renta bruta disponible de los hogares procedente de la Contabilidad Regional del INE, para estimar a continuación la renta municipal, que se obtiene en pesetas corrientes de cada año. Por su parte, la estimación por secciones censales, distritos y barrios de la capital está referida a 1997 y se basa únicamente en la información derivada del IRPF, disponible al nivel de sección censal.

Los principales resultados de la estimación de renta per capita municipal, para 1997, es la siguiente:

Mayor renta disponible por hab. (1997)		Menor renta disponible por hab. (1997)	
Municipio	Rta (mil.pta)	Municipio	Rta (mil.pta)
Pozuelo de Alarcón	2.563	Madarcos	972
Rozas de Madrid (Las)	2.378	Acebeda (La)	977
Boadilla del Monte	2.366	Brea de Tajo	978
Villanueva de la Cañada	2.333	Berzosa del Lozoya	985
Majadahonda	2.327	Horcajuelo de la Sierra	993
Torrelodones	2.260	Valdaracete	996
Tres Cantos	2.158	Canencia	1.007
Valdeolmos-Alalpardo	2.074	Hirueta (La)	1.008
Cobeña	2.010	Cenicientos	1.020
Venturada	1.990	Villaconejos	1.022

El IRPF, aunque corregido, no recoge adecuadamente las rentas de los hogares agrarios ni la economía sumergida, por lo que algunos municipios aparecen muy infravalorados en renta, y otros, debido al ajuste realizado con el dato de renta per capita de la Comunidad procedente de la Contabilidad Regional, están bastante sobrevalorados, con valores medios superiores a 2.300.000 pesetas, produciéndose mayores diferencias de renta de las realmente existentes.

➤ Departamento de Estadística del Ayuntamiento de Madrid

En junio de 2000, el Ayuntamiento de Madrid publicó los datos de renta familiar disponible por habitante de los distritos y barrios del municipio de Madrid, referida a 1996. Esta estimación parte de la metodología utilizada por el Instituto de Estadística de la Comunidad de Madrid (IECAM), aunque con algunas diferencias, ante la imposibilidad, en aquel momento, de encontrar información relativa al IRPF desagregada por ámbitos territoriales equiparables a los distritos y barrios de la capital. Efectivamente, dado que los datos del impuesto sólo estaban disponibles al nivel de las Delegaciones de Hacienda, ámbito espacial diferente de los distritos y barrios, se optó por un modelo de regresión múltiple que, a partir del dato de renta familiar disponible estimado por el IECAM para los municipios de la Comunidad, permitiese encontrar una relación funcional entre dicha renta con otras variables de la que se dispusiese de información para los ámbitos de distrito y barrio.

En concreto, se seleccionaron como posibles variables explicativas del modelo, aun sabiendo que la información recogida por algunas de ellas era redundante, las siguientes: tasas de actividad y paro,

tasas de analfabetismo, estudios primarios y universitarios, tasas de directivos, técnicos, trabajadores no cualificados, tasa de empresarios y tasas de trabajadores fijos, eventuales. A estas variables de carácter sociodemográfico, se añadió una variable más que ha venido empleándose en estudios similares, vehículos por habitante, que suele emplearse como un indicador de actividad económica.

El modelo finalmente seleccionado relaciona el dato provisional de renta familiar disponible por habitante municipal, propuesto por el IECAM para 1996, correspondiente a los 41 municipios de la Comunidad mayores de 5.000 habitantes (los municipios de menos población son entidades con estructuras socioeconómicas y demográficas poco comparables con las de los distritos madrileños), en función de dos variables explicativas, de la forma siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad \bar{R}^2 = 0,94 \quad A.IX.19.$$

siendo y_i : renta familiar disponible del municipio i , para $i = 1, \dots, 41$ municipios mayores de 5.000 habitantes de la Comunidad de Madrid (IECAM)
 x_1 : proporción de universitarios sobre población activa de 16 y más años (Padrón 1996)
 x_2 : vehículos por 1000 habitantes (Impuesto de vehículos de tracción mecánica).
 u_i : perturbación aleatoria del municipio i .

La aplicación de este modelo a los valores de las variables explicativas de los distritos y barrios de Madrid permite obtener la estimación de renta buscada.

Mayor renta disponible por hab. (1996)		Menor renta disponible por hab. (1996)	
Barrio (Distrito)	Rta (mil.pta)	Barrio (Distrito)	Rta (mil.pta)
Plantío (El) (Moncloa-Aravaca)	2.656	Orcasur (Usera)	1.034
Nueva España (Chamartín)	2.534	Entrevías (Puente de Vallecas)	1.060
Viso (El) (Chamartín)	2.510	Butarque (Villaverde)	1.094
Niño Jesús (Retiro)	2.495	Orcasitas (Usera)	1.115
Fuentalarreina (Fuencarral-El Pardo)	2.480	Casco H.Vallecas (Villa de Vallecas)	1.129
Vallehermoso (Chamberí)	2.420	Goloso (El) (Fuencarral-El Pardo)	1.130
Recoletos (Salamanca)	2.404	Portazgo (Puente de Vallecas)	1.132
Jerónimos (Retiro)	2.402	Arcos (San Blas)	1.138
Cdad.Universitaria (Moncloa-Aravaca)	2.398	San Cristóbal (Villaverde)	1.140
Atalaya (Ciudad Lineal)	2.391	Aeropuerto (Barajas)	1.142

Aunque los autores realizan un análisis de la distribución de los errores del modelo (normalidad, heteroscedasticidad), al no aplicar técnicas de econometría espacial, no abordan el problema de autocorrelación espacial realmente existente en el territorio madrileño, por lo que los coeficientes estimados pudieran estar sesgados. Por otro lado, al tomar como variable endógena el dato de renta per capita estimado por el IECAM, las estimaciones efectuadas para los distritos y barrios de la capital adolecen de los mismos problemas de sobrevaloración de las rentas más elevadas e infravaloración de las inferiores. Quizá por este motivo, los autores afirman que se produce en el interior del municipio una “enorme disparidad” entre el distrito más rico (Chamartín), que duplica el nivel de renta per capita obtenido por el más pobre (Vicálvaro), existiendo entre ambos una diferencia de renta per capita superior al millón de pesetas. Este tipo de situaciones aconsejan la utilización de intervalos de renta o niveles, que suaviza estas diferencias de renta en términos absolutos, solución que también se adopta en este estudio, que adopta los estratos de renta propuestos en el Anuario Comercial de España 2000, elaborado por el Instituto Lawrence R. Klein y patrocinado por “la Caixa”.

➤ **Ángel Alañón Pardo (Universidad Complutense de Madrid)**

El autor expone en su tesis doctoral un modelo de estimación de la renta familiar disponible por habitante para las comarcas y municipios de España, referida al año 1991. Junto con la estimación llevada a cabo por el Instituto L. R. Klein, que considera el problema de heterogeneidad espacial estimando un modelo de estructuras o regímenes espaciales, esta estimación también incorpora los métodos de econometría espacial en el proceso de estimación de renta per capita municipal.

El autor propone un modelo de regresión lineal en el que la renta disponible per capita de las provincias españolas está en función de tres variables explicativas: locales activos, edificios con 4 ó más plantas y población con estudios en curso de tercer grado, todas ellas relativizadas por la población. Tras una primera estimación del modelo por MCO, contrasta en el mismo la presencia de dependencia y/o heteroscedasticidad espacial, utilizando para ello tres especificaciones de la matriz de pesos espaciales: una matriz de contigüidades o interacciones espaciales, matriz de la distancia inversa y matriz de umbrales kilométricos. La existencia contrastada de dependencia sustantiva es lo que suele motivar la estimación de un modelo del retardo espacial (“spatial lag model”):

$$y_i = \rho W y_i + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad A.IX.20.$$

siendo y_i : renta bruta disponible de los hogares de la provincia i (INE)
 W : matriz de pesos o ponderaciones espaciales (para la que se han propuesto tres especificaciones)
 ρ : coeficiente autorregresivo espacial a estimar
 β : parámetro a estimar
 x_1 : locales activos por habitante
 x_2 : edificios con 4 ó más plantas por habitante
 x_3 : población con estudios en curso de tercer grado por habitante
 u : perturbación aleatoria

Sin embargo, tal como el autor pone de manifiesto, resulta imposible calcular, en cada municipio, el valor de la variable endógena espacialmente retardada, dado que precisamente es la renta disponible municipal el objetivo de la estimación. Por este motivo, se recomienda utilizar el método de filtrado espacial que elimine el efecto de dependencia en el modelo provincial y asegure unos estimadores insesgados, de la forma siguiente:

$$\hat{y} = (I - \rho W)^{-1} X \beta$$

$$\hat{y}_i = \hat{\beta}_0 (I - \rho W)^{-1} + \hat{\beta}_1 (I - \rho W)^{-1} x_{1i} + \hat{\beta}_2 (I - \rho W)^{-1} x_{2i} + \hat{\beta}_3 (I - \rho W)^{-1} x_{3i} \quad A.IX.21.$$

El autor, aunque conoce y aplica técnicas de econometría espacial, sorprende un poco por la selección que hace de unas variables explicativas sin antecedentes en la literatura, más ligadas a la producción que al poder adquisitivo de los hogares, máxime cuando escoge como período de estimación el año 1991, para el que dispone de un mayor volumen de información, por ser año censal. Por otro lado, queda a medio camino en su análisis al limitarse a reconocer los problemas que ocasionan las inferencias ecológicas sin intentar solucionarlos. Podría haberse incorporado la metodología propuesta por Otero y Fernández (1998) que, aunque no definitiva, ayuda a paliar esta dificultad.

➤ **Coro Chasco Yrigoyen (Instituto Lawrence R. Klein - UAM)**

La autora realizó, en 2001, una estimación del VAB de los municipios de la Comunidad de Madrid, referida al año 1999, que fue publicada por el Consejo Económico y Social (CES) de la Comunidad de Madrid, en un capítulo del informe “Situación Económica y Social de la Comunidad de Madrid 1999”. Dicha estimación fue llevada a cabo en las siguientes etapas:

1. Obtención de una relación matemática entre el VAB y el total de impuestos IAE de la Comunidad de Madrid, para 1999, a partir de un modelo de datos de panel SUR espacial, estimado por el método de máxima-verosimilitud (MV), en el que los coeficientes de la variable explicativa, β_i , son constantes en cada comunidad autónoma y varían para cada período temporal considerado (período 1996-99).

$$y_{it} = x_{it} \beta_i + \varepsilon_{it} \quad A.IX.22.$$

- y_{it} : VAB total per capita de la región i en el período t , para $i=1,\dots,17$; $t=1,\dots,4$ (Contabilidad Regional del INE).
 x_{it} : impuesto IAE per cápita de la región i en el período t (Instituto L. R. Klein).
 ε_{it} : perturbación aleatoria de la región i en el período t . $E[\varepsilon_{it} \cdot \varepsilon_{is}] = \sigma_{is}$.

2. A partir de la ecuación estimada para la Comunidad de Madrid, en 1999, se estiman los datos de VAB para los 179 municipios (i) de dicha Comunidad

$$y_i = 697,265 + 0,382 \cdot x_i$$

$t\text{-Stud.}(4,3)$ $t\text{-Stud.}(8,9)$

Los principales resultados del VAB per capita de los municipios⁷⁸ de la Comunidad de Madrid son:

Mayor VAB total por hab. (1999)		Menor VAB total por hab. (1999)	
Municipio	VAB (mil.pta)	Municipio	VAB (mil.pta)
Ajalvir	9.305	Villar del Olmo	1.380
Torrelodones	6.267	Belmonte de Tajo	1.439
Humanes de Madrid	5.734	Vellón (El)	1.488
Morata de Tajuña	5.483	Cenicientos	1.508
Daganzo de Arriba	3.891	Villaconejos	1.546
Arganda del Rey	3.804	Tielmes	1.606
Loeches	3.774	Carabaña	1.665
Camarma de Esteruelas	3.679	Molinos (Los)	1.671
Griñón	3.437	Robledo de Chavela	1.681
Pinto	3.348	Perales de Tajuña	1.690

⁷⁸ Para evitar resultados de municipios más o menos desconocidos, se ofrecen sólo los datos de los municipios de más de 1.000 habitantes, aunque la estimación se realizó para todos ellos.

Referencias bibliográficas

- . ACS, Z., L. ANSELIN y A. VARGA (2001), "*Patents and innovation counts as measures of regional production of new knowledge*". *Research Policy*, 1368; pp. 1–17.
- . ACHEN, C. y W. SHIVELY (1995), "*Cross-level inference*". Chicago: University of Chicago Press.
- . ALAÑÓN, A. (2001), "*La renta regional en España: análisis y estimación de sus determinantes*". Tesis doctoral, Universidad Complutense de Madrid.
- . AMEMIYA, T. (1971), "*The estimation of the variances in a variance-components model*". *International Economic Review*, 12; pp. 1-13.
- . AMEMIYA, T. (1977), "*A note on a heteroskedastic model*". *Journal of Econometrics*, 6; pp. 365-70.
- . AMEMIYA, T. (1985), "*Advance Econometrics*". Cambridge, MA.: Harvard University Press.
- . AMRHEIN, C. (1996), "*Searching for the elusive aggregation effect: Evidence from statistical simulations*". *Environment and Planning A*, 27; pp. 105-119.
- . ANSELIN, L. (1980), "*Estimation methods for spatial autoregressive structures*". Regional Science Dissertation and Monograph Series (Ithaca, NY).
- . ANSELIN, L. (1984), "*Spacification tests on the structure of interaction in spatial econometric models*". *Papers, Regional Science Association*, 54; pp. 165-82.
- . ANSELIN, L. (1986A), "*MicroQAP: a Microcomputer implementation of generalized measures of spatial association*". Department of Geography, University of California, Santa Barbara.
- . ANSELIN, L. (1986B), "*Some further notes on spatial models and regional science*". *Journal of Regional Science*, vol. 26; pp. 799-802
- . ANSELIN, L. (1988A), "*Spatial econometrics: methods and models*". Kluwer Academic Publishers.
- . ANSELIN, L. (1988B), "*Langrange Multiplier test diagnostics for spatial dependence and spatial heterogeneity*". *Geographical Analysis*, 20; pp. 1-17.
- . ANSELIN, L. (1990A), "*Some robust approaches to testing and estimation in spatial econometrics*", *Regional Science and Urban Economics*, 20; pp. 141-63.
- . ANSELIN, L. (1990B), "*Spatial dependence and spatial structural instability in applied regression analysis*". *Journal of Regional Science*, 30; pp. 185-207.
- . ANSELIN, L. (1992A), "*Discrete space autoregressive models*". En "*GIS and Environmental Modeling*", M.F. Goodchild, B. Parks, L.T. Steyaert (Eds.), Oxford: Oxford University Press.

- . ANSELIN, L. (1992B), "*Space and applied econometrics. Introduction*". Regional Science and Urban Economics, 22; pp. 307-16.
- ANSELIN, L. (1992C), "*Spatial dependence and spatial heterogeneity: Model specification issues in the spatial expansion paradigm*". En "Applications of the Expansion Method" (London: Routledge); pp. 334-54.
- . ANSELIN, L. (1992D), "*Space Stat tutorial. A workbook for using SpaceStat in the analysis of spatial data*". Technical Report S-92-1, National Center for Geographic Information and Analysis, University of California. Santa Barbara, CA.
- . ANSELIN, L. (1993A), "*Exploratory spatial data analysis and Geographic Information Systems*". Actas del Seminario DOSES/Eurostat ("New Tools for Spatial Analysis"). Lisboa (Portugal), 18-20 de noviembre.
- . ANSELIN, L. (1993B), "*The Moran Scatterplot as an ESDA tool to assess local instability in spatial association*". Paper presented at the GISDATA Specialist Meeting on GIS and Spatial Analysis, Amsterdam, The Netherlands, December 1-5 (West Virginia University, Regional Research Institute, Research Paper 9330).
- . ANSELIN, L. (1994A), "*SpaceStat Version 1.50: Revision notes*". Research Paper 9428, Regional Research Institute, West Virginia University.
- . ANSELIN, L. (1994B), "*Testing for spatial dependence in linear regression models: A review*". Regional Research Institute Research Paper. West Virginia University, Morgantown.
- . ANSELIN, L. (1995A), "*Space Stat version 1.80: Users' guide*". Regional Research Institute, West Virginia University, Morgantown, WV.
- . ANSELIN, L. (1995B), "*Local Indicators of Spatial Association-LISA*". Geographical Analysis, vol. 27(2); pp. 93-115.
- . ANSELIN, L. (1996), "*The Moran scatterplot as an ESDA tool to assess local instability in spatial association*". En Fischer, M., H. Scholten y D. Unwin, (eds), "Spatial analytical perspectives on GIS". London, Taylor and Francis.
- . ANSELIN, L. (1998), "*Exploratory spatial data analysis in a geocomputational environment*". Actas de la Conferencia de GeoComputación 1998, Bristol (UK), 17-19.
- . ANSELIN, L. (1999A), "*The future of spatial analysis in the social sciences*". Geographic Information Sciences, 5 (2); pp. 67-76.
- . ANSELIN, L. (1999B), "*Spatial Data Analysis with SpaceStatTM and ArcView®. Workbook (3rd Edition)*". Department of Agricultural and Consumer Economics, University of Illinois, Urbana, IL 61801.
- . ANSELIN, L. (2000A), "*Computing environments for spatial data analysis*". Journal of Geographical Systems 2 (3), pp. 201-225.
- . ANSELIN, L. (2000B), "*The alchemy of statistics, or creating data where no data exist*". Annals, Association of American Geographers, 90; pp. 586-592.
- . ANSELIN, L. (2001A), "*Spatial effects in econometric practice in environmental and resource economics*". American Journal of Agricultural Economics, 83 (3), pp. 705-710.

- . ANSELIN, L. (2001B), "*Spatial econometrics. 2001 Summer course*". University of Illinois, Urbana-Champaign, <http://www.spacestat.com>.
- . ANSELIN, L. (2001C) "*Rao's score test in spatial econometrics*". Journal of Statistical Planning and Inference (próxima publicación).
- . ANSELIN, L. (2001D), "*Spatial econometrics*". En "A companion to theoretical econometrics", ed. Baltagi, Oxford: Basil Blackwell; pp. 310-330.
- . ANSELIN, L. (2001E), "*Spatial externalities, spatial multipliers and spatial econometrics*". Discussion Paper del Regional Economics Applications Laboratory, REAL 01-T-11.
- . ANSELIN, L. y S. BAO (1997), "*Exploratory Spatial Data Analysis*". En "Recent developments in spatial analysis" (Eds. Fischer y Getis), Springer-Verlag, Berlín; pp. 35-59.
- . ANSELIN, L. y A. BERA (1998), "*Spatial dependence in linear regression models with an introduction to spatial econometrics*". Handbook of applied economic statistics", eds. A. Ullah y D. Giles, New York: Marcel Dekker, Inc.; pp. 237-289.
- . ANSELIN, L. y W. CHO (2002A) "*Spatial effects and ecological inference*". Political Analysis,10 (próxima publicación).
- . ANSELIN, L. y W. CHO (2002B) "*Reply: Conceptualizing space*". Political Analysis,10 (próxima publicación).
- . ANSELIN, L. y R. FLORAX (1995), "*New directions in spatial econometrics: Introduction*". En "New directions in spatial econometrics", L. Anselin y R.J.G.M. Florax editores. Springer; pp. 3-18.
- . ANSELIN, L. y R. FLORAX (1995B), "*Small sample properties of tests for spatial dependence in regression models*". En "New directions in spatial econometrics", L. Anselin y R.J.G.M. Florax editores. Springer; pp. 21-74.
- . ANSELIN, L. y R. FLORAX (eds.) (2000), "*Advances in Spatial Econometrics*". Heidelberg: Springer-Verlag.
- . ANSELIN, L. y D.A. GRIFFITH (1988), "*Do spatial effects really matter in regression analysis?*". Papers Regional Science Association, 65.
- . ANSELIN, L. y S. HUDAK (1992), "*Spatial econometrics in practice: a review of software options*". Regional Science and Urban Econometrics, 22; pp. 509-36.
- . ANSELIN, L. y R. MORENO (2001), "*Properties of tests for spatial error components*". Actas del 41º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Zagreb. CD-ROM.
- . ANSELIN, L. y J. O'LOUGHLIN (1992), "*Geography of international conflict and cooperation: spatial dependence and regional context in Africa*". En "The New Geopolitics", M. Ward (Ed.), London: London and Breach; pp. 399-475.
- . ANSELIN, L. y S. REY (1991), "*Properties of tests for spatial dependence in linear regression models*". Geographical Analysis, 23, 2; pp. 113-131.
- . ANSELIN, L. y S. REY (1997), "*Introduction to the special issue on spatial econometrics*". International Regional Science Review, 20(1,2), pp. 1-8.

- . ANSELIN, L. y O. SMIRNOV (1996), "*Efficient algorithms for constructing proper higher order spatial lag operators*". *Journal of Regional Science*, 36; pp. 67-89.
- . ANSELIN, L., R. BONGIOVANNI y J. LOWENBERG-DEBOER (2002), "*A spatial econometric approach to the economics of site-specific nitrogen corn production*". Discussion Paper de la Regional Economics Applications Laboratory REAL 02-T-2, enero.
- . ANSELIN, L., A. VARGA y Z. ACS (2000), "*Geographic and sectorial characteristics of academic knowledge externalities*". *Papers in Regional Science*, 79 (4); pp. 435-443.
- . ANSELIN, L., A. BERA, R. FLORAX y M. YOON (1996), "*Simple diagnostic tests for spatial dependence*". *Regional Science and Urban Economics*, 26; pp. 77-104.
- . ANSELIN, L., I. SYANBRI, O. SMIRNOV y Y. REN (2001), "*Visualizing spatial autocorrelation with dynamically linked windows*". Discussion Paper de la Regional Economics Applications Laboratory REAL 01-T-10, noviembre.
- . APARICIO, M.T. et al. (1984), "*Una metodología para la estimación de la renta disponible municipal*". Actas de la VII Reunión de Estudios Regionales. Bilbao.
- . ARBIA, G. (1989), "*Spatial data configuration in statistical analysis of regional economics and related problems*". Dordrecht: Kluwer.
- . ARCARONS, J., G. GARCÍA y M. PARELLADA (1994), "*Estimació de la renda familiar disponible a les comarques i municipis de Catalunya. 1991*". Departament d'Economia i Finances, Direcció General de Programació Econòmica (Generalitat de Catalunya). Barcelona.
- . ARENTZE, T., A. BORGERS y H. TIMMERMANS (1994a), "*Geographical Information Systems and the measurement of accessibility in the context of multipurpose travel: A new approach*". *Geographical Systems*, 1; pp. 87-102.
- . ARENTZE, T., A. BORGERS y H. TIMMERMANS (1994b), "*Multistop-based measurements of accesibility in a GIS environment*". *International Journal of Geographical Information Systems*, 8; pp. 343-56.
- . ARORA, S. y M. BROWN (1977), "*Alternative approaches to spatial autocorrelation: an improvement over current practice*". *International Regional Science Review*, 2; pp. 67-78.
- . AYUNTAMIENTO DE MADRID (2000), "*Hacia un indicador de la renta familiar disponible per capita para los distritos y barrios del municipio de Madrid*". Departamento de Estadística, Primera Tenencia de Alcaldía, Madrid.
- . AZNAR, A., J. MUR y F. J. TRÍVEZ (1996), "*Métodos econométricos en el análisis regional*". Actas de la XXII Reunión de Estudios Regionales, de la Asociación Española de Ciencia Regional. Pamplona/Iruña, 20-22 de noviembre de 1996; pp. 237-264.
- . BACHERO, J., J. ESTEBAN, I. LÓPEZ y F. RUIZ (1997), "*Estimación de la renta familiar disponible, per cápita, a nivel municipal en la Comunidad Valenciana (año 1995)*". Actas de la XXIII Reunión de Estudios Regionales de la Asociación Española de Ciencia Regional (AEER), Valencia.
- . BALESTRA, P. y M. NERLOVE (1966), "*Pooling cross-section and time series data in the estimation of a dynamic model: the demand for natural gas*". *Econometrica*, 34; pp. 585-612.

- . BALLER, R., L. ANSELIN, S. MESSNER y D. HAWKINS (2001), "*Structural covariates of U.S. county homicide rates: incorporating spatial effects*". Criminology (próxima publicación).
- . BALTAGI, B. (1995), "*Econometric analysis of panel data*". Ed. John Wiley & Sons. UK.
- . BALTAGI, B. y D. LI (2001), "*LM tests for functional form and spatial error correlation*". International Regional Science Review, 24 (2); pp.194–225.
- . BANCO CENTRAL-HISPANO (1994), "*Atlas Comercial de España 1994*". Obra elaborada por el Instituto L. R. Klein (Universidad Autónoma de Madrid). Madrid.
- . BANESTO (1993), "*Anuario del Mercado Español 1993*". Madrid.
- . BAO, S., L. ANSELIN, D. MARTIN y D. STRALBERG (2000), "*Seamless integration of spatial statistics and GIS: the S-Plus for ArcView and the S+Grassland links*". Journal of Geographical Systems 2 (3), pp. 287–306.
- . BARRINGER, T. y T. SMITH (1992), "*Experiments with central-limit properties of spatial samples from locally covariant random fields*". Regional Science and Urban Economics, 22; pag. 387-403.
- . BASU, S. y G. REINSEL (1994), "*Regression models with spatially correlated errors*". Journal of the American Statistician Association, 89 (425); pp. 88-99.
- . BAUMONT, C., C. ERTUR y J. LE GALLO (2000), "*Geographic spillover and growth. A spatial econometric analysis for European regions*". Actas del 6º Congreso mundial de la RSI (Asociación Internacional de Ciencia Regional), Lugano (Suiza).
- . BERON, K.J. y W.P.M. VIJVERBERG (2000), "*Probit in a spatial context: a Monte Carlo analysis*". En L. Anselin and R. Florax (eds.), "*Advances in Spatial Econometrics*". Heidelberg: Springer-Verlag.
- . BESAG, J. (1974), "*Spatial interaction and the statistical analysis of lattice systems*". Journal of the Royal Statistical Society B, 36; pp. 192-236.
- . BIVAND, R. (1984), "*Regression modelling with spatial dependence: an application of some class of selection and estimation methods*". Geographical Analysis, 16; pp. 25-37.
- . BIVAND, R. (1998), "*A review of spatial statistical techniques for location studies*". CEPR symposium on New Issues in Trade and Location, Lund (Suecia); 28-30 agosto, 1998.
- . BLANCO, J. (1994), "*Ingresos personales en los municipios de la provincia de Sevilla*". Cuadernos de Socioeconomía Sevillana, nº2. Diputación de Sevilla.
- . BLOOMMESTEIN, H. (1983), "*Specification and estimation of spatial econometric models: A discussion of alternative strategies for spatial economic modelling*". Regional Science and Urban Economics, 13; pp. 251-130.
- . BLOOMMESTEIN, H. (1985), "*Elimination of circular routes in spatial dynamic regression equations*". Regional Science and Urban Economics, 15; pp. 121-130.
- . BODSON, P. y D. PEETERS (1975), "*Estimation of the coefficients of a linear regression in the presence of spatial autocorrelation. An application to a Belgian labour-demand function*". Environment and Planning A, 7; pp. 455-472.

- . BOLDUC, D., R. LAFERRIÈRE y G. SANTAROSSA (1992), "*Spatial autoregressive error components in travel flow models*". *Regional Science and Urban Economics*, 22; pp. 371-385.
- . BOLDUC, D., R. LAFERRIÈRE y G. SANTAROSSA (1995), "*Spatial autoregressive error components in travel flow models: An application to aggregate mode choice*". En "New directions in spatial econometrics", L. Anselin y R.J.G.M. Florax editores. Springer; pp. 96-108.
- . BOOTS, B.N. (1985), "*Size effects in the spatial patterning of nonprincipal eigenvectors of planar networks*". *Geographical analysis*, 17; pp. 74-81.
- . BOOTS, B.N. y G.F. ROYLE (1991), "*A conjecture on the maximum value of the principal eigenvalue of a planar graph*". *Geographical Analysis*, 23; pp. 276-82.
- . BOWDEN, R y D. TURKINGTON (1984) "*Instrumental Variables*", Cambridge University Press.
- . BOX, G. y G. JENKINGS (1976), "*Time series analysis, forecasting and control*". San Francisco: Holden Day.
- . BRETT, C. y J. PINKSE (1997), "*Those taxes all over the map! A test for spatial independence of municipal tax rates in British Columbia*". *International Regional Science Review*, 20; pp. 131-51.
- . BREUSCH, T. y A. PAGAN (1979), "*A simple test for heteroskedasticity and random coefficient variation*". *Econometrica* 47; pp. 1287-94.
- . BRUNSDON, C., A.S. FOTHERINGHAM y M.E. CHARLTON (1998), "*An investigation of methods for visualizing highly multivariate datasets*". En Unwin, D. y P. Fisher (eds), "Case studies of visualization in the social sciences, advisory group on computer graphics"; pp. 55-79.
- . BRYK, A. y S. RAUDENBUSH (1992), "*Hierarchical linear models*". Newbury Park: Sage Publications.
- . BUETTNER, T. y M. ZEW (1999), "*Local capital income taxation and competition for capital: the choice of the tax rate*". Actas del 39º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Dublín.
- . BURDEN, B. y KIMBALL, D. (1998), "*A new approach to the study of ticket splitting*". *American Political Science Review*, 92 (3); pp. 533-544.
- . BURRIDGE, P. (1980), "*On the Cliff-Ord test for spatial autocorrelation*"; *Journal of the Royal Statistic Society B*, 42; pp. 107-8.
- . BURRIDGE, P. (1981), "*Testing for a common factor in a spatial autorregressive model*". *Environment and Planning A*, 13; pp. 795-800.
- . BUTTON, K. y R. KULKARNI (2001), "*Spatial and distance statistics of the trucking and warehousing industries using GIS tools*". 41º Congreso de la Asociación Europea de Ciencia Regional (ERSA). Zagreb. CD-ROM.
- . CANO, R., J. CHICA y J. HERMOSO (1997), "*Estimación espacial del valor catastral del suelo urbano*". Actas de la XII Reunión de Asepelt-España. Bilbao.
- . CASE, A. (1992), "*Neighbouring influence and technological change*". *Regional Science and Urban Economics*, 22; pp. 491-508.

- . CASE, A., H. ROSEN y J. HINES (1993), "*Budget spillovers and fiscal policy interdependence: evidence from the states*". *Journal of Public Economics*, 52; pp. 285-307.
- . CASSETTI, E. (1972), "*Generating models by the expansion method: applications to geographical research*". *Geographical Analysis*, 4; pp. 81-91.
- . CASSETTI, E. (1986), "*The dual expansion method: an application for evaluating the effects of population growth on development*". *IEEE Transactions on Systems, Man, and Cybernetics SMC-15*; pp. 29-39.
- . CASSETTI, E. y J.P. JONES (1988), "*Spatial parameter variation by orthogonal trend surface expansions: an application to the analysis of welfare program participation rates*". *Social Science Research*, 16; pp. 285-300.
- . CASSETTI, E. (1997A), "*Mixed estimation and the expansion method: An application to the spatial modelling of the AIDS epidemic*". En "*Recent developments in spatial analysis*", eds. M. Fischer y A. Getis, Springer-Verlag, Berlín; pp. 15-34.
- . CASSETTI, E. (1997B), "*The expansion method, mathematical modeling, and Spatial Econometrics*". *International Regional Science Review*, 20 (1,2); pp. 9-34.
- . CASSETTI, E. y J. POON (1995), "*Economic models and spatial parametric instability: Relevant concepts and an instability index*". En "*New directions in spatial econometrics*", eds. L. Anselin y R. Florax, Springer-Verlag, Berlín; pp. 301-321.
- . CAZAR (1984), "*Renta municipal de Aragón. Año 1981*".
- . CAZAR (1986), "*Renta municipal de Aragón. Año 1984*".
- . CECCATO, V., R. HAINING y P. SIGNORETA (2002), "*Exploring offence statistics in Stockholm city using spatial analysis tools*". *Annals of the American Association of Geographers*, 92 (1); pp. 29-51.
- . CEREJEIRA, J. (1998), "*A inclusão do espaço em modelos de regressão linear: algumas notas introductórias*". *Comunications V National Meeting of APDR (Associação Portuguesa para o Desenvolvimento Regional), Regional Science Methodologies, Parallel Session 14. Coimbra, 18-20 de junio de 1998.*
- . CHASCO, C. (1997), "*Modelos de determinación de áreas de mercado del comercio al por menor*". *Tesina. Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid.*
- . CHASCO, C. (2001), "*Situación y crecimiento de los municipios de la Comunidad de Madrid*". En "*Situación económica y social de la Comunidad de Madrid en 1999*", ed. Consejo Económico y Social de la Comunidad de Madrid, Madrid.
- . CHASCO, C. (2001), "*Coyuntura municipal*". En "*Situación económica y social de la Comunidad de Madrid en 1999*", ed. Consejo Económico y Social de la Comunidad de Madrid, Madrid.
- . CHASCO, C. y J. VICÉNS (1998), "*Spatial interaction models applied to the design of retail trade areas*". *38º Congreso de la Asociación Europea de Ciencia Regional (ERSA) y Universidad de Viena. Viena. CD-ROM.*

- . CHASCO, C. y J. VICÉNS (2000), “*EU-membership impacts in the Spanish province income convergence: a spatial autocorrelation perspective*”. Anales de Economía Aplicada XIV Reunión de ASEPELT-España, Oviedo.
- . CHASCO, P. (1988), “*Modelos de localización de centros comerciales*”. Tesis doctoral. Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid.
- . CHICA, J. (1994), “*Teoría de las variables regionalizadas. Aplicación en economía espacial y valoración inmobiliaria*”. Biblioteca de Económicas y Empresariales. Servicio de Publicaciones de la Universidad de Granada. Granada.
- . CHILDERS, T. (1983), “*The nazi voter: The social foundations of fascism in Germany, 1919-1933*”. Chapel Hill: University of North Carolina.
- . CHILÈS, J-P. y P. DELFINER (1999). “*Geostatistics: Modeling spatial uncertainty*”. New York: John Wiley.
- . CHO, W. (1998), “*Iff the assumption fits...: a comment on the King ecological inference solution*”. Political Analysis, 7; pp. 143-163.
- . CHO, W. (2001), “*Latent groups and cross-level inferences*”. Electoral Studies, 20; pp. 243-263.
- . CHOU, Y., R. A. MINNICH, L. A. SALAZAR, J. D. POWER y R. J. DEZZANI (1990), “*Spatial autocorrelation of wildfire distribution in the Idyllwild Quadrangle, San Jacinto Mountain, California*”. Photogrammetric Engineering and Remote Sensing, 56; pp. 1507-13.
- . CHOU, Y. (1991), “*Map resolution and spatial autocorrelation*”. Geographical Analysis, 23, 3; pp. 229-246.
- . CIISA (1988), “*Renda municipal en Galicia*”. Fundación Caixa Galicia, Santiago de Compostela.
- . CLEVELAND, W. (1993), “*Visualizing data*”. Summit, NJ: Hobart Press.
- . CLIFF, A. y J. ORD (1972), “*Testing for spatial autocorrelation among regression residuals*”. Geographical Analysis, 4; pp. 267-284.
- . CLIFF, A. y J. ORD (1973), “*Spatial autocorrelation*”. London: Pion.
- . CLIFF, A. y J. ORD (1981), “*Spatial processes, models and applications*”. London: Pion.
- . CONSEJO ECONÓMICO Y SOCIAL DE LA COMUNIDAD DE MADRID (2001), “*Situación económica y social de la Comunidad de Madrid 1999*”. Madrid.
- . CONSEJO ECONÓMICO Y SOCIAL DE LA REGIÓN DE MURCIA (1999), “*Informe sobre la distribución intermunicipal de la renta. Disparidades intermunicipales de la Región de Murcia durante el período 1986-1996*”. Informe nº 1, Murcia.
- . COSTA, A. y C. ROVIRA (1994), “*La desagregación de la Cuenta de Renta de los Hogares en pequeñas áreas territoriales (comarcas y municipios)*”. Institut d’Estadística de Catalunya.
- . COSTANZO, C.M., L.J. HUBERT y R.G. GOLLEDGE (1983), “*A higher moment for spatial statistics*”. Geographical Analysis, 15; pp. 347-51.

- . COX, K. R. (1969), "*The voting decision in a spatial context*". Progress in Geography, 1. Eds. C. Board, R. J. Chorley, P. Hagget, D. R. Stoddart. Arnold, London; pp. 81-117.
- . CRESSIE, N. (1993), "*Statistics for spatial data..Revised edition*". New York: Wiley.
- . CUADRADO, J.R., MANCHA, T. y GARRIDO, R. (1998), "*Convergencia Regional en España. Hechos, tendencias y perspectivas*". Fundación Argentaria (Economía Española).
- . DACEY, M.F. (1969), "*Similarities in the areal distributions of houses in Japan and Puerto Rico*". Area, 3; pp. 35-37.
- . DAVIES, S. (2001), "*Quantitative methods: Advancement in ecological inference*". Progress in Human Geography, 25 (1); pp. 87-96.
- . DIPUTACIÓN DE SEVILLA (1995), "*Ingresos personales en los municipios de la provincia de Sevilla*". Cuadernos de Socioeconomía Sevillana, nº 2. Sevilla.
- . DYKES, J. (1998), "*Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv*". The Statistician, 47 (3); pp. 485-497.
- . DOREIAN, P. (1980), "*Linear models with spatially distributed data: spatial disturbances or spatial effects?*". Sociological Methods and Research, 13; pp. 243-269.
- . DORLING, D. (1993), "*From computer cartography to spatial visualization: A new cartogram algorithm*". 11th Symp. Computer Assisted Cartography, Minneapolis, Bethesda: ASPRS-ACSM; pp. 208-217.
- . DUBIN, R. (1995), "*Estimating logit models with spatial dependence*". En "New directions of spatial econometrics", eds. L. Anselin y R.J. Florax, Springer-Verlag, Berlín; pp. 229-242.
- . DUNCAN, O y B. DAVIS (1953), "*An alternative of ecological correlation*". American Sociological Review, 18; pp. 665-666.
- . DUNCAN, O., R. CUZZORT y DUNCAN, B. (1961), "*Statistical geography: Problems in analysing areal data*". The free press of Glancoe, Illinois.
- . DURBIN, J. y G.S. WATSON (1950), "*Testing for serial correlation in least squares regression I*". Biometrika, 37; pp. 409-428.
- . DURBIN, J. y G.S. WATSON (1951), "*Testing for serial correlation in least squares regression II*". Biometrika, 38; pp. 159-178.
- . DURBIN, J. y G.S. WATSON (1971), "*Testing for serial correlation in least squares regression III*". Biometrika, 58; pp. 1-9.
- . EFRON, B. (1979), "*Bootstrap methods: another look at the jackknife*". Annals of Statistics, 7; pp. 1-26.
- . EFRON, B. (1982), "*The jackknife, the bootstrap and other resampling plans*". Philadelphia, PA: SIAM).
- . EFRON, B. y R. TIBSHIRANI (1986), "*Bootstrap methods for standards errors, confidence intervals and other measures of statistical accuracy*". Statistical Science, 1; pp. 54-77.

- . ESRI (1999), *"ArcView GIS 3.2 Tutorial"*. Environmental Systems Research Institute, Nueva York.
- . FERNÁNDEZ, C. y SIERRA, Y. (1992), *"Estimación de la renta familiar disponible a nivel municipal. Una aplicación a La Rioja. Año 1985"*. Actas de la VI Reunión de Asepelt-España, Granada.
- . FINGLETON, B. (1999), *"Spurious spatial regression: Some Monte Carlo results with a spatial unit root and spatial cointegration"*. Journal of Regional Science, 39(1); pp. 1-19.
- . FISCHER, M. y A. GETIS (1997), *"Advances in Spatial Analysis"*. En "Recent developments in spatial analysis". eds. M. Fischer y A. Getis, Springer-Verlag, Berlín, pp. 1-14.
- . FISCHER, M. y A. VARGA (2001), *"Production of knowledge and geographically mediated spillovers from universities. A spatial econometric perspective and evidence from Austria"*. Actas del 41º Congreso de la ERS (Asociación Europea de Ciencia Regional), Zagreb. CD-ROM.
- . FLORAX, R. y H. FOLMER (1992), *"Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators"*. Regional Science and Urban Economics, 22; pp. 405-32.
- . FLORAX, R. y S. REY (1995), *"The impacts of misspecified spatial interaction in linear regression models"*. En "New directions in spatial econometrics". Ed. Springer, pp. 111-135.
- . FLORAX, R., H. FOLMER y S. REY (1998), *"The relevance of Hendry's econometric methodology"*. Discussion Paper T 98-125/4, Instituto Tinbergen.
- . FLORAX, R., H. FOLMER y S. REY (2000), *"Specification searches in spatial econometrics: The relevance of Hendry's econometric methodology"*. En "Advances in Spatial Econometrics", eds. L. Anselin y R. Florax, Heidelberg: Springer-Verlag.
- . FOTHERINGHAM, A. (2000), *"A bluffers guide to A Solution to the Ecological Inference Problem"*. Annals, Association of American Geographers, 90; pp. 582-586.
- . FOTHERINGHAM, A., M. CHARLTON y C. BRUNSDON (1997), *"Measuring spatial variations in relationships with Geographically Weighted Regression"*. En "Recent development in spatial analysis", eds. M. Fischer y A. Getis, Springer-Verlag, Berlín; pp. 60-82.
- . FREEDMAN, D. y S. PETERS (1984A), *"Bootstrapping a regression equation: some empirical results"*. Journal of the American Statistical Association, 79; pp. 97-106.
- . FREEDMAN, D. y S. PETERS (1984B), *"Bootstrapping an econometric model: some empirical results"*. Journal of Business and Economic Statistics, 2; pp. 150-58.
- . FREEDMAN, D., S. KLEIN, J. SACKS, C. SMYTH y C. EVERETT (1991), *"Ecological regression and voting rights"*. Evaluation Review, 15; pp. 673-711.
- . FREEDMAN, D., S. KLEIN, M. OSTLAND y M. ROBERTS (1999), *"Review of A Solution to the Ecological Inference Problem"*, Journal of the American Statistical Association, 94; pp. 355-357.
- . GARRISON, W. y D. MARBLE (1964), *"Factor-analytic study of the connectivity of a transportation network"*. Papers, Regional Science Association, 12; pp. 231-38.

- . GEARY, R. (1954), "*The contiguity ratio and statistical mapping*". The Incorporated Statistician, 5; pp. 115-45.
- . GEERTMAN, S. y J. RITSEMAN VAN ECK (1995), "*GIS and models of accessibility potential: An application in planning*". International Journal of Geographical Information Systems, 9; pp.67-80.
- . GENERALITAT DE CATALUNYA (1998), "*Estimació de l'indicador de renda familiar disponible de les comarques i els municipis de Catalunya. 1989-1995*". Departament d'Economia i Finances, Direcció General de Programació Econòmica. Barcelona.
- . GETIS, A. (1990), "*Screening for spatial dependence in regression analysis*"; Papers, Regional Science Association, 69; pp. 69-81.
- . GETIS, A. (1995), "*Spatial filtering in a regression framework: Examples using data on urban crime, regional inequality, and government expenditures*". En "New directions in spatial econometrics", eds. L. Anselin y R. Florax, Springer-Verlag, Belín; pp. 172-185.
- . GETIS, A. (2002), "*Comparative spatial filtering in regression analysis*". Geographical Analysis, 34 (1).
- . GETIS, A. y J. ORD (1992), "*The analysis of spatial association by use of distance statistics*". Geographical Analysis, 24; pp. 189-206.
- . GILKS, W., S. RICHARDSON, D. SPIEGELHALTER (1996). "*Markov chain Monte Carlo in practice*". London: Chapman and Hall.
- . GOICOLEA, A., J.A. HERCE y J.J. DE LUCIO (1998), "*Regional integration and growth: the Spanish case*". Actas del 38º Congreso de la Asociación Europea de Ciencia Regional (ERSA), ed. Universidad de Viena. Viena.
- . GOODCHILD, M.F. (1992), "*Geographic data modeling*". Computers and geosciences, 18; pp. 401-408.
- . GOODCHILD, M., L. ANSELIN, R. APPLEBAUM y B. HARTHORN (2000), "*Toward spatially integrated social science*". International Regional Science Review, 23(2); pp. 139-159
- . GOODMAN, L. (1953), "*Ecological regressions and the behavior of individuals*". American Sociological Review, 18; pp. 663-666.
- . GOOVAERTS, P. (1997). "*Geostatistics for natural resources evaluation*". New York: Oxford. University Press.
- . GRAAFF, T., R. FLORAX, P. NIJKAMP y A. REGGIANI (2001), "*A general misspecification test for spatial regression models: dependence, heterogeneity, and nonlinearity*". Journal of Regional Science, 41 (2), pp. 255-276.
- . GREEN, W.H. (1990), "*Econometric analysis*"; New York: Macmillan
- . GRIBOV, A., K. KRIVORUCHKO y J. VER HOEF (2001), "*Modified weighted least squares semivariogram and covariance model fitting algorithm*". En "Stochastic modeling and geostatistics. AAPG Computer Applications in Geology", vol. 2; ed. J.M. Yarus y R.L. Chambers.

- . GRIFFITH, D.A. (1987), "*Toward a theory of spatial statistics: Another step forward*". Geographical Analysis, vol. 19; pp. 69-82
- . GRIFFITH, D.A. (1988a), "*Advanced spatial statistics*". Dordrecht: Kluwer Academic.
- . GRIFFITH, D.A. (1988b), "*A reply to: Some comments on correction techniques for boundary effects and missing value techniques*". Geographical analysis, 20.
- . GRIFFITH, D.A. (1992), "*A spatially adjusted N-way ANOVA model*". Regional Science and Urban Economics, 22; pp. 347-369.
- . GRIFFITHS, W. (1972), "*Estimation of actual response coefficients in the Hildreth-Houck random coefficient model*". Journal of the American Statistical Association, 68; pp. 329-335.
- . HAGGET, P. (1981), "*Boundary problems in statistical geography*". En "Die Bedeutung von Grenzen in der Geographie", ed. H. Hishimoto, Zurich: Kummerly and Frey; pp. 59-67.
- . HAYNES, K., S. LALL, R. STOUGH y S. YILMAZ (2000), "*Network usage patterns and the substitution and complementarity effects between telecommunications and transportation: A demand side approach*". Actas del 6º Congreso de la RSI (Asociación Internacional de Ciencia Regional), Lugano (Suiza).
- . HAINING, R. (1986), "*Income diffusion and regional economics*". En "Transformations through space and time", eds. D.Griffith and R. Haining, pp. 59-80. Dordrecht: Martinus Nijhoff.
- . HAINING, R. (1994), "*Diagnostics for regression modeling in spatial econometrics*". Journal of Regional Science 34.
- . HAINING, R. (1995), "*Data problems in spatial econometric modeling*". En "New directions in spatial econometrics", eds. L. Anselin y R. Florax, Springer-Verlag, Berlín; pp. 156-171.
- . HAINING, R., S. WISE y J. MA (1998), "*Exploratory spatial data analysis in a geographic information system environment*". The Statistician, 48 (3); pp. 457-469.
- . HAINING, R., S. WISE y P. SIGNORETTA (2000), "*Providing scientific visualization for spatial data analysis: Criteria and an assessment of SAGE*". Journal of Geographical Systems, 2; pp.121-140.
- . HAUSMAN, J.A. (1978), "*Specification tests in econometrics*". Econometrica, 46; pp. 1251-1271.
- . HENRY, M., B. SCHMITT y V. PIGUET (2001), "*Spatial econometric models for simultaneous systems: Application to rural community growth in France*". International Regional Science Review, 24 (2); pp. 171-193.
- . HENRY, M., B. SCHMITT y V. PIGUET (1999), "*Spatial econometric models for simultaneous systems: Comparison of some models for selected regions in France*". Actas del 39º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Dublín.
- . HERAS, A. (1992), "*Un modelo general de estimación indirecta de la renta familiar disponible municipal. Su aplicación a la Comunidad Autónoma de Cantabria*". Tesis doctoral, Universidad de Cantabria. Santander.

- . HERAS, A., C. MURILLO y J.M. RODRÍGUEZ (1998), "*Información fiscal y estimación indirecta de la renta familiar disponible municipal en España*". Actas del I Congreso de Economía Aplicada. Barcelona.
- . HILDRETH, C. y J. HOUCK (1968), "*Some estimators for the linear model with random coefficients*". Journal of the American Statistical Association, 63; pp. 584-595.
- . HINTERDING, A., B. SLEPTSOV, A. WYTZISK, S.K. UHNE y U. STREIT (2000), "*Geostatistical and knowledge-based techniques in case of spatial interpolation of hourly precipitation*". Actas del 6º Congreso mundial de RSI, Lugano (Suiza).
- . HOOPER, P. y G. HEWINGS (1981), "*Some properties of space-time processes*". Geographical Analysis, 13; pp. 203-23.
- . HSIAO, C. (1986), "*Analysis of panel data*". Cambridge University Press, Cambridge.
- . HUBERT, L.J. (1985), "*Combinatorial data analysis: association and partial association*". Psychometrika, 50; pp. 449-67.
- . HUBERT, L.J. (1987), "*Assignment methods in combinatorial data analysis*". New York: Marcel Dekker.
- . HUBERT, L.J., R. GOLLEDGE y C.M. COSTANZO (1981), "*Generalized procedures for evaluating spatial autocorrelation*". Geographical Analysis, 13; pp. 224-33.
- . HUBERT, L.J., R. GOLLEDGE y N. GALE (1985), "*Measuring association between spatially defined variables: an alternative procedure*". Geographical Analysis, 17; pp. 36-46.
- . INSTITUTO DE ESTADÍSTICA DE ANDALUCÍA (1999), "*La renta de los municipios de Andalucía*". Estudios y Análisis, Junta de Andalucía. Sevilla.
- . INSTITUTO DE ESTADÍSTICA DE LA COMUNIDAD DE MADRID (2000), "*Estimación de las cuentas municipales de la industria madrileña 1997*". Consejería de Hacienda. Madrid.
- . INSTITUTO DE ESTADÍSTICA DE LA COMUNIDAD DE MADRID (2000), "*Estimación del Producto Municipal Bruto. 1996*". Consejería de Hacienda. Madrid.
- . INSTITUTO DE ESTADÍSTICA DE LA COMUNIDAD DE MADRID (2000), "*Indicador de renta familiar disponible municipal. 1995, 1996 (provisional) y 1997 (avance)*". Consejería de Hacienda. Madrid.
- . INSTITUTO DE ESTADÍSTICA DE LA COMUNIDAD DE MADRID (2001), "*Indicador de renta familiar disponible territorializada 1997, por municipios, distritos, barrios y secciones censales*". Consejería de Economía e Innovación Tecnológica. Madrid.
- . INSTITUTO NACIONAL DE ESTADÍSTICA DE ESPAÑA (2001), "*Contabilidad Regional de España. Base 1995. Cuentas de renta de los hogares. Serie 1995-1999*". www.ine.es
- . INSTITUTO NACIONAL DE ESTADÍSTICA DE PORTUGAL (1998), "*Estudo sobre o poder de compra concelhio 1997*". Gabinete de Estudos Regionais da Direcção Regional do Centro, número III. Lisboa.
- . INSTITUTO NACIONAL DE ESTADÍSTICA DE PORTUGAL (2000), "*Estudo sobre o poder de compra concelhio 2000*". Núcleo de Estudos Regionais da Direcção Regional do Centro, número IV. Lisboa.

- . ISARD, W. (1956), *“Location and space economy”*. John Wiley, New York.
- . ISARD, W. (1969), *“General Theory”*. Cambridge: MIT Press.
- . ISARD, W. y P. LIOSSATOS (1979), *“Spatial dynamics and optimal space-time development”*. New York: North Holland.
- . JOHNSTON, J. (1984), *“Econometric Methods.”* 3ª edición, McGraw-Hill.
- . JOHNSTON, R. y C. PATTIE (2001), *“On geographers and ecological inference”*. Annals, Association of American Geographers, 91 (2); pp, 281-282
- . JONES, J.P. y E. CASETTI (1992), *“Applications of the expansion method”*. London: Routledge.
- . JUDGE, G., W. GRIFFITHS, R. HILL, H. LUTKEPOHL y T-C. LEE (1985), *“The theory and practice of econometrics. 2nd Ed.”*. New York: Wiley.
- . KELEJIAN, H. y I. R. PRUCHA (2001), *“On the asymptotic distribution of the Moran I test statistic with applications”*. Journal of Econometrics 104, pp. 219 –257.
- . KELEJIAN, H. y D. ROBINSON (1992), *“Spatial autocorrelation: a new computationally simple test with an application to per capita county policy expenditures”*; Regional Science and Urban Economics, 22; pp. 317-31.
- . KELEJIAN, H. y D. P. ROBINSON (1995), *“Spatial correlation: A suggested alternative to the autoregressive model”*. En “New directions in spatial econometrics”, eds. L. Anselin y R. Florax, Springer-Verlag, Berlin; pp. 75-95.
- . KELEJIAN, H. y D. P. ROBINSON (1998), *“A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte Carlo results”*. Regional Science and Urban Economics, 28; pp. 389–417
- . KING, G. (1997). *“A solution to the ecological inference problem. Reconstructing individual behavior from aggregate data ”*. Princeton: Princeton University Press.
- . KING, G. (1999), *“The future of ecological inference research: A reply to Freedman et al.”*. Journal of the American Statistical Association, 94; pp. 1-10.
- . KING, G. (2000), *“Geography, statistics, and ecological inference”*. Annals of The Association of American Geographers 90; pp. 601-606.
- . KING, G., O. ROSEN y M. TANNER (1999), *“Binomial-Beta hierarchical models for ecological inference”*. Sociological Methods & Research, 28 (1); pp. 61-90.
- . KINGSLEY, H., S. LALL, R. STOUGH, S. YILMAZ (2000), *“Network usage patterns and the substitution and complementarity effects between telecommunications and transportation: A demand side approach”*. Actas del 6º Congreso mundial de la RSI, Lugano (Suiza). CD-ROM.
- . KLEIN, L.R. (1946), *“Remarks on the theory of aggregation”*. Econometrica, 14 (4); pp. 303-312.
- . KMENTA, J. (1971), *“Elements of econometrics.”*. New York: MacMillan.

- . KOENKER, R. (1981), "*A note on studentizing a test for heteroskedasticity*", *Journal of Econometrics*, 17; pp. 107-12.
- . KOENKER, R. y BASSETT (1982), "*Robust tests for heteroskedasticity based on regression quantiles*", *Econometrica*, 50; pp. 43-61.
- . KRISTENSEN, G. (1998), "*Spatial heterogeneity in Danish urban areas*". Actas del 38º Congreso de la Asociación Europea de Ciencia Regional (ERSA), ed. Universidad de Viena. Viena.
- . KRUGMAN, P. (1991), "*Increasing returns and Economic Geography*". *Journal of Political Economy*, 99; pp. 483-499.
- . KRUGMAN, P. (1993), "*First nature, second nature, and metropolitan location*". *Journal of Regional Science*, 33; pp. 129-144.
- . "LA CAIXA" (2001), "*Anuario Económico de España 2001*". Obra elaborada por el Instituto L. R. Klein (Universidad Autónoma de Madrid). Barcelona.
- . LANGFORD, I., A. LEYLAND, J. RASBASH y H. GOLDSTEIN (1999), "*Multilevel modeling of geographical distributions of diseases*". *Applied Statistics*, 48; pp. 253-268.
- . LAURIDSEN, J. (1999), "*Spatial cointegration analysis in econometric modelling*". Actas del 39º Congreso de la ERSa (Asociación Europea de Ciencia Regional), Dublín.
- . LAURIDSEN, J. (2000), "*Spatial autoregressively distributed lag models: Equivalent forms and estimation*". Actas del 40º Congreso de la ERSa (Asociación Europea de Ciencia Regional), Barcelona.
- . LAURIDSEN, J. y B. NAHRSTEDT (1998), "*Spatial patterns in intermunicipal Danish commuting*". Actas del 38º Congreso de la Asociación Europea de Ciencia Regional (ERSA), CD-ROM. Ed. Universidad de Viena. Viena.
- . LEGENDRE, P. (1993), "*Spatial autocorrelation: Trouble or new paradigm?*". *Ecology*, 74(6), pp. 1659-1673.
- . LÉGER, C., D. POLITIS y J. ROMANO (1992), "*Bootstrap technology and applications*". *Technometrics*, 34; pp. 378-98.
- . LELE, S. y K. ORD (1986), "*Besag's pseudo-likelihood: some optimality results*". Technical Report and Preprints nº 66, Department of Statistics, The Pennsylvania State University.
- . LESAGE, J. (1995), "*A multiprocess mixture model to estimate space-time dimensions of weekly pricing of certificates of deposit*". En "New directions in spatial econometrics", eds. L. Anselin y R.J. Florax, Springer-Verlag, Berlín; pp. 359-416.
- . LESAGE, J. (1999), "*Spatial econometrics*". Department of Economics, University of Toledo.
- . LESAGE, J. (2000). "*Bayesian estimation of limited dependent variable spatial autoregressive models*". *Geographical Analysis*, 32.
- . LESAGE, J. y R. KELLEY PACE (2000), "*Using matrix exponentials to explore spatial structure in regression relationships*". Discussion Paper de la Regional Economics Applications Laboratory REAL, octubre.

- . LESAGE, J.P. y A. KRIVELYOVA (1999), "*A spatial prior for Bayesian vector autoregressive models*". *Journal of Regional Science*, 39 (2); pp. 297-317.
- . LÓPEZ, A. (2002), "*Estructuras de producción sectoriales y complejos industriales en España. Aproximación al estudio de las divergencias regionales con la aplicación del análisis input-output*". Consejería de Economía y Empleo de la Comunidad de Madrid, Madrid.
- . LÓPEZ, F.A. y M.A. PALACIOS (2000), "*Distintos modelos de dependencia espacial*". *Anales de Economía Aplicada XIV Reunión de Asepelt-España*, Oviedo.
- . LÓPEZ, F.A., M.A. PALACIOS y M. RUIZ (2001), "*Modelos explicativos del desempleo en términos de localización. Una aplicación a las provincias españolas.*". *Anales de Economía Aplicada XV Reunión de Asepelt-España*, Santiago de Compostela.
- . LÓPEZ-BAZO, E., E. VAYÁ y R. MORENO (2001), "*Crecimiento y convergencia regional en España. ¿Qué estamos midiendo?*" En B. Cabrer, ed., "*Análisis Regional. El proyecto Hispalink*". Ed. Mundi-Prensa, Madrid.
- . LUBIN, J. (1994), "*Invited commentary: Lung cancer and exposure to residential radon*". *American Journal of Epidemiology*, 140; pp. 323-332.
- . MA, J., R. HAINING y S. WISE (1997), "*SAGE User's Guide*". Dpto. de Geografía y Sheffield Center for Geographic Information and Spatial Analysis", Universidad de Sheffield.
- . MACKINNON, J. y H. WHITE (1985), "*Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties*", *Journal of Econometrics*, 29; pp. 305-25.
- . MAGALHÃES, A., G. HEWINGS y C. AZZONI (2000), "*Spatial dependence and regional convergence in Brazil*". Discussion Paper de la Regional Economics Applications Laboratory, REAL 00-T-11, septiembre.
- . MAGNUS, J. (1982), "*Multivariate error components analysis of linear and non-linear regression models by maximum likelihood*". *Journal of Econometrics*, 7; pp. 281-312. (Corrigenda: *Journal of Econometrics*, 10; pp. 261).
- . MAHÍA, R. (1994), "*Modelos de gravitación: Áreas de atracción e intercambios comerciales*". Tesina. Instituto Lawrence R. Klein, Facultad de Ciencias Económicas y Empresariales de la Universidad Autónoma de Madrid.
- . MANTEL, N. (1967), "*The detection of disease clustering and a generalized regression approach*". *Cancer Research* 27; pp. 209-20.
- . MAPINFO (1995), "*MapInfo Professional. Guía del usuario*". MapInfo Corporation, Nueva York.
- . MAPINFO (1996), "*Vertical Mapper. Contour modeling and display software, vs. 1.5*". Northwood Geoscience Ltd., Ontario (Canada).
- . MARSCHAK, J. (1939), "*On combining market and budget data in demand studies: A suggestion*". *Econometrica*, 7; pp. 332-335.
- . MATHSOFT (1996), "*S+Spatialstats user's manual, version 1.0*". Seattle, WA, MathSoft, Inc.

- . MÁTYÁS, L. y P. SEVESTRE (1996), *"The econometrics of panel data: A handbook of the theory with applications"*. Kluwer Academic Publishers. Dordrecht (The Netherlands).
- . MCCUE, K. (2001), *"The statistical foundations of the EI method"*. The American Statistician, 55 (2); pp. 106-110.
- . MCMILLEN, D.P. (1995), *"Spatial effects in probit models: A Monte Carlo investigation"*. En "New directions in spatial econometrics", eds. L. Anselin y R.J. Florax, Springer-Verlag, Berlín; pp. 189-227.
- . MIELKE, P.W. (1979), *"On asymptotic non-normality of null distributions of the MRRP statistic"*. Communications in Statistics A, 8; pp. 1541-50.
- . MILLS, F.C. (1969), *"Métodos estadísticos"*. Ed. Aguilar.
- . MOLHO, I. (1995), *"Spatial autocorrelation in British unemployment"*. Journal of Regional Science, 35; pp. 641-658.
- . MORAN, P. (1948), *"The interpretation of statistical maps"*. Journal of the Royal Statistical Society B, vol. 10; pp. 243-251.
- . MORENO, R. y M. ARTÍS (1996), *"Further evidence about the complex nature of the link between public capital and productivity growth: the Spanish case"*. Actas de la XXII Reunión de Estudios Regionales, de la Asociación Española de Ciencia Regional. Pamplona/Iruña, 20-22 de noviembre de 1996; pp. 237-264.
- . MORENO, R., M. ARTÍS, E. LÓPEZ-BAZO y J. SURIÑACH (1997), *"Evidence on the complex link between infrastructure and regional growth"*. International Journal of Development Planning Literature, 12 (1&2), pp. 81-108.
- . MORENO, R., E. LÓPEZ-BAZO, E. VAYÁ y M. ARTÍS (2001), *"Externalities, public capital and costs of production"*. En "Advances in spatial econometrics", L. Anselin y R. Florax (Eds.), Heidelberg: Springer-Verlag (próxima publicación).
- . MORENO, R. y E. VAYÁ (2000), *"Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial"*. Edicions Universitat de Barcelona, colecció UB 44, manuals.
- . MUR, J. (1990), *"Correlación espacial: un estudio de Monte Carlo"*. Tesis doctoral. Departamento de Análisis Económico, Universidad de Zaragoza.
- . MUR, J. (1999a), *"Heterocedasticidad y autocorrelación espacial: una propuesta de series espaciales"*. Estadística Española, vol. 41, nº 144; pp. 79-104.
- . MUR, J. (1999b), *"Testing for spatial autocorrelation: moving average versus autorregressive processes"*. Environment and Planning A, 31(8), pp. 1371-1382.
- . MUR, J. (2000), *"Scale elements in spatial autocorrelation tests"*. Actas del 40º Congreso de la ERS (ASOCIACIÓN EUROPEA DE CIENCIA REGIONAL), Barcelona.
- . NOVALES, A. (1993), *"Econometría"*. Ed. McGraw-Hill. Madrid.
- . OBERHOFFER, W. y J. KMENTA (1974), *"A general procedure for obtaining maximum likelihood estimates in Generalized Regression Models"*. Econometrica, 42; pp. 549-90.

- . OGBURN, W. y I. GOLTRA (1919), "*How women vote: A study of an election in Portland, Oregon*". *Political Science Quarterly*, 34; pp. 413-433.
- . OKABE, A. y S. FUNAMOTO (2000), "*An exploratory method for detecting multi-level clumps in the distribution of points - a computational tool, VCM (variable clumping method)*". *Journal of Geographical Systems*, 2; pp. 111-120.
- . O'LOUGHLIN, J. y L. ANSELIN (1991), "*Bringing geography back to the study of international relations: spatial dependence and regional context in Africa, 1966-78*". *International Interactions*, 17.
- . O'LOUGHLIN, J., C. FLINT y L. ANSELIN (1994), "*The political geography of the nazi vote: context, confession, and class in the Reichstag Election of 1930*". *Annals, Association of American Geographers* 84; pp. 351-80.
- . OPENSHAW, S. (1984), "*Ecological fallacies and the analysis of areal census data*". *Environment and Planning*, 16; pp. 17-31.
- . OPENSHAW, S. (1993), "*Some suggestions concerning the development of artificial intelligence tools for spatial modelling and analysis in GIS*". En "*Geographic Information Systems, spatial modelling and policy evaluation*", ed. M.M. Fischer y P. Nijkamp, Berlin: Springer Verlag; pp. 17-33.
- . OPENSHAW, S. y P. TAYLOR (1979), "*A million of so correlation coefficients: three experiments on the modifiable areal unit problem*". En "*Statistical applications in the spatial sciences*", ed. N. Wrigley; Pion: London; pp. 127-144.
- . OPENSHAW, S. y P. TAYLOR (1981), "*The modifiable unit problem*". En "*Quantitative geography, a British view*", ed. N. Wrigley y R. Bennett, London: Routledge y Kegan; pp. 60-69.
- . ORD, J.K. (1975), "*Estimation methods for models of spatial interaction*". *Journal of the American Statistical Association*, 70; pp. 120-26.
- . ORD, J.K. y A. GETIS (1995), "*Local spatial autocorrelation statistics: distributional issues and an application*". *Geographical Analysis*, 27.4; pp. 286-306.
- . ORD, J.K. y A. GETIS (2001), "*Testing for local spatial autocorrelation in the presence of global autocorrelation*". *Journal of Regional Science*, 41(3); pp. 411-432.
- . OTERO, J.M. (1993), "*Econometría. Series temporales y predicción*". Editorial AC.
- . PACE, R.K. y R. BARRY (1997), "*Quick computation of spatial autoregressive estimators*". *Geographical Analysis*, 29(3); pp. 233-247.
- . PACE, R.K. y J. LESAGE (2002), "*Semiparametric maximum likelihood estimates of spatial dependence*". *Geographical Analysis*, 34 (2).
- . PAELINCK, J.H.P. (1985), "*Elements d'analyse économique spatiale*". Paris: Economica.
- . PAELINCK, J.H.P. (1990), "*Some new estimators in Spatial Econometrics*". En Griffith, D. (eds.): "*Spatial Statistics: Past, present and future*".
- . PAELINCK, J.H.P. y L.H. KLAASEN (1979), "*Spatial econometrics*". Saxon House.
- . PENA, B., dir (1996), "*Distribución personal de la renta en España*". Ed. Pirámide, Madrid.

- . PONS, J. y E. VILADECANS (1999), "*Kaldor's laws and spatial dependence. Evidence for the European Regions*". *Regional Studies*, vol. 33.5; pp. 443-451.
- . PULIDO, A. (2000), "*Una panorámica de diez años de análisis regional a través del proyecto Hispalink: sus expectativas de futuro*". En "Análisis Regional. El Proyecto Hispalink", ed. Bernardí Cabrer, Valencia.
- . PULIDO, A. y A. LÓPEZ (2000), "*Predicción y simulación aplicadas a la economía y gestión de empresas*". Ediciones Pirámide, S.A. Madrid.
- . PULIDO, A. y J. PÉREZ (2001), "*Modelos econométricos*". Ediciones Pirámide, S.A. Madrid.
- . QUANDT, R. (1958), "*The estimation of the parameters of a linear regression system obeying two separate regimes*". *Journal of the American Statistical Association*, 53; pp. 873-880.
- . RAO, J. (2000), "*Statistical methodology for indirect estimations in small areas*". Instituto de Estadística del País Vasco, Eustat, doc. 39.
- . REMÍREZ, J.A. (1991), "*Una estimación de la renta familiar disponible a nivel municipal. El caso de Navarra*". Fundación FIES, Documentos de trabajo, nº 76.
- . REY, S. (2001A), "*Spatial empirics for economic growth and convergence*". *Geographical Analysis* (próxima publicación).
- . REY, S. (2001B), "*Spatial dependence in the evolution of regional income distributions*". Discussion Paper de la Regional Economics Applications Laboratory REAL 01-T-03, marzo.
- . REY, S. (2001C), "*Spatial analysis of regional income inequality*". Discussion Paper de la Regional Economics Applications Laboratory REAL 01-T-9, octubre.
- . REY, S. y L. ANSELIN (2001), "*Regional science publication patterns 1990's*". *International Regional Science Review* (próxima publicación).
- . REY, S. y M. BOARNET (1998), "*A taxonomy of spatial econometric models for simultaneous equations systems*". Actas del 45º Encuentro Anual Norteamericano de la RSAI (Asociación Internacional de Ciencia Regional), Santa Fe (Nuevo Méjico).
- . REY, S. y B. MONTOURI (1999), "*US regional income convergence: a spatial econometric perspective*". *Regional Studies*, vol. 33.2; pp. 143-156.
- . RIVERS, D. (1998), "*A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*". *The American Political Science Review*, 92 (2); pp. 442-443.
- . ROBINSON, W. (1950), "*Ecological correlations and the behaviour of individuals*". *American Sociological Review*, 15 (3); pp. 351-357.
- . ROJO, J. y J. SANZ (2001), "*Una propuesta bayesiana para la distribución de Contabilidades regionales por procedimientos indirectos*". Actas de la XXVII Reunión de Estudios regionales de la AEER. Madrid.
- . ROYALTEY, H.E., E. ASTRACHAN y R. SOKAL (1975), "*Tests for pattern in geographic variation*". *Geographical Analysis*, 7; pp. 369-96.

- . SADEI (1994), *“La renta de los municipios asturianos 1994”*. Caja de Asturias, Avilés.
- . SAMPSON, R. (1987), *“Urban black violence: The effect of male joblessness and family disruption”*. American Journal of Sociology, 93; pp. 348-382.
- . SANDBERG, K. y J. JOHANSSON (2001), *“Estimation of hedonic prices for co-operative flats in the city of Umeå with spatial autoregressive GMM”*. Actas del 41º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Zagreb. CD-ROM.
- . SHIVELY, W. (1985), *“A strategy for cross-level inference under an assumption of breakage effects”*. Political Methodology, 11; pp. 167-179.
- . SMIRNOV, O. y L. ANSELIN (2001), *“Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach”*. Computational Statistics and Data Analysis 35 (3), pp. 301–319.
- . SMITH, T. (1980), *“A Central Limit Theorem for spatial samples”*. Geographical Analysis, 12; pag. 299-324.
- . SNEEK, J. y P. RIETVELD (1997), *“Higher order spatial ARMA models”*. Discussion Paper TI 97-043/3, Instituto Tinbergen.
- . SNEEK, J. y P. RIETVELD (1997), *“On the estimation of the spatial moving average model”*. Discussion Paper TI 97-049/4, Instituto Tinbergen.
- . SNELL, M. (1999), *“Estimating spatial models within a disequilibrium framework”*. Journal of Regional Science, 39 (4); pp. 591-612.
- . SOHN, J. (2002), *“Do birds of a feather flock together?: Economic linkage and geographic proximity”*. Discussion Paper de la Regional Economics Applications Laboratory REAL 02-T-4
- . SOKAL, R. (1979), *“Testing statistical significance of geographic variation patterns”*. Systematic Zoology, 28; pp. 227-32.
- . SOKAL, R.R., N.L. ODEN y B.A. THOMSON (1998), *“Local spatial autocorrelation in a biological model”*. Geographical Analysis, vol. 30.4; pp. 331-354.
- . SPSS Inc. (1999), *“Tutorial del programa SPSS para Windows, vs. 10.0.6”*. 27 de noviembre de 1999.
- . STATSOFT Inc. (2000), *“Electronic Text Book Statsoft”*. www.statsoft.com
- . STETZER, F. (1982), *“Specifying weights in spatial forecasting models: the results of some experiments”*. Environmental and Planning A, 14; pp. 571-584.
- . STOUGH, R.R., S.V. LALL y M.P. TRICE (1998), *“Public infrastructure and private capital: examination of the technology sector in the Washington DC metropolitan region”*. Actas del 38º Congreso de la Asociación Europea de Ciencia Regional (ERSA), . Ed. Universidad de Viena. Viena.
- . STOUGH, R.R., R. KULKARNI y J. PAELINCK (2001), *“Spatial-structural relations among technology industrial clusters: A comparative analysis of metropolitan regions in the U.S.”*. Actas del 41º Congreso de la Asociación Europea de Ciencia Regional (ERSA). Zagreb. CD-ROM.

- . SUI, D. (2000), "*New directions in ecological inference: An introduction*". Annals of the Association of American Geographers, 90 (3); pp. 579-606.
- . SUR, A. (2000A), "*Modelización del bloque de oferta de la Comunidad de Madrid*". En "Análisis Regional. El Proyecto Hispalink", ed. Bernardí Cabrer, Valencia.
- . SUR, A. (2000B), "*Elaboración de un indicador sintético para la Comunidad de Madrid (ISCAM)*". En "Análisis Regional. El Proyecto Hispalink", ed. Bernardí Cabrer, Valencia.
- . SUR, A. (2000C), "*Elaboración de series trimestrales del VAB de la Comunidad de Madrid*". En "Análisis Regional. El Proyecto Hispalink", ed. Bernardí Cabrer, Valencia.
- . SYMANZIC, J., T. KÖTTER, S. SCHMELZER, S. KLINKE, D. COOK y D. F. SWAYNE (1997), "*Spatial data analysis in the dynamically linked ArcView/XGobi/XploRe environment*". Computing science and statistics, Proc. of the 29th Symp. of the Interface. Fairfax Station, VA: Interface Foundation of North America, Inc.
- . TAGASHIRA, N. y A. OKABE (2002), "*The Modifiable Areal Unit Problem in a regression model whose independent variable is a distance from a predetermined point*". Geographical Analysis, 34 (1).
- . TALEN, E. (1995), "*The achievement of planning goals: A methodology for evaluating the success of planning*". Unpublished Ph.D. Dissertation, Department of Geography, University of California, Santa Barbara, CA.
- . THEIL, H. (1957), "*Linear aggregation of economic relations*". Amsterdam: North-Holland.
- . THOMAS, O. (2001), "*A few evidences about the current growth of French cities*". Actas del 41º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Zagreb. CD-ROM.
- . TIEFELSDORF, M. y B. BOOTS (1997), "*A note on the extremities of local Moran's I_s and their impact on global Moran's I^** ". Geographical Analysis, vol. 29.3; pp. 249-257.
- . TIEFELSDORF, M. (2000), "*Modelling spatial processes*". Springer-Verlag.
- . TOBLER, W. (1979), "*Cellular Geography*". En "Philosophy in Geography", ed. S. Gale y G. Olsson, Dordrecht: Reidel; pp. 379-86.
- . TORAL, A. (2001A), "*El factor espacial en la convergencia de las regiones de la UE: 1980-1996*". Tesis doctoral, ICADE, Universidad de Comillas. Madrid.
- . TORAL, A. (2001B), "*Regional growth and convergence in the Spanish provinces*". Actas del 41º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Zagreb. CD-ROM.
- . TUKEY, J.W. (1977), "*Exploratory Data Analysis*". Reading: Addison-Wesley.
- . TSUTSUMI, M., E. SHIMIZU y H. IDE (2000), "*A new approach to spatial interpolation*". Actas del 6º Congreso mundial de la RSI (Asociación Internacional de Ciencia Regional), Lugano (Suiza).
- . UNWIN, A. (1996), "*Exploratory spatial analysis and local statistics*". Computational Statistics, 11; pp. 387-400.

- . UNWIN, A. (2000), "*Using your eyes- making statistics more visible with computers*". Computational Statistics & Data Analysis, 32; pp. 303-312.
- . VAN DER KRUK, R. (2001), "*Economic Impacts of wetland amenities: A spatial econometric analysis of the Dutch housing market*". 41º Congreso de la ERSA (Asociación Europea de Ciencia Regional), Zagreb. CD-ROM.
- . VAYÁ, E. (1996), "*Efectos spillover regionales en la ecuación β -convergencia*". Actas de la XI Reunión de Asepelt-España. Albacete.
- . VAYÁ, E. (1998), "*Empirical analysis of spatial location of activity: a proposal of using spatial association statistics*". 38º Congreso de la Asociación Europea de Ciencia Regional (ERSA), ed. Universidad de Viena. Viena.
- . VAYÁ, E. y R. MORENO (2000), "*La utilidad de la econometría espacial en el ámbito de la ciencia regional*". Documento de Trabajo 2000-13, FEDEA. Madrid.
- . VAYÁ, E. y J. SURIÑACH (1996), "*Contrastes de autocorrelación espacial: una aplicación al ámbito de las provincias españolas*". Actas de la X Reunión Asepelt-España, , com. B-3-9. Ed. Junta de Comunidades de Castilla-La Mancha. Albacete.
- . VAYÁ, E., R. MORENO, E. LÓPEZ-BAZO y J. SURIÑACH (2001), "*Economic growth and spatial externalities*". En "Advances in spatial econometrics", L. Anselin y R. Florax (Eds.), Heidelberg: Springer-Verlag (próxima publicación).
- . VICÉNS, J. (1995), "*Modelos con variables cualitativas dicotómicas*". Instituto Lawrence R. Klein, Universidad Autónoma de Madrid, documento 95/05.
- . VICÉNS, J. (1996), "*Introducción a la modelización con datos de panel*". Instituto Lawrence R. Klein, Universidad Autónoma de Madrid, documento 96/03.
- . VICÉNS, J. (2001), "*Las nuevas tecnologías de la información en la Comunidad de Madrid. Resultados de una encuesta*". Fundación Madritel.
- . VICÉNS, J. y C. CHASCO (1998), "*Estimación de la renta familiar disponible municipal y regional de 1996*". Papeles de Discusión, nº 2, Servicio de Estudios de "la Caixa", Barcelona.
- . WALKER, R., E. MORAN y L. ANSELIN (2000), "*Deforestation and cattle ranching in the Brazilian Amazon: External capital and household processes*". World Development 28 (4); pp. 683-699.
- . WALL, P. y O. DEVINE (2000), "*Interactive analysis of the spatial distribution of disease using a geographic information system*". Journal of Geographical Systems, 2; pp. 243-256.
- . WALLACE, T.D. y A. HUSSAIN (1969), "*The use of error components models in combining cross-section and time-series data*". Econometrica, 37; pp. 55-72.
- . WARD, M.D. (1992), "*The New Geopolitics*". London: Gordon and Breach.
- . WARTENBERG, D. (1985), "*Multivariate spatial correlation: a method for Exploratory Geographical Analysis*". Geographical Analysis, vol. 17, nº 4; pp. 263-283.
- . WHEELER, C. (2001), "*A note on the spatial correlation structure on county-level growth in the US*". Journal of Regional Science, 41(3); pp. 433-449.

- . WHITE, H. (1980), "*A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity*". *Econometrica*, 48; pp. 817-838.
- . WHITTLE, P. (1954), "*On stationary processes in the plane*". *Biometrika*, 41, pp. 434-49.
- . WISE, S., R. HAINING y J. MA (1997), "*Regionalisation tools for the exploratory spatial analysis of health data*". En "Recent developments in spatial analysis", eds. M. Fischer y A. Getis, Springer-Verlag, Berlín; pp. 83-100.
- . WISE, S., R. HAINING y P. SIGNORETTA (1999), "*Scientific visualization and the exploratory analysis of area based data*". *Environment and Planning A*.
- . WOOD, C. y D. SKOLE (1998), "*Linking satellite, census, and survey data to study deforestation in the Brazilian Amazon*". En D. Liverman, E. Moran, R. Rindfuss, and P. Stern (eds.), "People and Pixels: Linking Remote Sensing and Social Science", pp. 70-93. Washington, DC: National Academies Press.
- . WRIGLEY, N., D. HOLT, D. STEEL y M. TRANMER (1996), "*Analyzing, modelling and resolving the ecological fallacy*". En "Spatial analysis: Modelling in a GIS environment"; ed. P. Longley y M. Batty, Cambridge UK; pp. 25-40.
- . YULE, U. y M. KENDALL (1950), "*An introduction to the theory of statistics*". Charles Griffin, Londres.
- . ZELLNER, A. (1962), "*An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias*". *Journal of the American Statistical Association*, 57; pp. 348-368.